A Microarchitectural Study on Apple's A11 Bionic Processor

Article · May 2018

CITATIONS
13

READS
11,825

1 author:

Debrath Banerjee
Arkansas State University - Jonesboro
5 PUBLICATIONS **17** CITATIONS

SEE PROFILE

A Microarchitectural Study on Apple's A11 Bionic Processor

Debrath Banerjee
Department of Computer Science
Arkansas State University
Jonesboro,AR,USA
debrath.banerjee@smail.astate.edu

***Abstract*—Over the 10 years of evolution in iPhone generations, world has experienced a revolutionary advancement in iPhone processor which was first brought into palm through iPhone first generation embedded with APL0098 processor. After a rapid progression in microarchitecture , currently iPhone market is dominated by Apple's new A11(SoC) Bionic processor chipped with iPhone 8 and iPhone X which is based on ARM big.LITLE architecture. Apple's new A11 is based of two performance cores to handle heavy duty multithreaded workloads and four efficiency cores to cover more mundane tasks when the requirements arises in order to preserve power consumption. A11 sports a new heavy duty performance controller which allows the chip to use these six cores at same time which is a great departure from A10 processor. The implementation of 10nm FinFET technology containing 4.3 billion transistors also led this processor's design up to the priority queue. This paper focuses on the insights of A11 processor respective to its performance compared to other predecessor in the iPhone family.**

Keywords—Bionic, big.LITTLE architecture, Neural Engine, FinFET process

I. INTRODUCTION

Apple put a dent in the mobile planet when they first came up with original iPhone back in 2007 and gradually after 10 years of evolution it becomes iconic to Bionic powered with ARM architecture. Before company moved to its iconic A Series, used package on package chipset with Power VR graphics chipset . Apple's first invented iPhone used APL0098 which had 412 MHz single core ARM11 CPU and a Power VR MBX Lite GPU manufactured by Samsung on 90nm process. This APL0098 processor has built in memory controller and run with single thread which has clock speed of 0.41GHz.Two widely used configuration are involved with its integrated circuit packaging where two or more memory only packages are on the stacked on each other and CPU package on the bottom and memory package on the top [1].

Apple moved to its well-known Ax series family on Systems on Chip only after the iPhone 3GS which used APL0298 , a last PoP version of iPhone generation. The following System on Chip version mainly used in the iPhone, iPad , iTouch and Apple TV which is designed by Apple and manufactured by Samsung & TSMC . A4 is the technically first A series processor combined with ARM Cortex A8 CPU with Power VR 535 GPU. The next in the queue A5 processor replaced the A4 which is used to power iPhone 4S , contained dual core ARM Cortex A9 CPU with ARM's advanced SIMD extension called NEON and a dual core Power VR SGX543MP2 GPU. According to Apple , the A5 was clocked at 1GHz on the iPad2 while it could dynamically adjust its frequency to save its battery life.A5 processor came up with two different variants of 45nm and 32nm ,where 32nm was said to provide 12% better battery life.

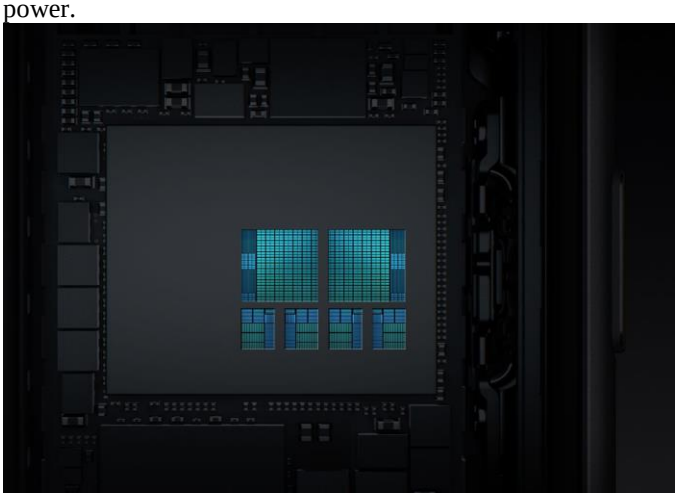
The high performance variant of Apple A5X was introduced when Apple launched third generation iPad. This SoC had a quadcore graphics unit instead of the previous dual core as well as quad core channel memory controller that provided a memory bandwidth of 12.8GB/sec which was about three times more than in the A5. The RAM in A5X was separate from main CPU package. A6 processor was landed with twice faster performance and twice the graphics powered compared to its predecessors. The A6 processor was manufactured by Samsung on high-K metal gate 32nm process. Apple designed custom ARMv7 based dual core CPU called Swift was clocked at 1.3Ghz ,integrated with 266MHz triple core PowerVR SGX 543MP3 GPU [1].

Apple A6X arrived at the same year with new quad core GPU PowerVR SGX554MP4 , quad channel memory and higher 1.4GHz CPU clock rate which 30 percent larger than A6 . A7 featured an Apple designed 1.3-1.4GHz 64bit ARMv8-A dual core CPU called Cyclone with four cluster configurations. The ARMv8-A architecture has 31 general purpose register that are each 64bits wide and 32 floating point/NEON registers of 128 bits wide. A8 featured an Apple designed 1.4GHz 64bit ARMv8-A[2] dual core CPU with four cluster configurations which contained 2 billion transistors that results its size reduced by 13 percent than A7.The A8X used a triple core CPU , a new Octa core dual channel memory and a higher 1.5GHz CPU clock rate . Afterwards the year of 2015 first saw a processor Apple A9 with dual sourced chipset on 14nm FinFET LPE process which had 70 percent more CPU performance compared to A8.Following to the A9X using 16nm process and 2 times faster GPU performance than A8X , A10 Fusion and A10X landed with new quadcore design with two high performance cores.

Finally, Apple introduced much awaited A11 Bionic powering iPhone X, iPhone 8 & iPhone 8 Plus is a 64bits ARM based system with two high performance cores and four high efficiency cores paired with Apple's new Neural Engine for handle AI tasks and face recognition. Rest of the paper will be drilling more insight to the architecture of A11 Bionic chip.

II. STATE OF ART A11 BIONIC PROCESSOR

In the age of octal core smartphones SoC, Apple is riding with six cores architecture into the battle which is once again proved itself as beast of the chip based on performance per watt characteristics. Apple's new A11 is based on ARM big.LITTLE architecture which features the 2x4 core structures. Two high performance core is called Monsoon and four high efficiency core is called as a Mistral. Monsoon are 25 percent and Mistrals are 75 percent faster than the high efficiency cores in the A10. A11 also has second generation performance controller which allows the chip to use all of its six core simultaneously to deliver the chip's top end horse power.



Two updated performance Core

A11 integrates an Apple's first internally designed three core graphics processing unit built into Bionic application processor and M11 motion coprocessor which gives 30 percent faster performance than A10. A11 also includes two completely innovative features which is combination of a new image processor and a dedicated neural network hardware. The image processor's job is to support lighting estimation , wide color capture and advanced pixel processing . The neural network hardware is called "Neural Engine" and can perform up to 600 billion operations per seconds.

A11 is manufactured by TSMC using 10nm [+]FinFET process which contains 4.3 billion transistors 2.53 GHz clock for 2 high performance CPU cores and 1.42 GHz clock for 4 low power CPU cores. A11 contain independent L2 cache ,meaning Mistral cores could be more independent than their A10 ancestors. The independence is underscored by the fact that Mistral cores share a common cluster id property while Monsoon core shares a distinct cluster id of their own. In depth microarchitecture of silicon engines are showcased in the following segments.

A. Six New CPU Cores & big.LITTLE Architecture

The major chunk of the A11 Bionic is the custom implementation of ARM architecture CPU cores. A10 processor got it names based on the capability to manage task between a pair of performance enabling flexibility between running at full power and efficient idleness. A11 processor is designed in such a way so that it can scale tasks across more low powers cores or to surge the workflow to its faster high power cores. Using asymmetric multiprocessing ,A11 Bionic can ramp up to activate any number of cores simultaneously.

A11 Bionic processor is designed based on ARM big.LITTLE technology [2]-[4] which is a heterogenous processing architecture uses two types of processor. "LITTLE" processors are designed for maximum power efficiency while "Big" processors are designed to provide maximum compute performance. Both types of processors are coherent and share same instruction set architecture. Using big.LITTLE technology each task can be dynamically allocated to big or little core depending on instantaneous requirement of the task. The key ingredient that makes big.LITTLE technology possible is coherency. This model requires transparent and performant transfer of the data between big and little processor. Hardware coherency enables this transparency to the processor .Without this hardware coherency the transfer of the data occur through memory which is slow and not power efficient. The idea of this technology is to dynamically allocate the task to the right processor. Different task have different constantly changing performance and requirements. In a typical system most task can be performed with "Little" cores however if the performance requirements goes above then it is sufficed by "Big" cores which provides great acceleration when needed . When performance requirement reduces, the big core could be turned off which results optimization in power consumption. Apple's A11 implemented this methodology in a very efficient way to optimize the performance based on dynamic task switching to the cores.

B. 3 Core GPU for Graphics – GPGPU

Deviating from the concept of imagination-based GPU , Apple introduced internally designed GPU which is built into A11 Bionic application processor claims to be faster than its predecessors. Normally GPU is for accelerating the graphics but for years they have been taken to perform other kind task with a similar repetitive nature, often referred as General Purpose GPU . Apple initially created Open CL as an API to perform GPGPU , and more recently folded GPGPU compute into its Metal API that is specifically optimized for GPU which Apple uses in its iOS device. Additionally, Apple is also forking into Machine Learning , one of the task which can be nailed by GPU very efficiently. Apple defines its A11 Bionic GPU family 4 graphics architecture as using Tile Based Deferred Rendering .On the desktop PC GPU "Immediate Mode" rendering is used on every triangle in the scene running through rasterization however TBDR breaks down the scene into tiles before it is analyzed that what needs to be rendered for each [2].

C. Dual Core ISP Nueral Engine

Apple's new A11 Bionic is incorporated with Neural Engine within its Image Signal Processor which can solve a problem like matching , analyzing , and calculating thousands of

[+]Fin Field Effect Transistor a type 3D transistor

reference points within a flood of image data rushing from the camera sensors. Neural Engine has two separate core which is designed to handle real time processing and capability of performing 600 billion operations per second. Neural Engine is responsible to give its name as Bionic which means that processor is enhanced with human like capabilities. This engine is designed to perform AI task like facial recognition with [1]"Deep Learning" techniques based on convolution operations. These processors are able to parallelize convolution operations leading to very fast computation. The secret sauce of the Neural engine , is its ability to handle matrix multiplications and floating point processing[4]-[5].

D. NVMe SSD storage controller

The six arithmetic cores of the A11 Bionic are assisted by SSD controller with an adapted error correcting code. This not only boost the iPhone also gives reliability to the storage. It enable the hardware side of reading and writing from solid state storage[5].

E. Video Encoder

Apple A9 first introduced with hardware based HEVC decoder enabling devices to efficiently playback H.265/High Efficiency video content.A11 introduced a hardware encoder, enabling iPhone's Biopic generations to create and save content in the high efficiency formats. This feature optimize the storage space while saving high definition photos and videos.

III. Strengths of A11 Bionic at a glance

Even though presence of Octa Core and Deca Core processor in market , A11 Bionic is much advanced compared to them.
The five parameter which is mostly used to determine the smartphones efficiency as follows

- Architecture
- Fabrication Procedure
- Number of Cores
- Number of Transistors and Clock speed
- Utilization , Intelligence and Power Consumption

A11 has been incorporated with two heavy lifting cores and four high efficiency core which gives an optimization to the distribution of the task which results less power consumption.
The usage of big.LITTLE architecture enables the A11 processor in such a way that it could handle the task using asymmetric multiprocessing. The GPU in the chip is three core Apple design which gives faster performance and uses less energy compared to other competitors in the mobile universe. The dual core Neural processor plays another important role by incorporating machine learning into the chip which is capable of 600 billion ops/second and appears to be pipelined through new ISP, which apparently has hardware noise reduction and faster face detection. It may also be linked with slow sync flash which can capture background details of the photo taken with flash in low light. Also, there is a new M11 motion coprocessor which features improved gyroscopes and inertial motion detection. There is also new video processor which segments the image into 2 million tiles and does motion and edge detection in conjunction with the hardware [2]HEIF and HEVC encoder/decoder. Finally, the A11 is manufactured by TSMC using 10nm FinFET process and contain 4.3 billion transistors on a die 87.66mm2 in size which gives several advantages like increased voltage headroom for circuits, lower gate resistance, higher current drive and higher gain.

IV. CONCLUSION

Over the ten years of domination in mobile universe ,Apple's new addition indeed set revolutionary benchmark respective to its performance and intelligence compared to another smart processor. Though most of the non-Apple chip processor also moving into FinFET process but the size of Apple chip is twice the size of non-Apple chip which means a lot to the performance. Deviating from the usage of third party PowerVR GPU which was used in the earlier apple ancestors, Apple's new Neural Engine combines both Neural networks and Machine Learning . Instead of explicitly programming the chip on how to perceive data, Machine Learning powers the device to read and learn from data. Apple always take the reference from its previous architecture and A11 is also not the exception which inherited property from A10 Fusion where high performance core and efficiency core were introduced but managing the core on 10nm CPU is one of the reason that make A11 Bionic more energy efficient. Through this advancement of the processor , Apple always target for more efficiency which could be achieved by silicon development. Based on the analysis of A11's six core architecture , it is evident that Apple is continuing pushing into Heterogenous Computing. Though we are unaware about the future generation of Apple processor , but it can be assumed that Apple's next design will have the intension to exploit more big.LITTLE architecture and introducing CPU core which are independently addressable, opening the door for mixed processor scenario. Any architectural changes ultimately circle back to improvements in some way. If Apple is making a change that includes the doubling of the lower power cores , it inevitable that they are going to spend more on die space with their own cache structures. In order to conclude , we can say this architectural change enables the modern CPU to manage performance aggressively and power consumption by dynamically changing clock speed, processor voltage and even disabling entire CPU cores by gating clocks and power to these cores which undoubtedly led Apple next generation processor ahead on the way to mobile planet.

V. REFERENCES

[1] Apple A11 - https://en.wikipedia.org/wiki/Apple_A11

[2] White paper on big. LITTLE Tech Copyright by ARM pp.1-5

[3]big.LITTLE Architecture by Subham Kamdar ;International Journal of Computer Application ; pp.35-37

[4]ARM Architecture Reference Manual Copyright by ARM pp.A32,34

[5] A11 Bionic Processor - https://appleinsider.com/

1-Broader family of machine learning based on learning data representations
2-High Efficiency Image File Format, Apple announced support on 2007