GPU Developments 2016

JPR
Jon Peddie Research
4 Saint Gabrielle Court
Tiburon, CA 94920-1619
(415) 435-9368
(415) 435-8214 Fax
www.jonpeddie.com

Table of Contents

Executive Summary 5
Introduction 6
Scope of report 9
First quarter 2016 9
AMD 9
ARM 10
Imagination Technologies 10
Intel 11
Nvidia 11
Qualcomm 11
Adjacent developments 12
Semiconductors 12
JSoftware 12
Other 12
Second quarter 2016 12
AMD 12
ARM 13
Nvidia 14
Qualcomm 15
Adjacent developments 15
Software 15
Third quarter 2016 16
AMD 16
Intel 17
Nvidia 17
Qualcomm 18
Adjacent developments 18
Fourth quarter 2016 19
AMD 19
Nvidia 19
Qualcomm 19
Think Silicon 19
Adjacent developments 20
Semiconductors 20
First transistor with a working 1-nanometer gate 20
Samsung starts production at 10nm 20
SUMMARY 21
Index 22

Table of Figures

Figure 1: Improvement in GFLOPS of GPUs over time 5
Figure 2: Relative die size of GPUs as the feature set gets smaller. Smaller die with more transistors. 7
Figure 3: Imagination Technologies' improved efficiency GPU IP (Imagination) 11
Figure 4: The previous generation of GPUs were scalable from 1 to 16 cores. Mali-G71 is scalable from 1 to 32 cores (ARM) 13
Figure 5: Nvidia's Tegra roadmap (Nvidia) 15
Figure 6: AMD revealed their GPU roadmap. 16
Figure 7: Nvidia's GPU roadmap (Nvidia) 18
Figure 8: Schematic of a transistor with molybdenum disulfide semiconductor and 1-nanometer carbon nanotube gate. (credit: Sujay Desai/Berkeley Lab) 20

Executive Summary

GPUs have been improving in performance, programmability, size, power consumption, and price since the term was first adopted in the late 1990s.

The advancement of performance, largely due to Moore's law, has be on all platforms, as illustrated in the following chart.

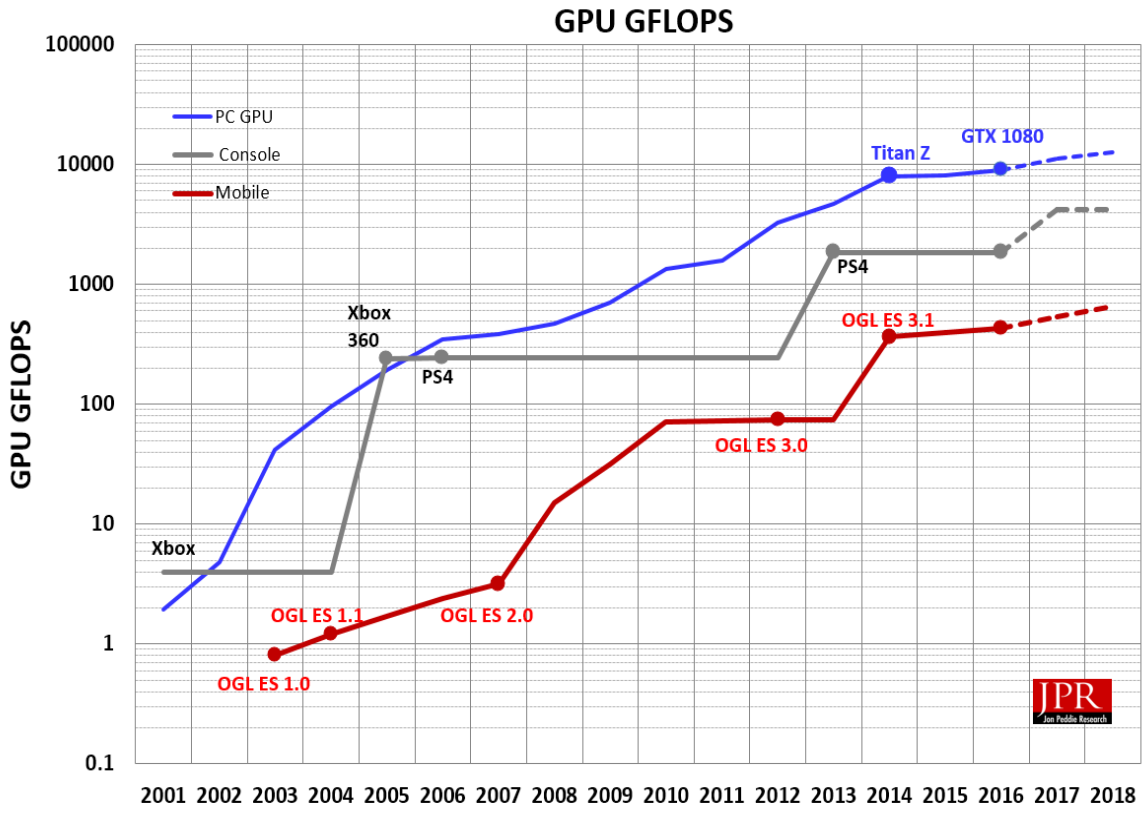

Figure 1: Improvement in GFLOPS of GPUs over time

The notion of smaller platforms (e.g., mobile devices), or integrated graphics (e.g., CPU with GPU) "catching up" to desktop PC GPUs is absurd—Moore's law works for all silicon devices. Intel's best integrated GPU today is capable of producing 1152 GFLOPS, which is almost equivalent to a 2010 desktop PC discrete GPU (i.e., 1300 GFLOPS).

Shaders are the other measure of a GPU's power, with mobile devices having 16 to 64, consoles 512 to 384, and desktop PCs up to 3800—more is better.

Clever techniques such as tessellation, deferred rendering, and texture compression have added to performance of GPUs by allowing them to deliver more frames per second while using less power.

The quest for higher performance, while staying within power budgets continues and even if Moore's law does slow down, the improvement of GPUs will not.

Introduction

The market for, and use of, GPUs stretches from supercomputers and medical instrumentation to gaming machines, mobile devices, automobiles, and wearables. Just about everyone in the industrialized world has at least a half dozen GPUs, and technophiles can easily count a dozen or more.

The ubiquity and invisibility of GPUs speaks to their success not just in marketing terms but in terms of real contribution to quality of life, safety, entertainment, and the advancement of science.

The GPU has evolved since its introduction in the late 1990s from a simple programmable geometry processor to an elaborate sea of 32-bit floating point processors running at multiple gigahertz speeds. The software supporting and exploiting the GPU, the programming tools, APIs, drivers, applications, and operating systems have also evolved in function, efficiency, and unavoidably, complexity.

The manufacturing of GPUs approaches science fiction with features that will move below 10nm next year and have a glide-path to 3nm, and some think even 1nm—Moore's law is far from dead, but is getting trickier to tease out of the Genie's bottle as we drive into subatomic realms that can only be modeled and not seen.

The partners to the GPU, memory, applications, and communications interfaces move at a slower pace and to some extent restrict the full advantages of GPU development. Such has always been the case, and such will always the situation.

The reduction of power consumption by these microscopic processors is perhaps the biggest challenge facing the developers. How to run at ever higher speeds, with ever smaller feature sizes, and not self-destruct due the heat generated in the process keeps engineers, scientists, programmers, and managers awake at night.

And yet at the end of the day, they've somehow done it—again. Every day in GPU land, a miracle happens, and several times a year new science is discovered. Ironically, GPUs have enabled the processing of the mathematics and data analysis needed to make those discoveries. The GPU is the father of future GPUs.

GPUs, or the elements thereof, have colonized associated and even orthogonal processors. It's almost impossible to find a non-GPU processor today that isn't heterogeneous. Almost all processors today have some form of similar-instruction, multiple-data (SIMD) parallel processing capability built into them. And many processors, from the tiniest IoT devices to biggest supercomputer processor have some type of pixel-pushing and raster-operations (ROPs) capability.

We acquire 90% of the information we digest through our eyes. Naturally we want and need abundant sources of information-generating devices to supply our unquenchable appetite for yet more information. And the machines we build to service and support us have a similar demand for information, albeit maybe not visible, although in some case such as in robots and autonomous

vehicles that's exactly what they need. The GPU can not only generate pixels, but it can also process photons captured by sensors.

No single device is good at everything, and specialized, dedicated processors such as a FPGA, can always outperform general purpose processors. The issue of when to use which is a function of economics, and life expectancy of the product or device. If a device is going to be made to monitor one or two, or even ten sensors, and never be changed in that activity, a dedicated device (assuming manufacturing costs and power consumption is in its favor) will be the better choice. But if that device is ever going to have to be expanded in its role, and/or if there is a probability that not all conditions could be thoroughly tested or simulated, it will likely become necessary to reprogram it, even if ever so slightly. And then the fixed function device loses its advantage. It's a trade-off and ROI analysis engineers and managers make every day.

The other big advantage the GPU has over most of its brethren processors is scalability. To date there doesn't seem to be an upper limit on how far a GPU can scale. The current crop of high-end GPUs has in excess of 3,000 32-bit floating-point processors, and the next generation will likely cross 5,000. That same design, if done properly, can be made with as few as two, or even one SIMD processor. The scalability of the GPU underscores the notion that one size does not fit all, nor does it have to. For example, Nvidia adopted a single GPU architecture with the Logan design in and used it in the Tegra 1. AMD used their Radeon GPUs in their APUs.

It's probably safe to say the GPU is the most versatile, scalable, manufacturable, and powerful processor ever built. Nvidia, which claims to have invented the term GPU (they didn't, 3Dlabs did in 1993), built their first device with programmable transform and lighting capability in 1999, at 220nm. Since then the GPU, from all suppliers, has ridden the Moore's law curve into ever smaller feature sizes, and in the process, delivering exponentially greater performance, prompting some observers to say GPUs have outpaced Moore's law. Today's high-end GPUs have over 15 billion transistors. The next generation is expected to feature as much as 32 GB with 2nd gen high-bandwidth memory (HBM2) VRAM and that will exceed 18 billion transistors.

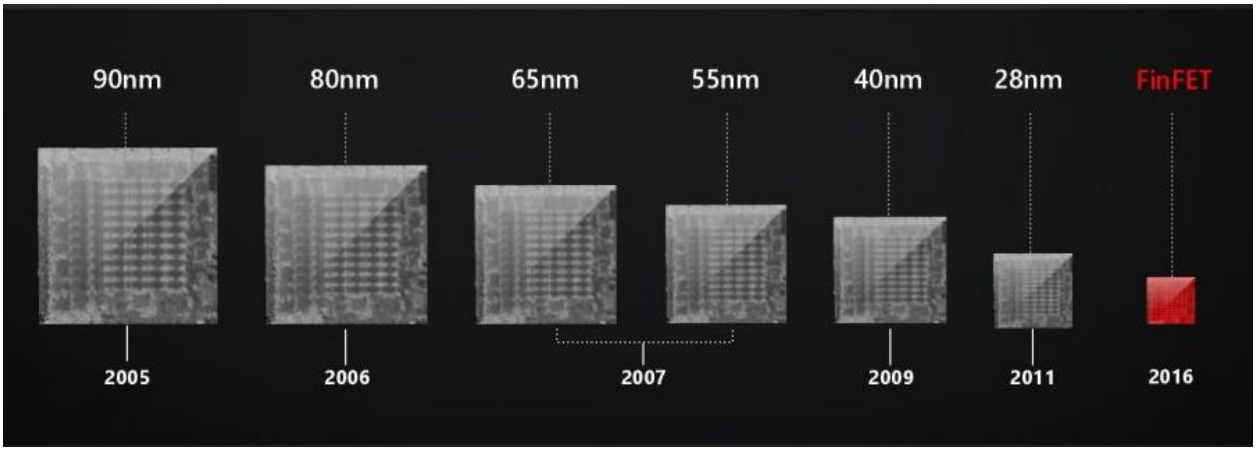

Figure 2: Relative die size of GPUs as the feature set gets smaller. Smaller die with more transistors.

Today GPUs are being produced in mass production at 14 and 16nm using 3D FinFET transistors. The next node is expected to be 10nm, and Intel is already producing test chips at that level,

Samsung is said to be doing the same, while TSMC is expected to offer 10nm in late 2017. GlobalFoundries says it will jump from its current production at 14nm to 7nm and not develop a 10nm line. And then the future has been questioned, is 7nm the end of Moore's law. The short answer is no. At IMEC in Belgium, they are doing statistical modeling of transistors at 5nm, and even 3nm. However, the chip industry has acknowledged that it's prepared for transistors to stop shrinking. Earlier this year, the Semiconductor Industry Association published a report announcing that by 2021 it will not be economically efficient to reduce the size of silicon transistors any further. Instead, chips look set to change in different ways. Different ways can mean new materials.

One such example is the announcement from Lawrence Berkeley National Laboratory (Berkeley Lab) of the first transistor with a working 1nm gate. That breaks through the 5, to 3nm quantum tunneling threshold; and may allow for Moore's law to continue. Until now, a transistor gate size less than 5 nanometers has been considered impossible because of quantum tunneling effects. (One nanometer is the diameter of a glucose molecule.).

Berkeley lab created a 2D (flat) semiconductor field-effect transistor using molybdenum disulfide (MoS2) instead of silicon and a 1D single-walled carbon nanotube (SWCNT) as a gate electrode, instead of various metals. (SWCNTs are hollow cylindrical tubes with diameters as small as 1 nanometer.)

Over the past ten months we have seen a few new, and some clever adaptations of GPUs that show the path for future developments, and subsequent applications.

Scope of report

For this report we have segmented the application and deployment of the GPU into five platforms: Supercomputers, Workstations and servers, PCs, Mobile devices, and embedded devices which include gaming consoles and vehicles.

You could enter that list anywhere and go in either direction, and find the transition and similarities very subtle. A GPU used in a digital watch will share many of the same characteristics, behaviors, and program paradigms as one used in a supercomputer. One will have an architectural and system configuration organizational similarity as the other. Not many, if any other processors can make such a claim.

Because of these similarities, the development work in GPUs is primarily done within the scope of two primary platforms, PCs, and mobile devices. All the other platforms are the beneficiary of those developments and the scalability of the design.

In the case of the PC, the development emphasis is about performance, and in some cases, performance at any cost. In the realm of supercomputers, performance, as measured with teraflops (TFLOPS) is first, power consumption second, and price third or fourth if you also include programmability or software tools.

In the mobile segment it's just the opposite, where power consumption is the most important factor, then price, and last is performance. As you go smaller and less expensive, into the land of the IoT, the GPU is reduced to the barest of functionality and performance is almost never an issue. In between you find PCs, game consoles, TVs, industrial and scientific instruments, and various vehicles from autonomous fork-lifts to F35 Joint Strike Fighter.

Let us now look at some of the key developments of GPUs during 2016 and what those developments suggest for future directions and trends.

First quarter 2016

Several announcements were made in the first quarter, dominated by the CES conference in January in Las Vegas, the Mobile World Congress in Barcelona, and the Game Developer's Conference in San Francisco.

AMD

In the first quarter of 2016, AMD announced the latest version of its graphics core next (GCN) architecture, its new Polaris GPU, would be fabricated in a 14nm process by GlobalFoundries (GF). Questioning the wisdom of committing to a fab that was late to get to a smaller node, AMD told us they were also getting parts from Samsung, and that the 14nm technology GF was using was from Samsung as part of the "copy-smart" agreed to in 2014.

The company introduced its A10 series of APUs, with R5 equivalent GPU that has 512 stream processors.

AMD also introduced the industry's first hardware segmented virtualized GPU. Later in the fourth quarter of 2016, AMD revealed Alibaba, one of the world's largest virtual GPU users and providers, had selected AMD's virtualized GPU, nudging Nvidia out of being the sole supplier to Alibaba.

Sony announced it would produce the PS4 Pro with 8 2.1GHz 'Jaguar CPU, and 36 'improved GCN compute units (3204 shaders cores) at 911MHz, yielding 2.3x FLOPS from the first generation PS4.

The company also introduced Xconnect (based on Thunderbolt) for external GPU with notebooks.

ARM

The company announced its Mali-DP650 supports several advanced features including up to seven display layer composition, rotation, high-quality scaling and energy-efficient technologies such as ARM Frame Buffer Compression (AFBC) within a very small silicon area. By doubling the size of AXI bus to 128-bit and providing a MMU pre-fetcher solution, Mali-DP650 can support more 4K composition layers and offers higher memory system latency tolerance than the Mali-DP550 display processor. Mali-DP650 is designed to connect to ARM's System MMU (ARM CoreLink MMU-500) enabling systems to produce up to 4K at 60fps display resolutions.

ARM also released (v1.0) of its VR SDK comprising of APIs, libraries, sample codes and tutorials for ARM Mali VR applications on Android.

Imagination Technologies

IMG demonstrated its PowerVR Wizard GR6500 ray tracing chip at CES.

The PowerVR Series7XT Plus GPUs w/Open CL support, the GT 7200 Plus, and the GT 7400 Plus; both with 128 ALUs, were introduced at MWC.

The company also announced its PowerVR Series 8XE family and claims the 8XE series can double the fill rate performance offered per mm^2. There are two designs GE8200 and GE8300. The 8200 generates 2 pixels/clock, and has 8 pipelines, and the 8300 has 16 and 4 pipelines and pixels/clock respectively.

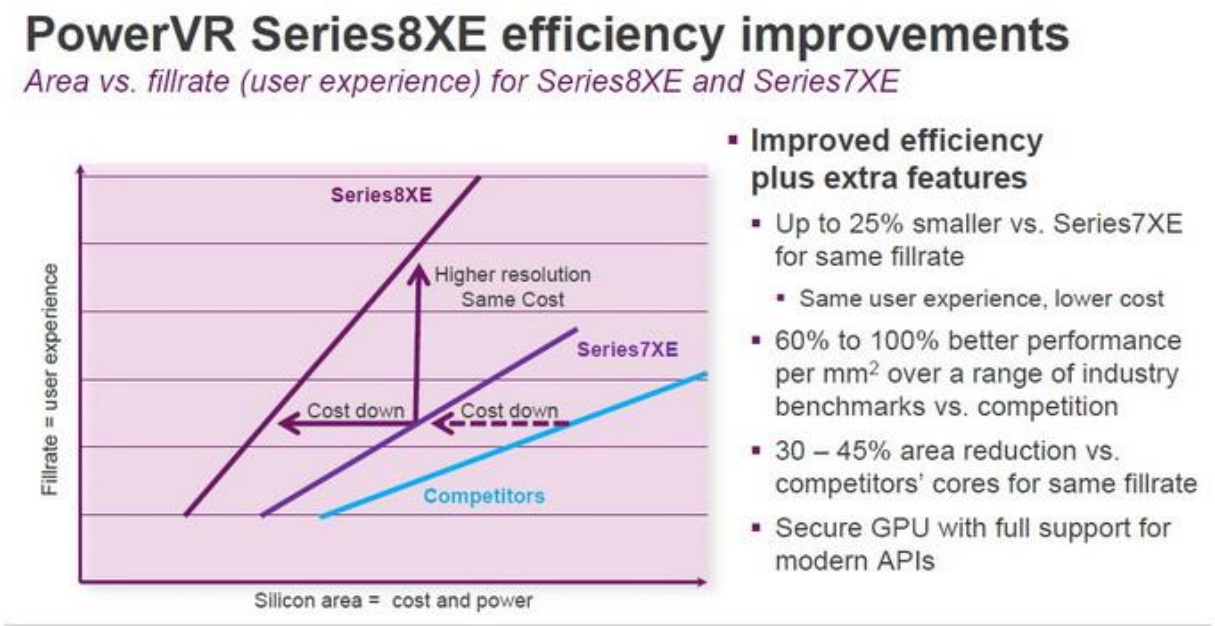

Figure 3: Imagination Technologies' improved efficiency GPU IP (Imagination)

The company also said that its PowerVR Rogue Series 6 GPUs received Khronos' OpenVX 1.0.1 conformance.

Intel

The company announced its 6th Gen Intel Core vPro processor with up to 2.5 times the performance, 3 times the battery life and a 30-times increase in graphics performance over a 5-year-old system.

Nvidia

Announced at GTC in 2013, Nvidia introduced its Tegra Parker architecture at CES 2016 in the Drive PX2 automotive computer system. Containing 256 CUDA cores, and triple display pipeline, and 2 Denver ARM V8 CPUs. The SoC (Parker) is built in 16nm at TMSC.

Qualcomm

Qualcomm announced in the first quarter that its SD602 had been selected by Audi for its infotainment systems in its2017 vehicles. That too was a surprise as Nvidia and Audi had been development partners.

The company also announced it had ported Vulkan to the Adreno 550.

Announced the SD 625 with Adreno 506 GPU (~130 GFLOPS), the SD 435 with Adreno 505 (48.6 GFLOPS), and the SD425 with Adreno 308 (>22 GFLOPS).

First Qualcomm Snapdragon 820 powered smartphones announced.

Snapdragon 821 introduced with Kryo quad-core CPU, reaching speeds up to 2.4GHz

Adjacent developments

Semiconductors

ARM and TSMC announced a multi-year agreement to collaborate on a 7nm FinFET process technology which includes a design solution for future low-power, high-performance compute SoCs.

JSoftware

Ghent University released Quasar GPU compute complier for high-level GPU-compute programming.

Geomerics, an ARM company, announced its Enlighten real-time global illumination technology will deliver large-scale dynamic lighting to open-world games. It includes advanced level of detail mechanisms for terrain, non-terrain light maps and probes. By solving the global illumination for distant geometry at lower resolutions than nearby geometry, users can achieve higher quality dynamic global illumination within the same map size and performance budget, or improve performance without sacrificing the user experience.

Other

European union funds research into low power graphics GPU consortium awarded €2.97m R&D grant to research and develop power and performance analysis for applications running on low power graphics processor units.

Second quarter 2016

The second quarter is usually a slow period with announcements of mid-life kicker products, and one major conference, Computex, to be used for new product introductions.

AMD

The 28nm 7th generation Carrizo APU the A10, was introduced, with 512 stream processors at 866 MHz.

The company introduced a new line of workstation boards, the first with 32 GB of memory.

One of the advantages that some of Intel Corp.'s integrated graphics processors (IGPs) have compared to AMD's IGPs is a large level four cache that is used to store frequently used data. While at present AMD's integrate graphics adapters are still faster compared to Intel's, in the future the latter may become considerably more competitive. In a bid to ensure that its IGPs are the fastest on the market, AMD reportedly plans to equip them with high-bandwidth memory that will act like cache.

ARM

ARM introduced its Mali-G71 based on its new Bifrost architecture, with support for Vulkan and OpenCL 2.0. The new design features "claused shaders," which the company says allows grouping sets of instructions together into defined blocks that will run to completion atomically and uninterrupted. This means all external dependencies are in place prior to clause execution and execution units can be designed to allow temporary results to bypass accesses to the register bank. This reduces the pressure on the register file, decreasing the amount of power it consumes and also contributes to area reduction by simplifying the control logic in the execution units.

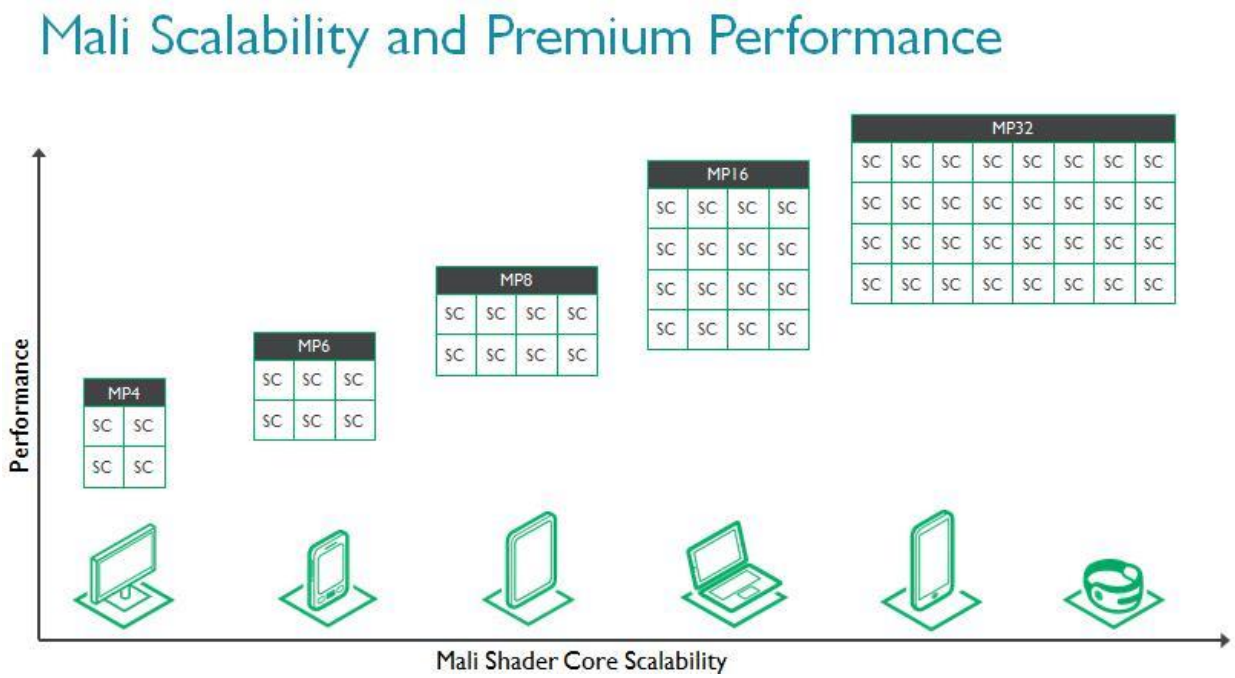

Figure 4: The previous generation of GPUs were scalable from 1 to 16 cores. Mali-G71 is scalable from 1 to 32 cores (ARM)

Bifrost architecture uses Quad based vectorization. Midgard, ARM's previous architecture, used SIMD vectorization which executed one thread at a time in the pipeline stage and was very dependent on the shader code executing vector instructions. Quad vectorization allows four threads to be executed together, sharing control logic. This makes it much easier to fill the execution units, achieving close to 100% utilization and better fits recent advances in how developers are writing shader code. Mali-G71 delivers 20% higher energy efficiency compared to Mali-T880 under similar conditions.

ARM acquired Apical Limited, an imaging and embedded computer vision IP company. Apical's technology will complement Mali graphics, display and video processor roadmap with dedicated silicon IP blocks that deliver an on-chip computer vision capability by converting raw sensor data or video into a machine-readable representation of an image. Apical's Assertive Display enables screens to adapt to changes in light by overcoming brightness limitations while reducing power consumption. Assertive Camera for ISPs and software packages, for high dynamic range, noise reduction and color management.

In July, SoftBank bought ARM for £24.4bn ($34b).

Nvidia

Nvidia announced its new Pascal GPUs with the GTX 1080, 2560 CUDA cores at 9 TFLOPS, would be available in Q2'16.

The GTX 1080 has been compared to AMD's RX 480, even though they are not in the same class.

	RX 480 (Polaris 10)	GTX 1080 (Pascal GP104)
Die arera MM2	232	315
Transistors (bil)	5.7	7.2
Density (million transistors /mm2)	24.57	22.86

Technical and Performance Metrics
GP104 is a 36% larger die.
GP104 has 21% more transistors.
Polaris 10 is 7.5% more dense than GP104.
GTX 1070 and 1080 are 70-80% more efficient depending on 1080p or 1440p.
GTX 1080 is 75-85% faster depending on 1080p or 1440p.
GTX 1070 is 50% faster at every resolution.

Perf/$
RX 480 8gb is 66% more cost effective (perf/$) than GTX 1080 (current prices).
Crossfire RX 480 8gb is 17% more cost effective than GTX 1080.
RX 480 8gb is 25% more cost effective than GTX 1070.
GTX 1070 is 9-14% more cost effective than crossfire RX 480 8gb.
RX 480 4gb is 50% more cost effective than GTX 1070 and twice as cost effective as a GTX 1080.

The company also showed its plans for its smaller Tegra SoC.

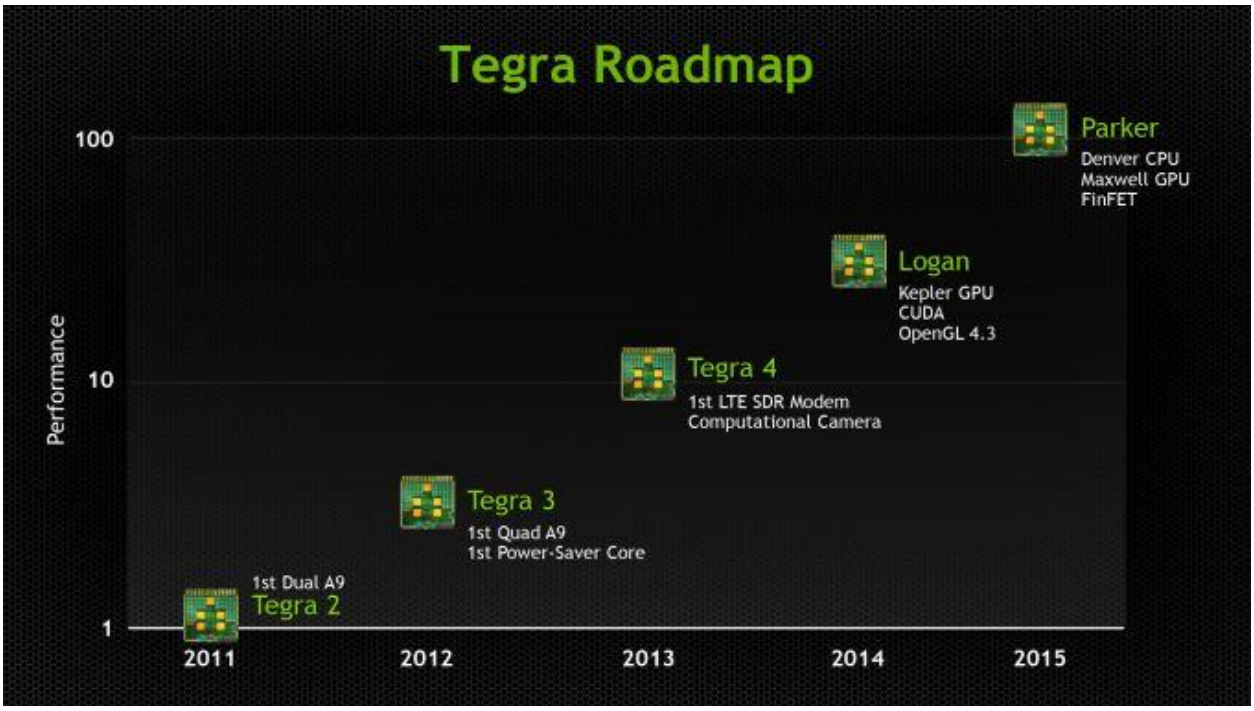

Figure 5: Nvidia's Tegra roadmap (Nvidia)

Nvidia has repositioned the Tegra into the automotive sector.

Qualcomm

The company announced it is working with Google on an initiative bringing Android OS embedded into the car. The initiative aims to help car makers create infotainment systems using Android as a common platform, making it easier to add connected services and applications with a safer and intuitive driving experience. The goal is to accelerate innovation in the car with an approach that offers openness, customization, and scale. The concept car functions demonstrated at Google I/O run on the Snapdragon 820 Automotive processor for connected cars and infotainment.

Adjacent developments

Few developments in Q2

Software

Imagination Technologies releases PowerVR Graphics SDK v4.1 with new functionality, examples, documentation, and support for Vulkan 1.0.

Qualcomm announces a deep learning software development kit (SDK) for devices powered by Snapdragon 820 processors.

Third quarter 2016

The first part of the third quarter is slow due to vacations, with Siggraph being the major event in the period. The Hotchips conference often is used to announce or introduce new semiconductors.

AMD

AMD introduced its new RX 480 with 2048 stream processors at 5.8 TFLOPS, built with 14nm FinFETs at GF. The GPU is based on AMD's GCN 4.0 architecture and has a number of fundamental features that define the it including:

- Primitive Discard Accelerator
- Hardware Scheduler
- Instruction Pre-Fetch
- Improved Shader Efficiency
- Memory Compression

The new GPU also incorporates h.265 decode at up to 4K and encode at 4K and 60 FPS. It does not incorporate HBM as the previous generation did.

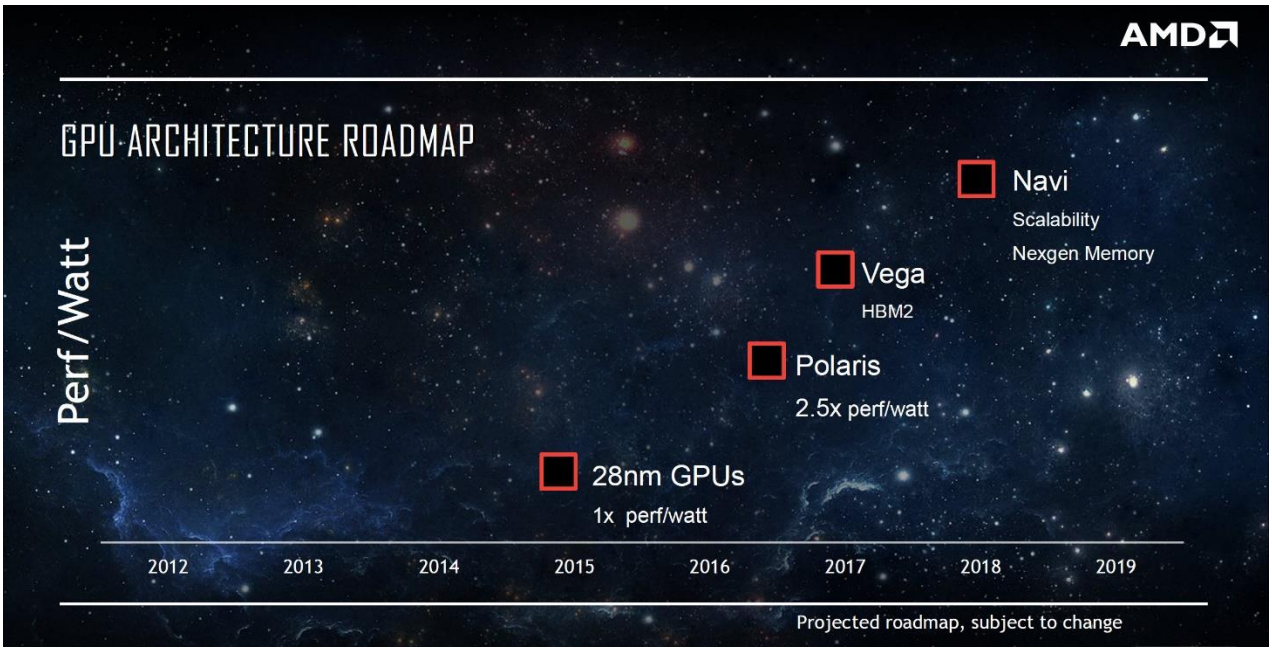

Figure 6: AMD revealed their GPU roadmap.

AMD's upcoming Vega GPU which has also been known as Greenland will feature 4096 stream processors. These are not the current generation stream processors but utilize the advancements made in the IP v9.0 generation of graphics SOCs under development by AMD. It is also noted that this chip is the "Leading Chip" of the first graphic IP v9.0 chip generation. The Vega 10 GPU is expected to have as much as 32 GB of HBM2 VRAM and use 18 billion transistors

Introduced A12 APU with R7 GPU with 512 stream processors. This will be last APU with last generation GCN architecture.

The Radeon Pro WX Polaris-based workstation series was introduced: 7100 (8GB, >5 TFLOPS), 5100 (8GB, >4 TFLOPS), 4100 (4GB, 2.5 TFLOPS), as well as the Radeon SSG with 1TB SSD. The Vega 10 GPU with 32 MB next generation of HBM2 memory will be aiming at the heart of HPC computing in FirePro cards and HPC APUs.

The Radeon Pro Solid State Graphics (SSG), workstation-class AIB was announced which includes M.2 slots for adding NAND SSDs.

Alibaba announced it would use AMD's FirePro S7150 x2 GPU in its cloud server systems. It's a major coupe for AMD to gain traction in the segment Nvidia has been dominating, and in particular with a big Nvidia customer like Alibaba.

AMD Announced Embedded Radeon E9260 & E9550 Polaris for Embedded Markets.

Company introduced its midrange Radeon RX 470 and RX 460.

Intel

Intel introduced the 7th generation core processors, basically a process node jump to 14 nm. The company claims the integrated HD graphics offers 3.5× better 3D graphics performance than a 5-year-old PC. The processor includes HEVC 10-bit decode capability up to 4K UHD, and a new VP9 decode capability.

Nvidia

Nvidia added the Nvidia GeForce GTX 1060 with a starting price of $249 to its Pascal family of gaming GPUs, complementing the GTX 1080 and 1070 following their launch two months earlier. The GTX 1060 has 1280 CUDA cores, 6GB of GDDR5 memory running at 8Gbps and a boost clock of 1.7GHz, which can be easily overclocked to 2GHz for further performance.

The GTX 1060 also supports Nvidia Ansel technology, a game-capture tool that allows users to explore, capture and compose gameplay shots, pointing the camera in any direction, from any vantage point within a gaming world, and then capture 360-degree stereo photospheres for viewing with a VR headset or Google Cardboard.

Nvidia Also announced Xavier, a-new SoC based on the company's next-gen Volta GPU, which Nvidia hopes will be the processor in future self-driving cars. Xavier features a high performance GPU, and the latest ARM CPU, yet has great energy efficiency according to the company.

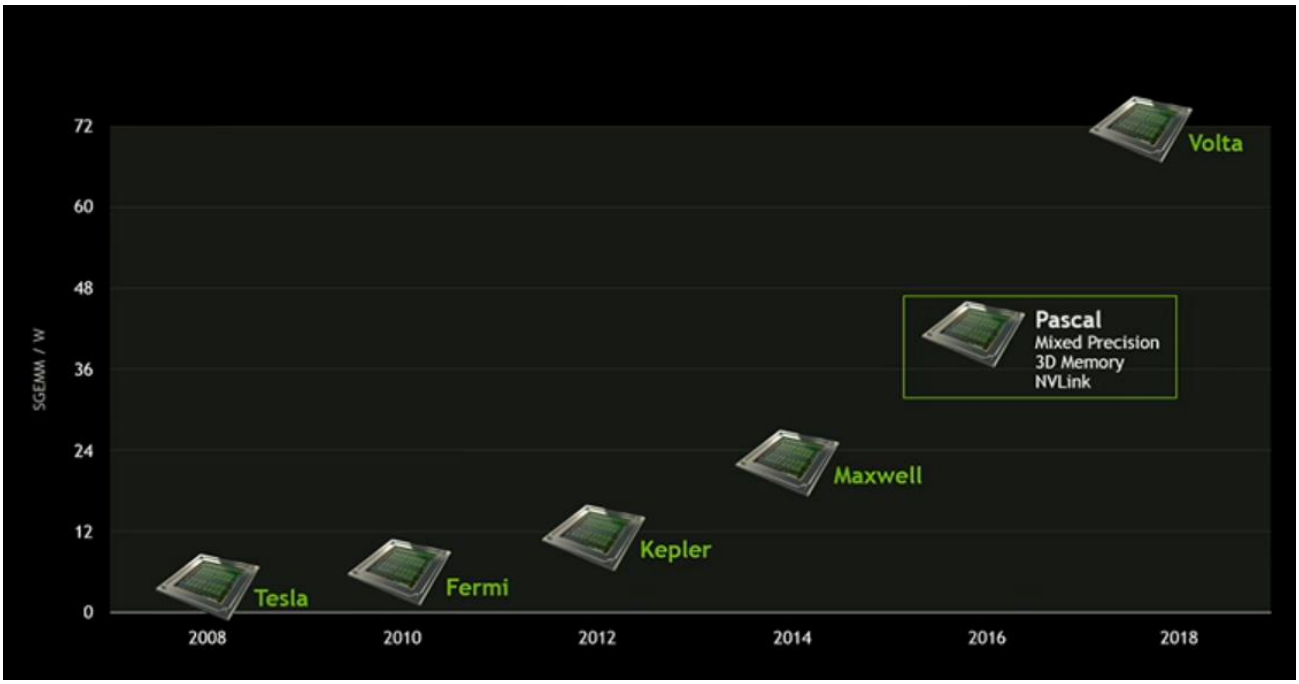

Figure 7: Nvidia's GPU roadmap (Nvidia)

Using the expanded 512-core Volta GPU in Xavier, the chip, is designed to support deep-learning features important to the automotive market, says the company. A single Xavier-based AI car supercomputer will be able to replace today's fully configured Drive PX 2 with two Parker SoCs and two Pascal GPUs. Xavier will be built using 16nm FinFET process and have seven-BILLION transistors—which has to be the biggest chip ever built anywhere. Xavier samples will be available the fourth quarter of 2017 to automakers, tier 1 suppliers, startups and research institutions who are developing self-driving cars.

Qualcomm

Qualcomm introduced its VR reference platform, the Snapdragon VR820. Developed with Goertek, an ODM. The VR820 includes integrated eye tracking with two cameras, dual front facing cameras for 6DOF and see-through applications, four microphones, gyro, accelerometer, and magnetometer sensors. The reference platform combines Goertek's system design including mechanical, electrical, optical, acoustic, and firmware development with computing, graphics, sensor fusion, audio, connectivity, power management, VR software and system capabilities of the Snapdragon 820.

Adjacent developments

GlobalFoundries will skip development of the 10nm process, and jump to 7nm using resources from its acquisition of the IBM NY fab. The company currently operates a 14nm FinFET node. 7nm FinFET technology is expected to deliver more than twice the logic density and a 30 percent performance boost compared to today's 16/14nm foundry FinFET offerings. The technology is expected to be ready for customer product design starts in the second half of 2017, with ramp to risk production in early 2018.

Fourth quarter 2016

The fourth quarter is when companies try to stir up interest and enthusiasm for their products to get the consumers to buy them for holiday gifts.

AMD

Microsoft is planning to release the new Xbox Project Scorpio in 2017, based on AMD APU.

AMD announced its GPUs would be used by Google in its cloud platform.

Company introduces its GPU-compute line, Radeon Instinct MI6 based on the company's Polaris architecture. It will offer 5.7 teraflops of performance and 224GBps of memory bandwidth.

Major software libraries, drivers, and tools introduced software update called Crimson ReLive,

Sony announced its PS4 Pro VR console, based on AMD APU.

Nvidia

The company filled out its product line and introduced mainstream AIBs based on the Pascal GP107 GPU. The 2GB GTX 1050 (640 cores, 75W) sells for $109, and the 4GB GTX 1050TI (768 cores, 75W) sells for $139.

Board partner Aitec Defense Systems, announced a ruggedized GPGPU sub-system computer, the C535 Typhoon, is based on Nvidia's Jetson TX1 SoM with 256 CUDA cores Maxwell GPU. The board also features 4 GB of LPDDR4 RAM.

Nvidia entered the Top 500 with its super computer DGX-1, while Oak Ridge National Laboratory's "Summit" and Lawrence Livermore National Laboratory's "Sierra" will use Tesla GPU accelerators in the next-generation IBM Power servers.

Nintendo announces new Switch console based on Nvidia Tegra

Qualcomm

Qualcomm announces three new mid-range Snapdragon chipsets: 653, 626, and 427. The 653, 626, and 427 all receive a new X9 LTE modem and support for Quick Charge 3.0 and dual cameras.

Think Silicon

Think Silicon announced the NEMA GFX API designed to accelerate high quality GUI development for embedded and wearable devices.

Adjacent developments

Semiconductors

First transistor with a working 1-nanometer gate

The breakthrough was achieved by Lawrence Berkeley National Laboratory (Berkeley Lab) scientists, by creating a 2D FET using molybdenum disulfide (MoS2) instead of silicon and a 1D single-walled carbon nanotube (SWCNT) as a gate electrode, instead of various metals. (SWCNTs are hollow cylindrical tubes with diameters as small as 1 nanometer.)

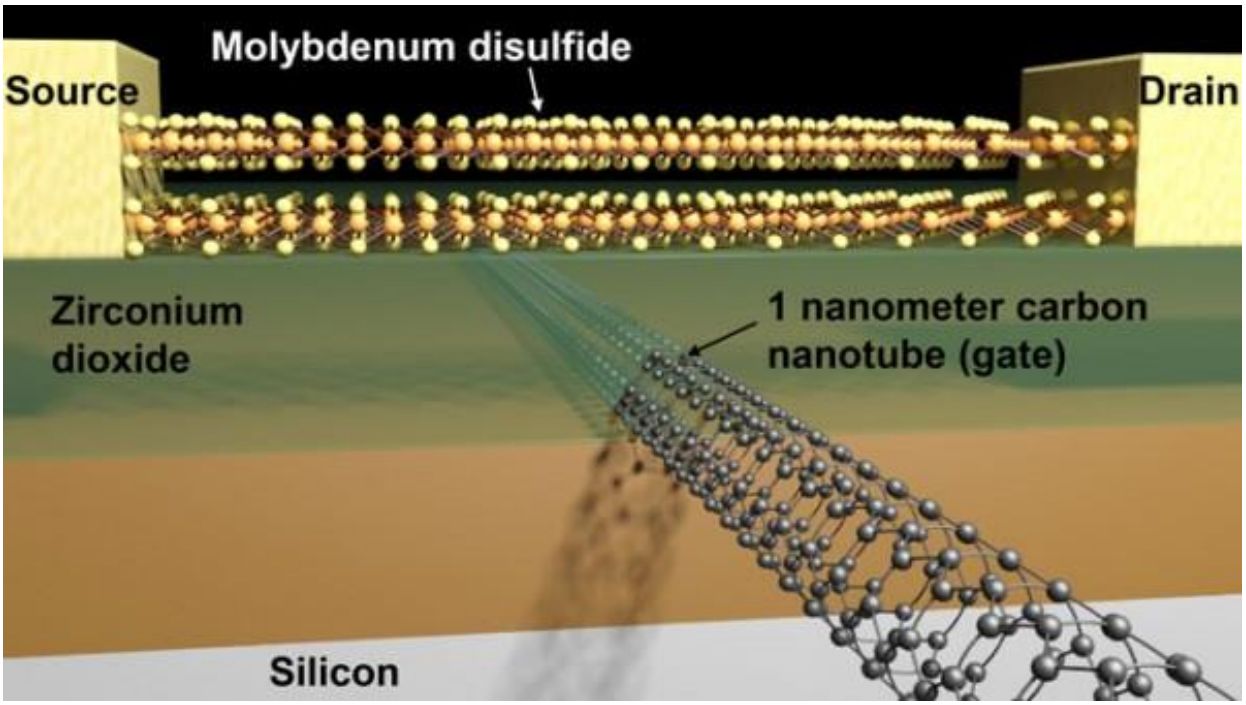

Figure 8: Schematic of a transistor with molybdenum disulfide semiconductor and 1-nanometer carbon nanotube gate. (credit: Sujay Desai/Berkeley Lab)

Compared with MoS2, electrons flowing through silicon are lighter and encounter less resistance . But with a gate length below 5 nanometers in length, a quantum mechanical phenomenon called tunneling kicks in, and the gate barrier is no longer able to keep the electrons from barging through from the source to the drain terminals, so the transistor cannot be turned off.

Samsung starts production at 10nm

Samsung's new 10nm FinFET process (10LPE) adopts an advanced 3D transistor structure with additional enhancements in both process technology and design enablement compared to its 14nm predecessor, allowing up to 30-percent increase in area efficiency with 27-percent higher performance or 40-percent lower power consumption.

Intel and TSMC are expected to start producing out volume-levels of 10nm FinFET chips in early 2017. Intel's 10nm Cannonlake x86 processors are due to arrive next year as are 10nm ARM-compatible CPUs fabricated by Intel.

SUMMARY

GPUs represent such a phenomenal performance to any parameter (price, power, size, etc.) that they are under constant development and improvement, as well as expansion. Originally developed for workstations, when applied to gaming, the volumes increased dramatically and costs were reduced, creating a price elasticity that further fueled their adoption, scaling, and application expansion.

The trends for GPUs will be basically more of the same, as they get employed in supercomputer, mobile phones, and game machines.

Index

6DOF, 18
7th generation core processors, 17
Alibaba, 17
Apical, 13
Berkeley Lab, 20
Bifrost, 13
Cannonlake, 20
Carrizo APU, 12
Denver ARM V8 CPU, 11
DGX-1, 19
Drive PX2, 11
Eye tracking, 18
FinFET, 18
Frame Buffer Compression, 10
Geomerics, 12
GlobalFoundries 10nm, 18
GlobalFoundries 14nm, 9
Google cloud platform., 19
GP107, 19
GTX 1050, 19
GTX 1050TI, 19
GTX 1060, 17
GTX 1070, 14
GTX 1080, 14
HBM2, 7, 16
HEVC 10-bit, 17
Hotchips, 16
IGPs, 12
Intel 10nm, 20
Intel 6th Gen Core vPro, 11
Kryo quad-core CPU, 11
Mali-DP650, 10
Mali-G71, 13
Microsoft, 19
Midgard, 13
Moore's law, 5, 20
NEMA GFX API, 19
Nintendo Switch, 19
Nvidia Xavie, 17
Oak Ridge National Laboratory, 19
Pascal GPU, 14
Polaris GPU, 9
Polaris-based workstation series, 17
PowerVR Graphics SDK v4.1, 15
PowerVR Wizard GR6500, 10
PS4 Pro VR, 19
Qualcomm SD602, 11
Quasar GPU compute, 12
Radeon E9260 & E955, 17
Radeon Instinct MI6, 19
Radeon Pro Solid State Graphics (SSG, 17
Ray tracing, 10
ROPs, 6
RX 460, 17
RX 470, 17
RX 480, 16
Samsung 10nm FinFET, 20
Scalability, 7
Sierra, 19
SIMD, 6
Snapdragon 427, 19
Snapdragon 626, 19
Snapdragon 653, 19
Snapdragon 821, 11
SoftBank, 14
Summit, 19
Tegra Parker architectur, 11
TFLOPS, 9
Thunderbolt, 10
Top 500, 19
TSMC 7nm FinFET, 12
Vega GPU, 16
Virtualized GPU, 10
Volta GPU, 17
VR reference platform, 18
Vulkan, 11, 15
Xbox Project Scorpio, 19