Title: Will GPU replace CPU in the future or the combination of them should be the final trend

Name: Ruoyu Jia

Student ID: G23371040

Abstract & Introduction:

In traditional SoC design, CPU takes up a large space, but along with the rapid enhancement of processing capacity in modern application, CPU occupies less and less, on the contrary, GPU takes up a larger space. It is primarily because of the strong capability of parallel processing of GPU, more and more calculations depend on GPU. In the future, the boundary of the CPU and GPU will may become very vague, under the support of specific API and development platform, GPU may replace the CPU to process heavy parallel and floating-point calculation task. Meanwhile, since in the personal computer area, CPU and GPU always supplement each other, when both of them meet bottleneck, combination may be a better choice.

Keywords:

GPU, CPU, Replacement, Combination

Background

From 1999 when NVIDIA firstly came up with the concept of GPU, the high floating-point calculation capability of GPU triggered a lot of topics, like GPU will replace CPU. Based on the distinctiveness of GPU, researchers are doing the relevant researches in order to adequately take advantage of the high calculation speed of GPU. Along with the update of computer graphics hardware, that is the update of computer graphics processors. Applications of graphics processor have already been the hot topic.

Processor industry is having a big change that CPU and GPU compensate each other and are going to merge. Although in a certain period of time they cannot replace each other, but merging has already been the irreversible trend.

Content Table

1. Can GPU replace CPU

The development trend of CPU is to constantly integrate more functions and modules, from coprocessor to cache and finally to memory controller and even the whole north bridge. Currently, all mainstream CPU of AMD and Intel have already integrated memory controller, the most up-to-date Lynnfield (Core i7 8XX and i5 7XX) of Intel has already integrated the whole north bridge included PCIE controller, while Clarkdale (Core i5 6XX and i3 5XX) also integrated GPU.

1.1 The development trend of GPU: constantly encroach on functions of CPU

As for GPU, to some extent, it itself is a coprocessor, which mainly used for image, video, 3D acceleration. The reason why it still not be replaced by CPU over the years is because GPU is too complicated, because of the restriction of current manufacturing technique, CPU cannot integrate GPU, because the scale of GPU is much larger than CPU, CPU at most can integrate a mainstream mid-range GPU, but this kind of product can only be entry-level, which cannot satisfy the need of game players and high performance calculation.

From the birth of GPU, GPU has been continuously encroaching functions that originally belong to CPU, or in other words, it helps CPU alleviate burdens, to handle those tasks that CPU is not good at. Such as the T&L (Transforming and Lighting) from the very beginning, VCD\DVD\HD\BD video decoding, physical acceleration, geometry shader. And from now and future, GPU will take away one of the most

important functions of CPU - parallel computing, high precision floating point arithmetic.

Use GPU to process non-graphics calculation. Multi-core processor has already been the mainstream in the industry. However, the isomorphism processor of quad core is uncertain to perform 4 times of the performance. Use the strongest functional Core i7 processor of Intel as an example, the architecture of it is totally different from the previous generation Core 2 Quad: it imports three-level buffer, high speed QPI (QuickPath Interconnect) bus, three-channel DDR3 memory controller, Hyper-Threading Technology and a lot of optimization of core and instruction set; however, the test shows that the combined effect of these technologies is: there are only 20% overall performance enhancement of i7 965 compared with its previous generation QX9770, it is really hard to have that tremendous enhancement from signal-core Pentium D to dual-core Core 2 Duo. Furthermore, compare Phenom II processor of AMD with Phenom processor, the enhancement of performance is mainly because of the high frequency by technology, the optimization of core architecture only contributes 5%.

Lately, a simulation test results from America Sandia national laboratory showed that: as for supercomputer, because of the limitation of storage mechanism and memory bandwidth, the 16-core, 32-core and even 64-core after 8-core processor may cause largely reduction of efficiency. This means that the core number of multi-core

processor cannot increase without restriction.

In this circumstance, GPU suddenly rises. In the past, when processing graphics rendering, according to indicate the numbers of triangle in 3D or the difference of texture clarity during coloring, the load processed by each level will have changes. In the traditional architecture, since the number of calculation unit processed by each level is determined in advance, therefore, when load has changed, fixed calculation unit number will be the bottleneck to hinder the enhancement of the whole processing capability of the system.

In recent years, researchers did deep research of graphic instruction structure. They found the proportion of scalar data flow was increasing year by year, if the industry still insists the design of SIMD (Single Instruction Multiple Data) will decrease the efficiency. In the early years of GPU, the crucial calculation specially for graphic processing separated processing unit into vertex shader, rasterization engine, texture map unit and other different parts, these parts accomplished different calculation tasks separately. While the new generation GPU provided an uniform rendering structure, uniform calculation unit replaced the different units above-mentioned. This structure integrated multiple multistage processing calculation units that support vertex coordinate calculation and triangle coloring, tasks for every calculation unit can adjust based on the load of each level. This kind of uniform calculation unit is called uniform scalar shaders, which is also called stream processor. Each stream processor

only achieves operations of the one-dimensional scalar.

For this purpose, researchers did revolutions to GPU: stream processor will not specially be designed for vector, but change to be scalar ALU (Arithmetic Logical Unit) unit. That is to say, it will completely disassemble the arithmetic unit ALU inside the Shader unit of GPU, and design them to be independent stream processors and assign corresponding instruction transmitting terminal and control unit. This type of structure can guarantee the highest execution efficiency no matter what kind of command (include combined command) will face. Which means this structure not only has a strong graphic processing capability, it can process non-graphic calculation commands.

The appearance of the new structure leaded the research that regarded floating-point calculation of general processing as center to use GPU. Theoretically, all floating-point calculation instructions can be delivered to GPU to process. In order to transform GPU to be a real general processor, researchers largely increased the number of flow processor and optimized and improved the kernel structure at the same time, which can make them be more capable to process parallel data on a very large scale.

1.2 GPU has a promising future: the future of floating-point calculation

As mentioned above, in the traditional architecture, since the number of calculation

unit processed by each level is determined in advance, therefore, when load has change, fixed calculation unit number will be the bottleneck to hinder the enhancement of the whole processing capability of system. The appearance of uniform rendering structure made it possible that the general processing regarded floating-point calculation as the center can also use GPU. After adopting this kind of architecture, GPU calculation unit can read instruction and data each time when it is processing, which enhanced its universality. Hence, it can also say that GPU enhance the execution speed of floating-point calculation instruction of computer.

As for floating-point calculation, GPU uses specialized arithmetic unit, so it can process parallel processing at high speed, thus improving the computing speed. Use Tokyo Institute of Technology as example, in October, 2008, they adopted 170 C1070 processors to make the system comprehensive operating speed of their supercomputer T SUBAME jump from 67 millions times per second to 77 millions times. Because of this, people began to call GPU as acceleration processor.

It is known that the first one CPU integrates is coprocessor which is specially used for accelerating floating-point calculation, hereafter all generations of SSE instruction set are all in order to strengthen the floating-point calculation performance of SIMD (Single Instruction Multiple Data) of CPU. But GPU is designed to be SIMD architecture from the very beginning (until now Cypress is still this type of architecture), GPU is a really strong capable processor in floating-point calculation.

At present, the floating-point calculation capability of GPU is also dozens even hundreds of times than multi-core CPU.

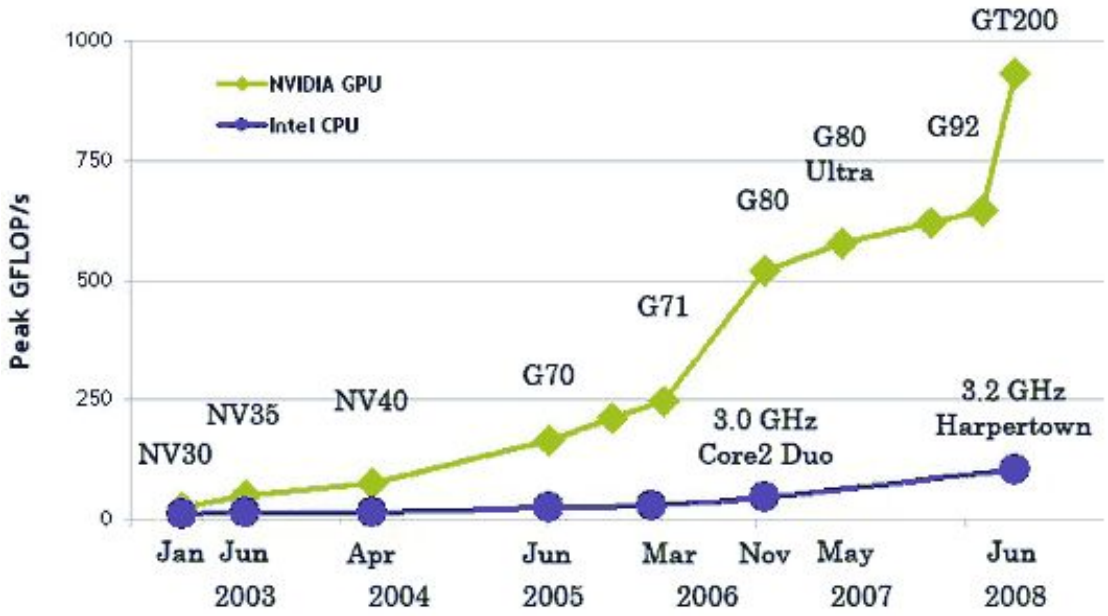

Figure 1 The comparison of NVIDIA GPU and Intel CPU

The floating-point calculation capability of CPU and GPU

CPU can never catch up with the development speed of GPU, hence, it is obviously that GPU is the most proper one to operate floating-point calculation, it makes no sense of CPU to continue increasing core number, thus the whole industry is trying every way to dig potential of GPU and transferring all parallel computing tasks to GPU. Even if Intel also sees the wide prospect of GPU and starts researching and developing GPU.

Previously, because of the restriction of API and software, it is very difficult for GPU to develop in application of parallel calculation field, although the promotion of CUDA by NVIDIA itself got some achievements, there is still not enough. Fortunately,

the two releases API OpenCL and DirectCompute made the future of parallel calculation be suddenly enlightened, and at that time ATI and NVIDIA stood on the same starting line again, so it is clear that who can get stronger performance and wider application depends on whose architecture is more proper for parallel calculation, From the analysis of this paper, the architecture of ATI is still concentrating on traditional graphics rendering which is not suitable for parallel calculation; while the architecture of NVIDIA can completely optimize design directly through general computation API and instruction set, which guarantees to perform the maximum efficacy close to the theoretical value, and provides the strongest floating-point calculation performance.

However, the problem GPU is facing nowadays is that it can only read from its own storage, it cannot read from the main storage of the computer. GPU will copy the data it needs to storage that specially used for itself, and call functions executing in GPU; hereafter, GPU will use multiple cores to process in parallel towards independent data based on the instruction from processor; finally, processor will get the result from the storage used by GPU. Just because of this, in some situations, GPU cannot adequately perform its advantage of extremely high speed in floating-point calculation.

As mentioned above, GPU currently can be seen as a multifunctional parallel calculation processor. Some experts expect that personal computers all over the world will adopt GPU to compute by the end of 2010.

1.3 CPU is facing inflextion: reinforce integer performance, leave floating-point calculation to GPU

AMD possesses CPU and GPU at the same time, also, in the aspect of technology, AMD always can lead the industry, therefore, the future development plan of it is really deserved to be considered. According to the newest product roadmap of AMD, the highlight of the core of the next generation high-end CPU Bulldozer is that every core has double integer computing units, the design of integer and floating-point is asymmetrical.

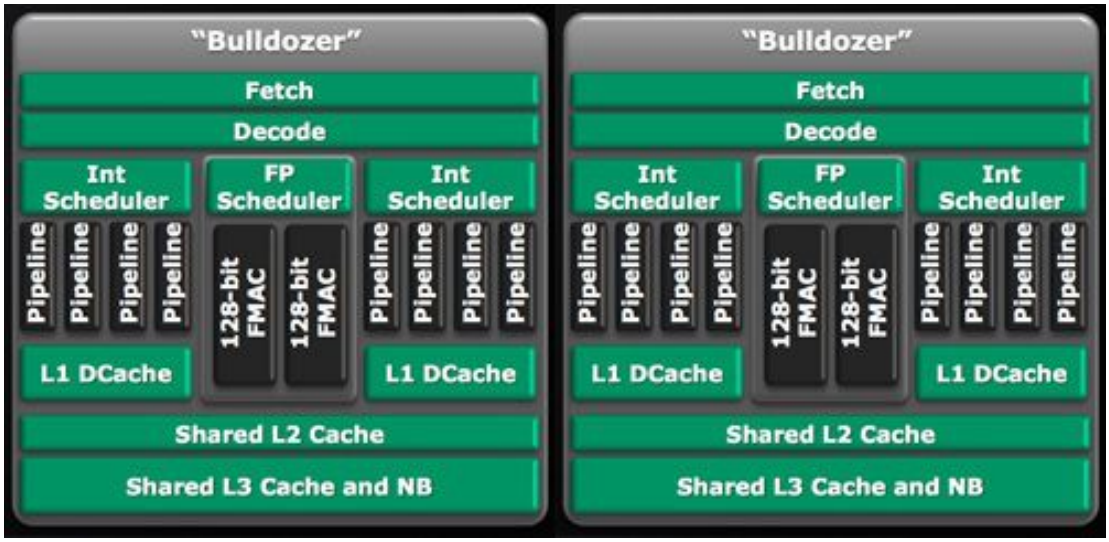

Figure 2 The design of Bulldozer

The next generation Bulldozer architecture largely reinforce the integer computing unit:

In one Bulldozer module, there are two independent integer cores, everyone has their own instructions, data cache, which is scheduling/reordering logic unit. In addition, the throughput capacity of anyone of these two integer units is stronger than the existing integer process unit on Phenom II. The core architecture of Intel no matter it

is integer or floating-point, they all adopt uniform scheduler to dispatch instructions. While the architecture of AMD uses independent integer and floating-point scheduler.

Nowadays, 80% of operations in the server are all pure integer operations, moreover, along with the development of heterogeneous computing applications of CPU and GPU, GPU will burden more and more operations of floating-point calculation, in the next 3-5 years, it can be predicted that all floating-point calculation will be delivered to GPU, since it is the best processor to operate floating-point calculation, this is also the real target of Bulldozer to strengthen integer computation.

Of course, AMD and Intel will both release products integrated CPU and GPU, but their goals are not to wipe out the graphics card and CPU, but to provide strong floating-point calculation capability to CPU through built-in GPU. However, due to the restriction of the manufacture process, the GPU that is integrated with CPU can only be integrated card and low-end card, they can only satisfy fundamental requirements. So in order to get stronger game performance and parallel calculation performance, design a new generation architecture of GPU products specially for floating-point calculation is the wisest choice.

In order to be more clear that whether GPU can replace CPU or not, we need to know more advantages of GPU and more differences between GPU and CPU.

2. Functions of GPU

Graphic Processing Unit is graphics processor. In 1999, when NVIDIA released GeForce 256 graphics processor chip, it was the first time to raise the concept of GPU. A standard GPU contains 2D unit, 3D unit, video processing unit, FSAA(Full Scene Anti—aliasing) unit and memory management unit, etc. The goal of design is to achieve graphics acceleration, now the main target is to accomplish 3D graphics acceleration, hence, the design optimization is basically around the related operations of 3D graphics acceleration, such as blanking, texture mapping, the position change of graphics coordinate and lighting computing and so on. The development of GPU is rapid these years, in 2007, NVIDIA even came up with the argument that "emphasize on GPU core, pay less attention on CPU frequency", is that means the era of GPU will come?

3. The comparison of GPU and CPU

We need to discuss the advantages and disadvantages of both CPU and GPU in the speed of operating calculation and efficiency.

GPU attracts people by its high speed of floating-point calculation capability. How huge is the computing capability? The floating-point calculation capability of CPU is generally under 10 Gflops(which means it can process 10 hundred billions floating-point calculations per second), while the calculation capability of GeForce6 series is around 40 Gflops, GeForce7950GX2 is 384 Gflops; in the field of vector

computing, the calculation efficiency of GPU is 10 times faster than CPU. It benefits by reason that it is customized to process graphics. Parallel computing of GPU is stronger, it has rapid storage system in the internal, NVIDIA 8800 has 128 processors, moreover, the hardware design of GPU can manage thousands of parallel threads, GPU creates and manages all of these thousands of threads without the need of developers to do any coding and management. However, GPU is still giving assistance to CPU to process graphics nowadays,it indeed wastes plenty of operational capability of GPU.

CPU and GPU are chips that both capable to compute, CPU is more like a “generalist” -- contribute more on instruction operation(execution) plus numeric calculation, while GPU is more like a specialist -- graphical numeric calculation is the core. The speed of different types of calculations determined their capabilities -- “be good or not good at those areas”. The speed of chips is mainly determined by three fields: micro-architecture, main frequency and IPC(Instruction per Clock).

What we need to mention is, GPU has such strong computing capability that has pertinence to graphics calculation. These types of calculation are all aim at parallel data, the calculation data is huge, but the calculation types are not complex and also have homophyly, the computing capability is strong but logic is simple, such as matrix manipulation, it is the typical characteristic of graphics operation. But CPU is designed to process, manage and calculate universal tasks, and control the core of the

system, the micro-architecture of CPU is optimized to efficiently operate calculations that have low data relevance and complicated non-calculation class. Therefore, nowadays CPU and GPU perform their own duties in their own orbits.

a. Micro-architecture

The special hardware architecture of GPU highlights its advantages towards CPU: GPU has high bandwidth independent graphics card; high performance of floating-point calculation; high capability of geometric processing. GPU is suitable to process parallel computing tasks; suitable to do repeated calculation; suitable to process image and video tasks. GPU can sharply reduce system cost.

From the aspect of micro-architecture, CPU and GPU are completely designed by different thoughts, micro-architecture of modern CPU is designed by the thought that giving consideration to both “execute parallel instructions” and “compute parallel data”, which means it needs take account of parallelism, universality and balance of program execution and data operation. Micro-architecture of CPU focuses on the efficiency of program execution, it will not pursue extreme fast operation speed but sacrifice the efficiency of program execution.

The micro-architecture of CPU is designed to achieve the high efficiency of the instruction oriented execution, thus, CPU is the most complicated chip in computer design. Compared with GPU, the core of CPU does not have so many repeated design

parts, this complexity cannot be measured only by the number of transistor, this complexity is from accomplishment: such as prediction of program branches, predict execution, execution of multinest branch, instruction and data correlation when processing parallel execution, data consistency and other complex logic when multi-core are processing together.

GPU is actually a set of graphical functions implemented by hardware, these functions are mainly using on drawing calculations needed by various graphics. These calculations that relevant to pixel, light-shadow treatment, 3D coordinate transformation are implemented by GPU hardware acceleration. The characteristic of graphic computing is large amount of dense computing in the same type -- such as matrix manipulation of graphic data, the micro-architecture of GPU is designed to calculate data that is suited to matrix type, there are plenty of repeated designed computing units, this type of computing can be separated into a lot of independent numeric calculations -- a large number of threads of numeric calculation, and there is no logical relevance between data like program execution.

The complexity of micro-architecture of GPU is not high, although there are many transistors. From the aspect of application, the main work of how to well use the parallel computing capability of GPU is to well develop its driver program. Strengths and weaknesses of GPU driver program largely influence the performance of GPU in practice.

Hence, from the aspect of micro-architecture, CPU is good at operating system, system software and general application these kinds of programming tasks that have complex instruction dispatch, circulation, branch, logic judgement and execution. The advantage of parallel of it is on the execution layer, the complexity of program logic also limits the parallelism of instruction when executing the program, the thread of hundreds of parallel programs basically cannot be seen. GPU is good at high parallel numerical calculation that is belonging to graphical or non-graphic class, GPU can contain thousands of numerical calculation threads that have no logical relationships between each other, the advantage of it is the parallel calculation between non-logical relationship data.

b. main frequency

In addition, the speed of GPU executing every numerical calculation is not faster than CPU, this can be seen from the main frequency of modern CPU and GPU currently, the main frequency of CPU is beyond 1GHz, 2GHz, even 3GHz, while the main frequency of GPU is less than 1GHz, the mainstream of GPU main frequency is between 500 and 600 MHz. It is known that 1GHz = 1000MHz. Therefore, when executing a few numbers of threads, GPU cannot exceed CPU.

At present, the main numerical calculation advantage of GPU is floating-point, the high speed of it when executing floating-point calculation depends on a huge amount

of parallel, but the parallelism of this numerical calculation is utterly useless when executing program logic.

c. IPC (Instruction per Clock)

In this aspect, CPU and GPU cannot be compared, because most of the instructions of GPU are numerical calculation oriented, few of the control instructions also cannot be used by operating system and software directly. If comparing IPC of data of control instruction, it is obvious that CPU is much more higher. The reason is simple, CPU emphasizes on parallelism when executing instructions.

Furthermore, there are some GPU can support quite complicated control instructions, such as conditional jump, branch, circulation and subroutine call, but the increment of GPU in the aspect of program control and the capability needed for supporting operating system are still far more uncompetitive than CPU, and the efficiency of executing instruction also cannot be mentioned in the same breath as CPU.

Finally, we can conclude that:

CPU is good at: operating system, system software, application, general computation, system control; artificial intelligence in game, physical simulation; 3D modeling - ray tracing rendering; virtualization technology - abstract hardware, operate multiple operating systems or multiple replicas of one operating system at the same time, etc.

GPU is good at: graphic class matrix operation, non-graphic class parallel numerical calculation, high-end 3D game.

All sum up, in a balance computing computer system, CPU and GPU are still perform their own duties, except graphic computing, GPU may mainly focus on high performance parallel numerical computing with high efficiency and low cost in the future, and help CPU share this kind of computation, enhance performance of system in this area. While typical application of GPU nowadays is still high-end 3D games, an efficient GPU cooperates and an efficient CPU can guarantee the overall efficiency of 3D games. “high-end 3D games only need high-end graphics card” or “high-end 3D games only need CPU” are both groundless statements.

4. The current situation of GPU

GPU triggered the revolution of computer visualization. Processor tycoon Intel felt the strong impact of GPU, Intel specially developed a set of brand new architecture that is programmable oriented display computing general architecture chip -- Larrabee architecture. The effect it can bring can come out by the speak of Pat Gelsinger in Intel Developer Forum who is the Senior Vice President of Intel. “Programmable oriented display computing general architecture chip is a revolution, it will overturn the graphics card industry which has already lasted over decades, it will not replace graphics card immediately, but 3 and 4 years later, along with the mature release of relevant technologies and products, graphics card industry will die out.”

According to the opinions of Intel, along with the development of programmable oriented display computing general architecture chip, it will gradually replace GPU, graphics card will slowly be replaced by integration, which will have less and less living space as independent hardware.

It is really a waste of GPU if it is only treated as graphics card, hence, NVIDIA released CUDA(Compute Unified Device Architecture), which can make graphics card work on other areas not just on image computing. Also, CUDA redefined the functions of GPU, it was a revolutionary computing architecture and computing idea, which can make GPU be an expert in the application field such as consumption, commercial and technology, solve complicated calculation problems. CUDA can use properties of GPU that is growing more efficiently, effectively utilize the high speed calculation capability of GPU and cooperate CPU to process the high performance general calculation.

GPU and high parallel processor are racing at the same time and rapidly developing forward to grasp a future market.

5. The application of GPU

Research of GPU application is around high floating-point calculation capability, programmability and parallel computing. Up to now, cooperation of GPU and CUDA

is mainly used on commercial high-end computation or supercomputing. Such as Tesla high performance calculation, GPU accelerates Matlab’ s high performance calculation and generates medical image, etc.

Since the assembly instructions of GPU are quite complex, and there are also some incompatible problems among different versions of hardware, some advanced languages (such as GLSL, HLSL) have already been developed by some major manufacturers, which accelerated the application research of GPU programming. From the aspect of system architecture, GPU is highly parallel data flow processor that specially optimizes vector calculation, it includes two kinds of flow process units: multiple instruction multiple data (MIMD) process unit - Vertex Shader and single instruction multiple data (SIMD) process unit - Pixel Shader.

This kind of processor uses data flow as process unit, it can obtain higher efficiency when processing data flow, hence, many researchers begin working on a new research field: GPGPU (General-Purpose Computation on Graphics Processors), besides graphic processing, the main research content of it considers a broader application computation.

The initial design idea of GPU brings problem to this new field. GPU is a processor that is specially designed for graphic processing, it has its own storage cell and has some particularities of the way to store and get data, besides, usually there are only

some game makers that develop GPU, they use nonstandard programming model, programming environment and architecture are mostly been seen as commercial secrets that are not public to researchers. Therefore, materials that can reference are limited, and researchers must research parallel algorithm and how to get the highest performance when operating general computing at the same time. Against the above problems, researchers has already raised GPU general programming model and method, it is no doubt that it has pushed the application in the field of non-graphics.

In recent years, there are plenty of achievements in the aspect of GPGPU, such as algebra calculation and fluid simulation, database operation, spectrum transformation and filtering. There are also some achievements in the area of software programming, advanced drawing language and real-time shading language (the idea of drawing programming originates from RenderMan drawing software designed by Pixar in the early years, this software has been widely used in the production of Hollywood films over the years), OpenGL Shading Language, RTSL (real-time shading language) by Stanford, HLSL (high-level shading language) by Microsoft and Cg by NVIDIA are all have great influence on this field; stream processing programming environment and tools have already developed to expand programming of GPU.

6. The combination of CPU+GPU has obvious advantages

CPU and GPU have their own advantages. Generally, CPU is good at processing irregular data structures, unpredictable access pattern, recursive algorithm,

branch-intensive code and single-threaded program. This type of program has complicated instruction scheduling, circulation, branch, logical judgment and execution and other steps. For example, operating system, word processing, debugging of interactive applications, general computation, system control, virtualization technology and other system softwares and general applications. While GPU is good at regular data structure and predictable access pattern. For instance, application of light and shade treatment, 3D coordinate transformation, oil-gas exploration, financial analysis, medical imaging, finite element, gene analysis, geographical information system and scientific computing, etc.

Although GPU performs excellently on many aspects, over a period of time, CPU and GPU will still develop independently, they could continue playing their roles on the field they are good at separately, while the evolution direction in the future will be that CPU and GPU will complement each other and finally combine together. From the aspect of CPU, in order to enhance processing capability, previously it was multithreading, now is multi-core, the development direction in the future will be many-core. CPU is developing to constantly increasing throughput and enhancing energy efficiency; while from the aspect of GPU, the programmable performance of it is originally the program solidified inside the chip, then it can develop to be local programmable, and finally be fully programmable. In other words, GPU is enhancing the throughput and developing in the direction of general processing at the same time.

From now on, the heterogeneous computing structure of CPU+GPU will lead the development direction of processor, this is also the development direction of the next generation of supercomputer. The guiding ideology of designing CPU+GPU architecture platform at present is: let the resources of CPU to be used more in cache, and let the resources of GPU to be used in data calculation. Put these two together, not only the cost of transmission bandwidth could reduce, CPU and GPU that both have the highest speed in computation of personal computer can help each other. The reason is because: the arithmetic unit of CPU only has several ALU, but GPU has a lot of more ALU than CPU. Moreover, there are relatively more cache in CPU, while the number of cache in GPU is much smaller than CPU. When necessary, CPU can help GPU share some works of software rendering, on the other hand GPU can use mainstream programming language to solve general calculation problems. This also means CPU adds a strong floating-point calculation component, and GPU adds a pixel processing unit.

7. Conclusion

Currently, cooperation of CUDA and GPU does not have a tacit understanding, there still needs some time for them to meet the civilian market, the market of graphic processing is becoming bigger and bigger, which gives a powerful driving force to GPU development. There are some progresses of GPU applied in the general purpose computation field, GPU will gradually appear in the civilian market.

What the final result of GPU and CPU will be? There are influences from numerous factors that we cannot get a solid conclusion. Will GPU like the time of 386, the math co-processor 387 which firstly is used as independent chip finally merged into CPU and extinct, or NVIDIA can really develop a GPU that can process various general purpose computation just like CPU, the era of GPU may truly begin.

So, CPU and GPU cannot replace each other, there is a complementary relationship between them, only CPU and GPU work together to process tasks they are good at separately that the strongest performance of computer can be performed. CPU will integrate GPU, but only for those low-end products; GPU will replace CPU to process floating-point computation, but it still needs CPU to operate operating system and control the whole computer. Only when manufacture technology has already developed to a certain level that CPU and GPU can combine together perfectly, at that time, it is also hard to say that if CPU replaces GPU or GPU replaces CPU, but it might be unimportant that who will be replaced.

Processor industry is having an important change that CPU and GPU complement each other and combine together. Even if in a certain period of time they cannot replace each other, but the combination of them is seems like an irreversible trend. Countless imaginations arise in our head, so it is better for the market to eventually determine this controversial topic.

References:

[1] Jiadong Wu, Bo Hong, "Collocating CPU-only Jobs with GPU-assisted Jobs on GPU-assisted HPC," 2014 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pp. 418-425, 2013 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, 2013

[2] Changmin Lee, Won W. Ro, Jean-Luc Gaudiot, "Cooperative heterogeneous computing for parallel processing on CPU/GPU hybrids," 2012 16th Workshop on Interaction between Compilers and Computer Architectures (INTERACT), pp. 33-40, 2012 16th Workshop on Interaction between Compilers and Computer Architectures (INTERACT), 2012

[3] Sungpack, Hong, Tayo, Oguntebi, Kunle, Olukotun. Efficient Parallel Graph Exploration on Multi-Core CPU and GPU[J]. Parallel Architectures and Compilation Techniques(PACT), 2011, (10): 11-14

[4] Tayler H. Hetherington, Timothy G. Rogers, Lisa Hsu, Mike O'Connor, Tor M. Aamodt, "Characterizing and evaluating a key-value store application on heterogeneous CPU-GPU systems," 2014 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), pp. 88-98, 2012 IEEE International Symposium on Performance Analysis of Systems & Software, 2012

[5] John D. Owens, David Luebke, Naga Govindaraju, Mark Harris,

Jens Krüger, Aaron E. Lefohn, and Tim Purcell. A Survey of General-Purpose Computation

on Graphics Hardware. Computer Graphics Forum, 26(1):80–113, March 2007.

[6] Manish, Arora, Siddhartha, Nath, Subhra, Mazumdar, Scott, Baden, Dean, Tullsen. Redefining the Role of the CPU in the Era of CPU-GPU Integration[J]. Micro, IEEE, 2012, 32(6): 4-16

[7] Thomas B. Jablin, Prakash Prabhu, James A. Jablin, Nick P. Johnson, Stephen R. Beard, and David I. August. 2011. Automatic CPU-GPU communication management and optimization. SIGPLAN Not. 46, 6 (June 2011), 142-151.