

Bachelor's thesis
Information Technology
Hardware Design
2015

Sheng Huang

CURRENT GRAPHICS CARD ANALYSIS & NEW DESIGN CONCEPT



TURUN AMMATTIKORKEAKOULU
TURKU UNIVERSITY OF APPLIED SCIENCES

BACHELOR'S THESIS | ABSTRACT
TURKU UNIVERSITY OF APPLIED SCIENCES

Information Technology | Hardware Design

May 2015 | 43

Instructor: Patric Granholm

Sheng Huang

CURRENT GRAPHICS CARD ANALYSIS & NEW DESIGN CONCEPT

Dedicated graphics cards play an essential role in many areas in which the performance of the graphics card always directly reflects the efficiency of the tasks. At the same time, in daily life, people's computer experiences are affected by graphics card to a very large extent.

This thesis aims at analyzing the current architecture designs of the graphics card and a new design concept which will enable possible and significant improvements compared to the old designs, according to the market structure and technology innovation.

KEYWORDS:

Hardware design, Hardware analysis, Maxwell architecture, SLI implement, Power dissipation optimization, Quantization concept, Market friendly

CONTENTS

1 INTRODUCTION TO GRAPHICS CARD	1
1.1 Dedicated graphics card	1
1.2 Supplier brief introduction	2
2 PERFORMANCE JUDGMENT & ARCHITECTURE ANALYSIS	3
2.1 Performance judgement	3
2.2 Architecture design of graphics card	5
3 THE WAR OF TWO GIANTS	9
3.1 Competition	9
3.2 Design strategy	11
4 TWO GRAPHICS CARD = ONE + ONE?	13
4.1 Graphics card paralleling technology	13
4.2 SLI status	14
5 DISADVANTAGES & PROBLEMS	15
5.1 Situation analysis	16
5.1.1 Performance waste	16
5.1.2 The neglected SLI technology	17
5.2 Situation results	18
6 NEW HOPE--- MAXWELL ARCHITECTURE	18
6.1 Maxwell architecture	18
6.2 Performance analysis	19
6.2.1 Statistics analysis	19
6.2.2 Theory test & practical test analysis	20
6.2.3 Power dissipation & core area	21
6.2.4 The SLI efficiency of Maxwell Architecture	24
7 NEW DESIGN CONCEPT: QUANTIZATION GRAPHICS CARD	26
7.1 Q-card	27
7.2 Q-board	30
7.3 Q-controller	32
7.4 Upgrades & Adjustments	33
7.5 The PCB issue	34
7.6 Future improvements & influence	35
REFERENCES	37

PICTURES

Picture 1. The result of GTX780 on 3DMark fire EX	4
Picture 2. The result of GTX780 on 3DMark 11 X mode	4
Picture 3. The result of GTX780 on 3DMark vantage X mode	5
Picture 4. The chip block of GK110	6
Picture 5. The chip block of GK104	7
Picture 6. The Streaming Multiprocessor Architecture of GK110	8
Picture 7. The Streaming Multiprocessor Architecture of GK104	9
Picture 8. The 3DMark 11 test results of GTX680 and HD7970	11
Picture 9. Comparison between Kepler control logic and Maxwell control logic	22
Picture 10. The full chip block diagram of GM204	23
Picture 11. The theoretical test about the GTX980 SLI	24
Picture 12. SLI theory test result in extreme situation.	25
Picture 13. Different number of SLI test results.	26
Picture 14. Brief structure of Q-card	28
Picture 15. Brief structure of Q-board	31

TABLES

Table 1. Comparison GTX660TI, GTX670 and GTX680	12
Table 2. The main statistics of GTX650TI, GTX750TI and GTX750	19
Table 3. Estimate statistics of Q-card	27

1. INTRODUCTION TO GRAPHICS CARD

1.1 Dedicated graphics card

A graphics card, as one of the most basic and important part of a computer, handles the task of transferring data into graph as well as output images onto the screen. Meanwhile, the graphics card can also support the CPU to improve the operating speed of the whole computer.

Graphics cards are divided into two different types: dedicated graphics cards and integrated graphics cards. In 1981, Intel, IEEE and EISA cooperated to develop ISA Bus, named Industry Standard Architecture Bus, and on this basis, the first dedicated graphics card was invented to achieve higher performance to meet more professional needs (Intel, Intel ISA Bus Specification and Application Notes, 1989). As the development of computers has been faster and more widespread, the dedicated graphics card has an unbelievably significant role nowadays in almost every areas, although the integrated graphics card still remains widely used.

The dedicated graphics card is considered to be more advanced than the integrated graphics card because the dedicated graphics card is an independent hardware attached to the mother board whereas the integrated one is a part of CPU or the mother board. This difference leads to the fact that the dedicated graphics card has much more room for its electronic parts to acquire both better technology and more perfect design. Furthermore, the integrated graphics card uses the RAM rather than the memory of the graphics card to process data which possibly decreases physical RAM can be used by CPU decreased while the interaction between the integrated graphics card and RAM reduces the process speed of CPU. Judging from all these facts listed above, the performance difference between the best representatives from these two types is very sharp, which is also reflected in their prices. The best dedicated graphics card nowadays can reach up to 1000 euros or even higher. Hence, the choice

between the dedicated graphics card and the integrated graphics card can be considered to be the choice between performance and price. This thesis will focus mostly on the dedicated graphics cards.

Dedicated graphics card for civil use are divided into two target areas: high-performance gaming and professional graph design. Both of them have a slightly different architecture from each other because they are focusing on different goals when processing data. Due to the fact that the usage of gaming graphics cards is much wider than the professional graph design ones and the performance difference between them is very small, in this thesis only the former will be discussed.

1.2 Supplier brief introduction

The two largest civil dedicated graphics card supply companies are AMD and NVIDIA and both of them produce gaming graphics cards and graph design graphics cards (Jon Peddie Research, 2014).

NVIDIA and AMD have been competing with each other for a very long time. The competition in the dedicated graphics card area has a significant point and it is the only point: the performance for a certain price. When NVIDIA and AMD provides their own products at similar prices, the performance of them will affect the result in sales, because every customer is willing to buy better products for the same price.


Considering that many types of customers have different needs when they choose the dedicated graphics cards, both NVIDIA and AMD have different graphics card models with different levels of performance from Basic Office use like GT210 and HD6450, to very high performance for high quality games, like GTX 980 and R9 290X. Furthermore, each model has no possibility to be upgraded to another one, in other words, they are all independent of others. The architecture of these models varies according to their performance which sometimes causes chaos for customers because they may have very little knowledge about the graphics cards and this situation can also cause a problem

because customers may buy a model that does not have the performance they need.

2. PERFORMANCE JUDGMENT & ARCHITECTURE ANALYSIS

2.1 Performance judgement

The performance of graphics cards is judged mainly by two methods: one is using professional simulation test software, namely 3DMark, for different tests in multiple high pressure situations, which will give results in points for different features; the other one is using high quality games, for instance Battle Field 4, for practical tests, including the frame per second (FPS) comparison for extreme quality settings and the frames forming speed. These methods mostly have minor differences in the results, although some games may have bad optimization so that the practical result may be worse than the theoretical result. The theoretical result is considered as more important for the final test result. The 3DMark has various versions and the result of each can only compared to the result from the same 3DMark version not with the result from other versions. The graphics card with higher scores has better performance.

GPU	NVIDIA GeForce GTX 780	CPU	Intel Core i7 4770K
Fire Strike 适合高性能游戏电脑 		有效分数 ? 分数 8304	
		显卡分数 9257	
		物理分数 11056	
		显卡测试1 44.22	
		显卡测试2 36.93	
		物理测试 35.10	
		综合测试 18.01	

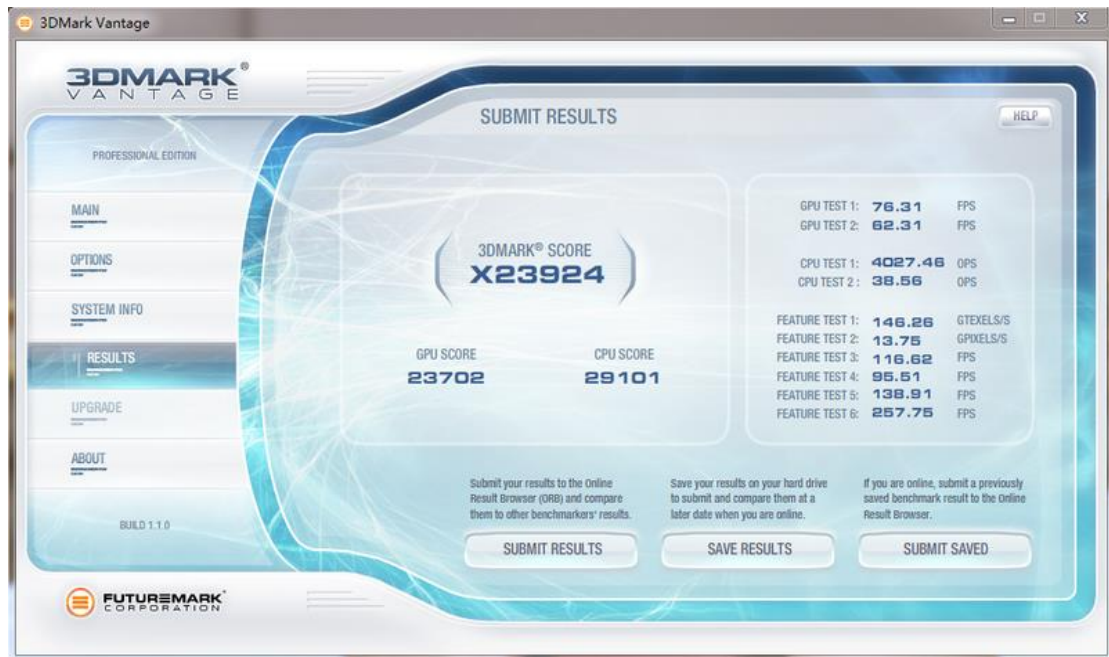
Picture 1. The result of GTX780 on 3DMark fire EX

The 3DMark will provide the results in multiple features, including: the physical result, which is related to the CPU performance, the graphical result, which is related to the graphics card result, and the combination result as well as other important categories, such as double precision float calculation results.

The different version 3DMarks may have different sequences results when comparing two graphics cards at almost the same level because they place more emphasis on different categories.



Picture 2. The result of GTX780 on 3DMark 11 X mode



Picture 3. The result of GTX780 on 3DMark vantage X mode

2.2 Architecture design of graphics card

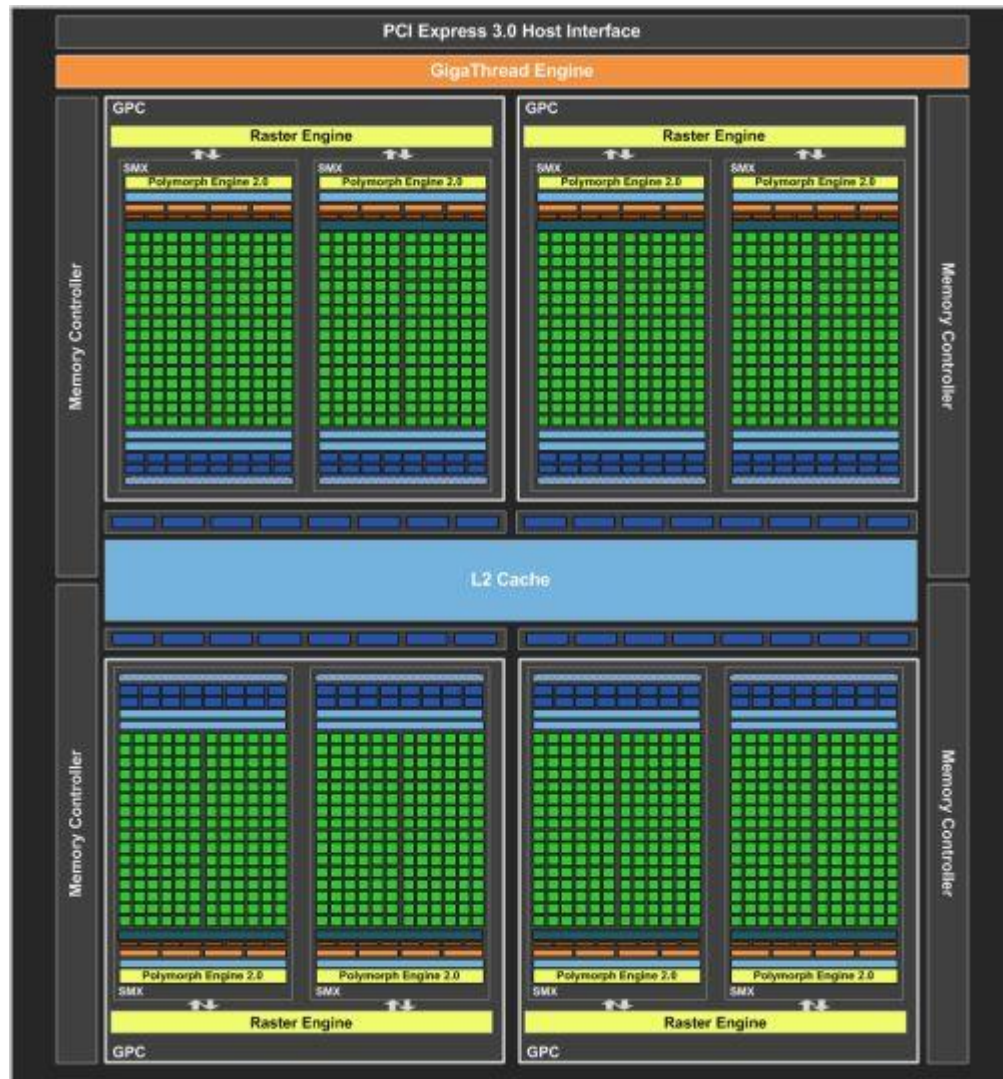
Both NVIDIA and AMD suggest one year as a period of lifetime for the graphics cards of every generation, although sometimes the period lasts longer or shorter. Each generation usually indicates a new architecture of the graphics card. Sometimes the new architecture is very different from the old one, for example the GTX600 series uses the new Kepler architecture instead of the Fermi architecture of the GTX500 series. Nevertheless there exists also the situation that the new architecture is only thoroughly developed from the old architecture, like the GTX700 series using advanced Kepler architecture in the place of the old Kepler architecture of the GTX600 series. Either of the new architecture types has one feature in common: they are much better than the old ones. In addition, every generation has not much connection with each other, for the reason that different architecture reflects totally different performance and features. Even the improved architecture has significant difference from the old version so that the new generation should undoubtedly be better. Another

flexible feature is that even though two graphics cards are the same model, they can have different characteristics due to micro adjustments.



Picture 4. The chip block of GK110 (advanced version Kepler architecture used on GTX780TI) (NVIDIA, 2013)

GTX780TI uses the final version of GK110 which is the most powerful version of the Kepler architecture without any weakening adjustment.



Picture 5. The chip block of GK104 (old version Kepler architecture used on GTX680) (NVIDIA 2012)

Both GTX780TI and GTX680 use Kepler architecture but GTX780TI has an advanced version and both GTX780TI and GTX680 are the flagship graphics cards of their generation, which means that they use the full architecture of their own generation (NVIDIA, 2014).

The difference from the graph (Picture 4 & Picture 5) makes it so obvious that GK110 has 15 Streaming Multiprocessors compared to the 8 of GK104 and this is one of the main reasons why GTX780TI is far better than GTX680. More Streaming Multiprocessors means more efficiency when dealing with high level

work, like high quality games (Weiss, 2012).

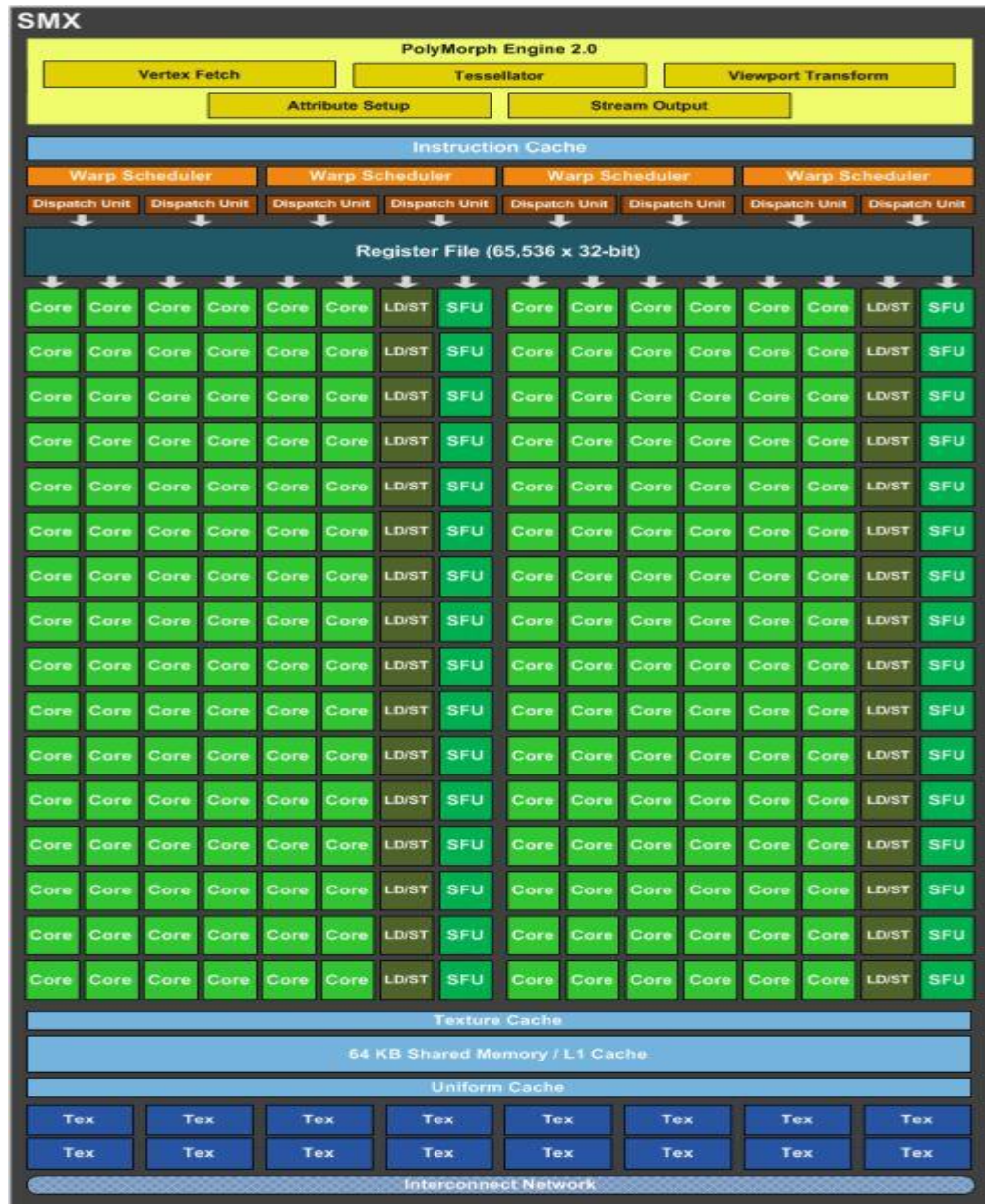


Picture 6. The Streaming Multiprocessor Architecture of GK110 (NVIDIA 2013)

When it comes to the architecture of the Streaming Multiprocessor, the difference between GK110 and GK104 is not huge but important: GK110 has inserted DP units (Double Precision Float) to let GTX780TI have much better double precision float calculation than GTX680, which was infamous when compared to its opponent, HD7970.

The minor changes in the same architecture with different version can create huge difference in the final performance. The changes between different

architectures have much more influence in the performance of graphics cards.



Picture 7. The Streaming Multiprocessor Architecture of GK104 (NVIDIA, 2014)

3. THE WAR OF TWO GIANTS

3.1 Competition

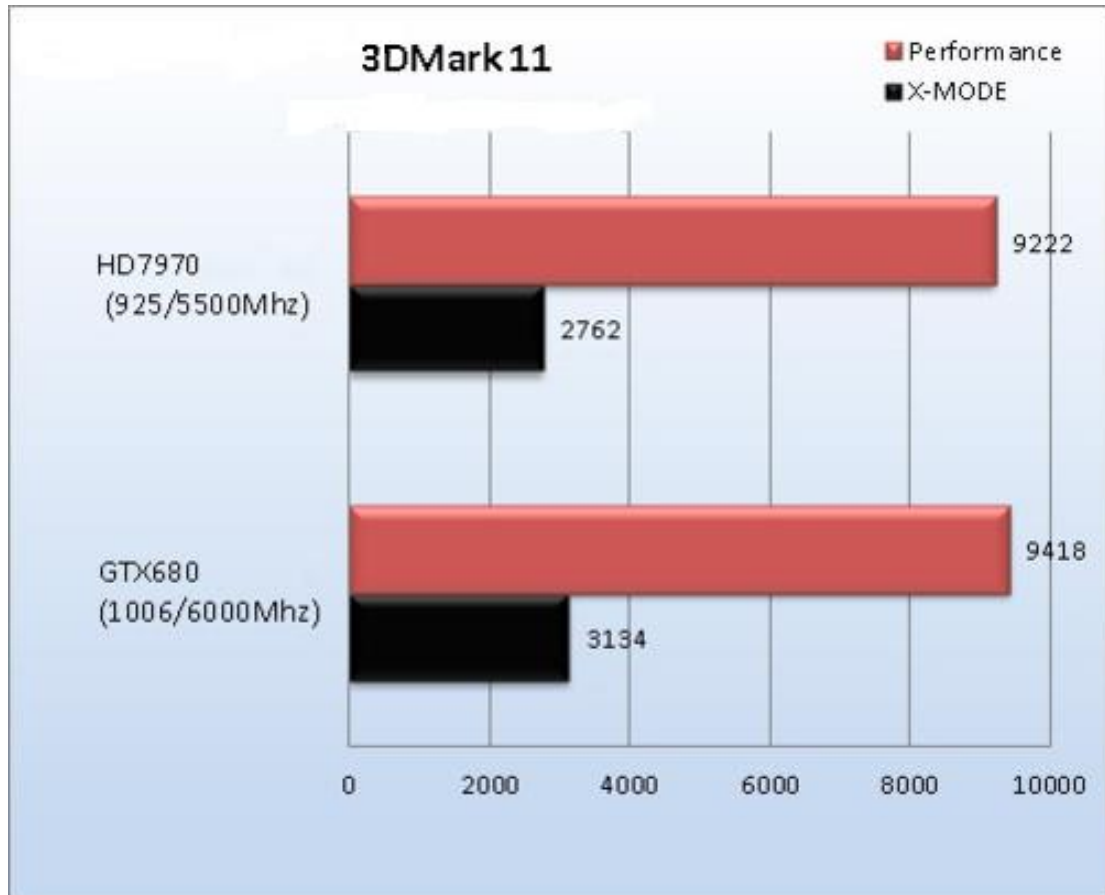
During their long time competition, AMD used to be in the leading position,

especially when the model HD5870 was developed on 23rd, September, 2009. AMD managed to achieve their peak with the brilliant performance of HD5870, but they did not maintain the lead for a long time. After HD5870, NVIDIA managed to narrow down the gap between the performance of AMD and NVIDIA products in the following years and with the birth of GTX690 on 29th, April, 2012, NVIDIA has successfully taken the lead in the competition and manage to maintain the position till now (Jon Peddie Research, 2014).

The competition in graphics cards from these two companies is not as easy as it seems to be although the main point is only about the performance of the graphics cards. The competition is more like a duel with a certain tacit agreement rather than normal business war. Like some products in other areas, graphics cards are divided into generations according to their model names.

The competition in the graphics cards always takes place between the same generations from both NVIDIA and AMD. For instance, in the 2012 generation competition, the GTX600 series using Kepler architecture from NVIDIA has their opponents HD7000 series using Graphics Core Next architecture from AMD.

Each generation from both companies has multiple models according to their different level of performance. The new generation is always claimed with the declaration of the flagship graphics card of itself, like the GTX680 of the GTX600 series and the HD7970 of the HD7000 series. The status of the flagship cards is crucial because for those customers with certain knowledge of graphics cards, the duel between the flagship cards from NVIDIA and AMD to a very high extent reflects the victory or defeat of this generation's competition.



Picture 8. The 3DMark 11 test results of GTX680 and HD7970

From the result we can see GTX680 has a slightly better performance than HD7970.

This phenomenon is logical because the flagship cards, as the term “flagship” suggests, always use the best technology and manufacturing skills in order to reflect the best performance the architecture of this generation can provide. The comparison between the flagship cards usually only considers the performance without the price, the power dissipation and any other factors, even though one has an unignorable higher price or power dissipation than the other.

3.2 Design strategy

The competition of the whole generation is not over by only comparing the flagship cards. The majority of customers has neither enough money nor need to

buy a flagship card, so for the whole market, different level graphics cards with different performance as well as price are necessary. Judging from the market strategy, the flagship cards have more meaning as a symbol, not as a product. However both NVIDIA and AMD are willing to invest huge amounts of money and human resources as well as technology into flagship cards to earn the reputation in such an ethereal way.

When the comparison comes to the graphics cards except for the flagship cards, many factors are taken into consideration, such as the power dissipation, the price, and the outlook although the performance is still the main point.

These graphics cards usually use the same architecture with some difference from the flagship. Judging from the technology and business strategy, both NVIDIA and AMD are using the same method to make the graphics cards with lower performance. This method is to reduce the performance of architecture by crippling certain technology features, like the clock rate, the number of stream processor, the bus width and the memory size.

Table 1. Comparison GTX660TI, GTX670 and GTX680 with the same GK104 architecture

Graphics card	GeForce GTX 660 Ti	GeForce GTX 670	GeForce GTX 680
Fabrication node	28nm	28nm	28nm
Shader processors	1344	1344	1536
Streaming Multiprocessors (SM)	7	7	8
Texture Units	112	112	128
ROP units	24	32	32
Graphics Clock (Core)	915 / 980MHz	915 / 980MHz	1006/1058MHz
Shader Processor Clock	915 / 980MHz	915 / 980MHz	1006/1058MHz
Memory Clock / Data rate MHz	1502 / 6008 MHz	1502 / 6008 MHz	1502 / 6008 MHz
Graphics memory	2048 MB	2048 MB	2048 MB
Memory interface	192-bit	256-bit	256-bit
Memory bandwidth	144 GB/s	192 GB/s	192 GB/s
Power connectors	2x6-pin PEG	2x6-pin PEG	2x6-pin PEG
Max board power (TDP)	150 Watts	170 Watts	170 Watts
Recommended Power supply	450 Watts	500 Watts	550 Watts
GPU Thermal Threshold	98 degrees C	98 degrees C	98 degrees C

Needless to say, graphics cards with weakened architecture have much lower price according to their performance as well as the customers' needs. At each

level, both NVIDIA and AMD have their products to offer to the corresponding customers. In the mid-high level market, providing the most profit for both NVIDIA and AMD, which is targeting normal game players who do not intend to play games with ultra-quality graphic effects, NVIDIA has GTX660TI while AMD has HD7870 with some other models. In the low-level market which constitutes the largest customer group, NVIDIA has GT630 while AMD has HD6670 with some other models.

There is one point that needs to be declared, the weakened architecture cannot be restored to the full version architecture. In other words, although the architectures these graphics cards use are similar, they have no connection with each other. When customers want to upgrade their low performance graphics card, the only way is to desert the old one and buy a new one with better performance.

4. TWO GRAPHICS CARDS = ONE + ONE?

4.1 Graphics card paralleling technology

Concerning the upgrade of performance, there exists another way rather than buying a new graphics card when a customer is using a mid-high level graphics card. Both NVIDIA and AMD have designed graphics cards paralleling for the graphics cards above mid-high level. NVIDIA calls its paralleling technology Scalable Link Interface (SLI) while AMD calls its paralleling technology CrossFire. This thesis mainly focuses on NVIDIA's SLI.

SLI technology is provided from mid-high level graphic card to the flagship card, as an example from the GTX600 series: from GTX660 to GTX680 (NVIDIA, 2013). The main purpose of SLI is to provide a way to increase more performance of graphics cards than the one single card. For the flagship cards, this technology can give those "hardware addicts" a way to achieve an extreme performance meanwhile for mid-high level graphics cards, customers can upgrade their old graphics card to fit new needs. Although this is called graphics

cards paralleling, the amount of performance increase of SLI is not just a $1+1=2$ calculation. The SLI efficiency varies according to the models of graphics cards from 50% to 80% or even higher which means a 2-card SLI's performance may lead to 1.5 to 1.8 times than one card's performance.

4.2 SLI status

SLI technology is not a panacea because it has some restrictions. Firstly, SLI technology only supports paralleling graphics cards of the same model, and this rule has to be strictly followed. For example, a GTX660 card cannot be paralleled with a GTX670 card by using SLI. The reason for this rule is that even minor architecture differences are not tolerated in SLI technology.

Secondly, when two graphics cards of the same model are used in SLI, the one with the higher clock rate will be adjusted to the lower clock rate. The graphics cards need to function under the same rate without exception.

Thirdly, SLI technology nowadays only supports four cards paralleling maximum and the maximum is only provided on flagship cards while other models only have two cards paralleling SLI. This is the most unusual restriction in all rules. The number limit of SLI is not plausible for SLI technology has the ability to afford many more than four cards paralleling. The only reason for this restriction is obvious: multiple graphics cards SLI is not necessary for most customers. Actually, SLI technology is not necessary for most customers because the performance of a single graphics card is so powerful that when customers need to upgrade them to meet their increased needs, the next generation graphics cards with better performance at the same price may already come out.

Here is an example: a GTX670 costs 400 euros when a customer buys it. Its price drops to 300 euros when the new generation GTX770 comes out while GTX770 costs 400 euros. The choice for the customer who needs better performance than one single GTX670 is simple: buying another GTX670 to form SLI or buying a GTX770 and abandon the old GTX670. Most customers would choose the latter even though the performance of SLI of 2 GTX670 is slightly

better than one single GTX770 because most customers prefer a new product. Furthermore, the drives for the graphics cards are always more adapted to the new products.

Judging from the features and restrictions of the SLI technology, SLI is not widely used nowadays though it is an advanced technology.

5. DISADVANTAGES & PROBLEMS

The overall status of graphics cards is very stable judging from all points of view. The market has two large graphics card suppliers, which prevent the possibility of monopoly. Although the price of graphics cards is still much higher than their prime cost, concretely the price is up to 5 times more than the prime cost at least. Fortunately the competition between NVIDIA and AMD has restricted the price rise, especially for AMD currently, for they are losing in performance so they are using a price war strategy.

Due to this competition, the technology updating is also stable but swift. Both NVIDIA and AMD do not want to lose to each other in attracting customers with flagship cards. Although they divide each generation with the same length, there still exists space between the publishing dates of flagship cards. This space provides a window for the competitor of the publishing company to adjust their flagship card according to its rival's statistics to achieve better performance. This situation ensures that during any period of time, one of the flagship cards from NVIDIA and AMD will be better than the other one, so that the other company will devote their best to claim the flagship advantage back. The spiral development in the technology causes that customers can get better experience much faster than before.

As mentioned above, gaming is one of the two most important areas civil graphics targets. Due to this, at least the flagship cards of every generation should be competent to run all games smoothly. Call of Duty: Advanced Warfare, which came out on 4th, November, 2014, as the first new generation computer

game, it has far better hardware requirements than the preceding generation and more games with such high requirements will appear in the very near future and this situation is a great catalyzer to stimulate graphics card development as it used to doing so.

Judging from above, the overall needs for graphics cards will be still steady while accelerating.

5.1 Situation analysis

The overall design of the graphics cards is focusing on the performance of every single graphics card which leads to multiple problems and disadvantages.

5.1.1 Performance waste

GTX780, the flagship card of GTX700 series, has its maximum 900MHz core clock rate when it is fully functioning for high quality games, like Battle Field 4, to fit the needs in multiple features, like anti-aliasing, shades processing and light rendering, but it will only work at the minimum 324MHz core clock rate for Office software.

What is the reason for the difference? When a graphics card is working, it will adjust its core clock rate according to the task it is handling at the moment. This working feature determines the fact that graphics cards are always in a dynamic situation. Furthermore this design was aiming at increasing the life length of the graphics cards. However the fact does not seem to be as good as it should be and in fact the waste of performance is the most extrusive problem for the graphics cards nowadays.

Although the performance as well as the functioning speed of other parts of the graphics card decreases when handling low pressure work load, they still need to function not totally shutting down. So the durability is still worn down at a reduced speed. For example, when a very experienced professor in mathematics is asked to teach basic math to the children in kindergarten, he is surely not using his entire skills, he still needs to spend his energy. In addition,

as for GTX780, even its minimum 324MHz core clock rate is still too high for normal Office software because the graphics card has to maintain the core rate at a certain level in order to increase its performance to the maximum when needed as fast as possible.

From another point, the dynamic core clock rate design surely increases the life length of graphics card to some extents, but it seems so unnecessary. The reason is that the life length of a graphics card is seldom longer than its real usage time. As indicated above, the evolution of graphics card is very fast with each new generation per year so that the average real usage time of graphics cards is around 2-4 years which is much shorter than its life length, around 5-6 years. So the life length of graphics card is not a feature that needs such considerate designs. Thus, even with the huge waste of performance, graphics cards are still having shorter usage time than their life length for most of time. In addition, as discussed above, the updating of graphics card requires buying a new one so that the waste can never be compensated.

This problem is more crucial for the graphics cards with mid-high or high level performance which are mostly bought by game players. For normal game players, although they spend more time than other customers on games so that their graphics cards have more time at full performance level, their computers still have so much time idling or performing low pressure functions like chatting and browsing websites. The performance waste of medium-high and high level graphics cards is more than those of low level models because medium-high and high level graphics cards have more range between their maximum performance and minimum performance.

5.1.2 The neglected SLI technology

Another problem concerns the SLI technology. SLI technology has almost the same length as the graphics cards but it still has so low priority and status among all the features. The main problem is the efficiency of SLI. The current SLI efficiency is around 70%-80% for double graphics cards, and the efficiency will drop when the number of cards increases to 3 or 4. This problem can be

easily solved if NVIDIA devote enough research into the architecture adjusting but they have not done that for so many years. Basically, the SLI technology is not following the design nowadays which is focusing on performance of single graphics card. For non-flagship graphics cards, SLI technology is unnecessary compared to a higher model so that both customers and NVIDIA do not value it much.

5.2 Situation results

The analysis about the current graphics card situation shows that graphics card, as the representative of human technology, has many features that can be improved. The performance waste and the power dissipation are the 2 most serious problems at the moment. These two problems not only cause chaos to the models of graphics cards which are needed to face different level of customers, but also result in huge energy loss in the 21st century although this contradiction is not so obvious yet.

Still, NVIDIA and AMD are not able to solve the problems with the technology before a great breakthrough. However, the improvements in the architecture come more swiftly than expected.

6. NEW HOPE--- MAXWELL ARCHITECTURE

6.1 Maxwell architecture

On 19th, February, 2014, NVIDIA announced the new generation architecture of their graphic card, the Maxwell architecture, while it also began to sell the first two products using the alpha version (still 28nm process) of Maxwell architecture, named GTX750TI and GTX750. The action taken by NVIDIA is a signal to show the world that they have already had enough technology to step into the next generation of graphic card (NVIDIA White Paper, 2013).

The Maxwell architecture is named after the great physicist James Clerk Maxwell just like the former architectures named as Johann's Kepler and Enrica

Fermi. Before the Maxwell architecture was announced, there was a rumor saying the Maxwell architecture will not have so many improvements as the Kepler architecture and that the Maxwell architecture would be a failure, because the Kepler architecture was so successful when compared to the Fermi architecture while the Maxwell architecture may not reach such a goal when compared to Kepler.

Nevertheless the fact has slammed those troublemakers badly after NVIDIA formally announced the Maxwell architecture. The first test about GTX750Ti and GTX750 has given the whole world a great shock: the Maxwell architecture is so powerful.

6.2 Performance analysis

6.2.1 Statistics analysis

NVIDIA changes its plan when a new architecture comes out. Instead of announcing a brand new flagship graphic card, NVIDIA only brought out two medium-end graphic cards. Although we cannot see the full power of Maxwell architecture, we can still see the improvements by comparing GTX750TI with the medium-end graphic card of the old Kepler Generation.

Table 2. The main statistics of GTX650TI, GTX750TI and GTX750 (VideoCardz, 2014)

VIDEOCARDZ.COM GeForce GTX 750 Series Specifications			
	GeForce GTX 650 Ti	GeForce GTX 750 Ti	GeForce GTX 750
GPU	28nm GK106	28nm GM107-400	28nm GM107-300
GPU Config	768 : 64 : 16	640 : 40 : 16	512 : 32 : 16
GPU Clock	928 MHz	1020 / 1085 MHz	1020 / 1085 MHz
Memory Clock	1350 MHz	1350 MHz	1250 MHz
Video Memory	1GB GDDR5	2GB GDDR5	1GB GDDR5
Memory Bus	128-bit	128-bit	128-bit
TDP	110W	60W	55W

The test was made between three graphic cards targeting the medium-end market, namely GTX750TI, GTX750 and GTX650TI among which GTX650TI was using the Kepler architecture. The price of these three graphic cards is

similar to each other, around \$150 while GTX750 is slightly cheaper.

GTX750Ti has 640 CUDA cores with a 1020 MHz base clock which can be boosted to 1085 MHz or even higher for the non-reference design version while GTX750 has 512 CUDA cores with the same base and boost clock as GTX750TI. When it comes to the memory interface, GTX750TI has 2048 MB while GTX750 has 1024 MB with the same 128-bit memory interface width.

The Kepler architecture graphic cards' features seem to be kind of similar or even better than their new architecture brothers. GTX650TI has 768 CUDA cores with 980 MHz base clock and a 1033 MHz boost clock while it also has 2048 MB memory interface with the same 128-bit memory interface width.

Only judging from the features above, if the Maxwell architecture does not have a great amount of improvements, GTX750TI will not be able to take the place of GTX650TI.

6.2.2 Theory test & practical test analysis

The test starts the first stage, testing with the 3DMark Firestrike, GTX750TI has gained 3727 points while GTX750 has gained 3296 points. At this stage, the old medium-end GTX650Ti only receives 2895 points.

After tested in the simulation, the next stage is to test in different games. Those games are specially selected to acquire at least medium-end hardware to play and they are always used to test the performance of a graphic card. Because the optimization of each game is different, the test will normally cover several different games and combine the results together to reach the final performance conclusion. The standard is to compare the average FPS between different graphic cards.

The first game is Metro: Last Light which requires very high level hardware to get all features on maximum, but for the test, it is only necessary to compare the FPS between the candidates. The test circumstance is under 1920x1080 with high quality and 16x AF. During the test, GTX750TI has gained 36.4 averages

FPS while GTX750 has gained 31.5 FPS, meanwhile GTX650TI only receives 27.7 FPS. In addition, human eyes can sense the spikes of the output on screen when the FPS is below 30, so GTX650TI is not able to provide ideal game entertainment (Li JiaSheng, 2014).

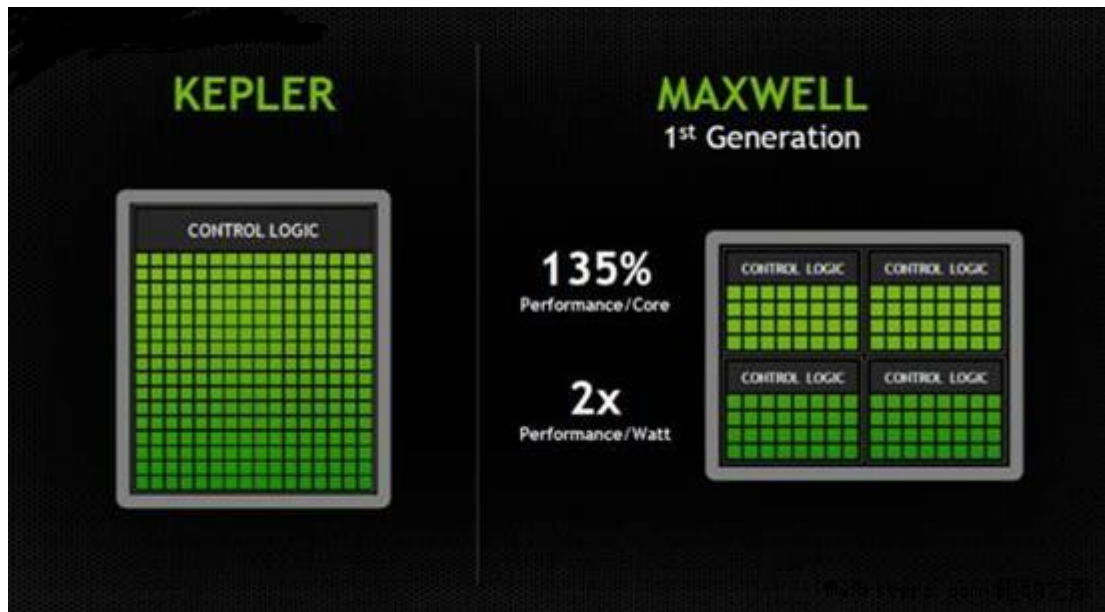
The second game is Crysis 3. The test circumstance is also 1920x1080 with high quality and 16x AF. The global results seem to be better than those of the previous game: GTX750TI has gained 43.4 average FPS while GTX750 has gained 38.5 FPS and GTX650TI only has 35 FPS, still lower than its two new brothers.

In the rest of the games, GTX650Ti also fails to defeat the Maxwell architecture graphic cards. It is very obvious to conclude that GTX750TI and GTX750's performance are significantly improved compared to GTX650TI.

These are two different graphic cards with almost the same features, while the GTX650TI even has better statistics in some features, they both perform so differently due to the Maxwell architecture's greatly improved design.

6.2.3 Power dissipation & core area

Power dissipation has so much difference. Judging from the statistics even GTX750TI has a better performance than GTX650TI. GTX750TI with Maxwell architecture can provide better performance with only half power dissipation of GTX650TI. This is due to the Maxwell architecture design focusing on reducing the power dissipation, and furthermore, GTX750TI does not need extra attached power cable to provide the power it needs.



Picture 9. Comparison between Kepler control logic and Maxwell control logic (NVIDIA, 2014)

Maxwell architecture has changed the original 192 cores per control logic from the Kepler architecture into 4 smaller control logics and each of them only needs to take charge of 32 cores. The performance becomes more different when computer is handling multiple tasks because the pressure on each control logic decreases so much due to the increasing number of it so that the efficiency increases greatly.

In addition, the L1 cache has doubled while the L2 cache of GPC increases to 2M so that the effective usage is increased significantly.



Picture 10. The full chip block diagram of GM204 (NVIDIA, 2014)

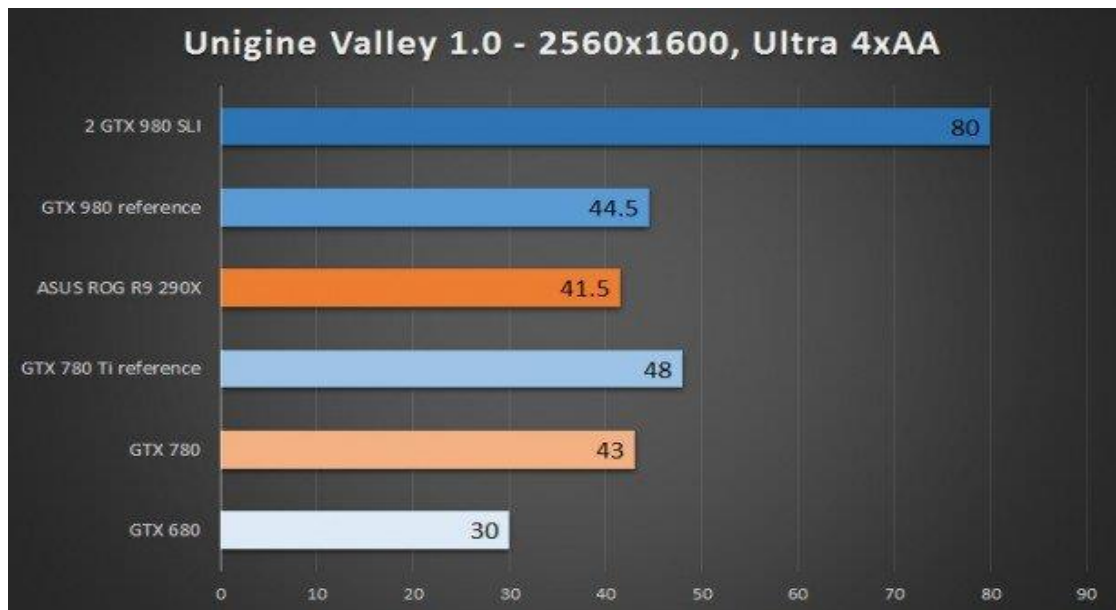
In addition, another two important features, which are not directly related to the performance of a graphic card but closely related to the limitation of architecture, are the core area and the transistor number. Because on a certain PCB board exists the limitation space for the transistor number and the core area these two features are affected mainly by the architecture, which means the performance limitation of a high-end graphic card (usually the flagship graphic card) depends on how good the architecture is.

The core area on GTX750Ti and GTX750 are both 148mm^2 with 1870 million transistors while GTX650Ti has used a 221mm^2 core area with 2540 million transistors but still cannot achieve as good its performance as the Maxwell architecture graphic cards. The comparison clearly demonstrates that Maxwell architecture is a brilliant new architecture with significant and thorough

improvements.

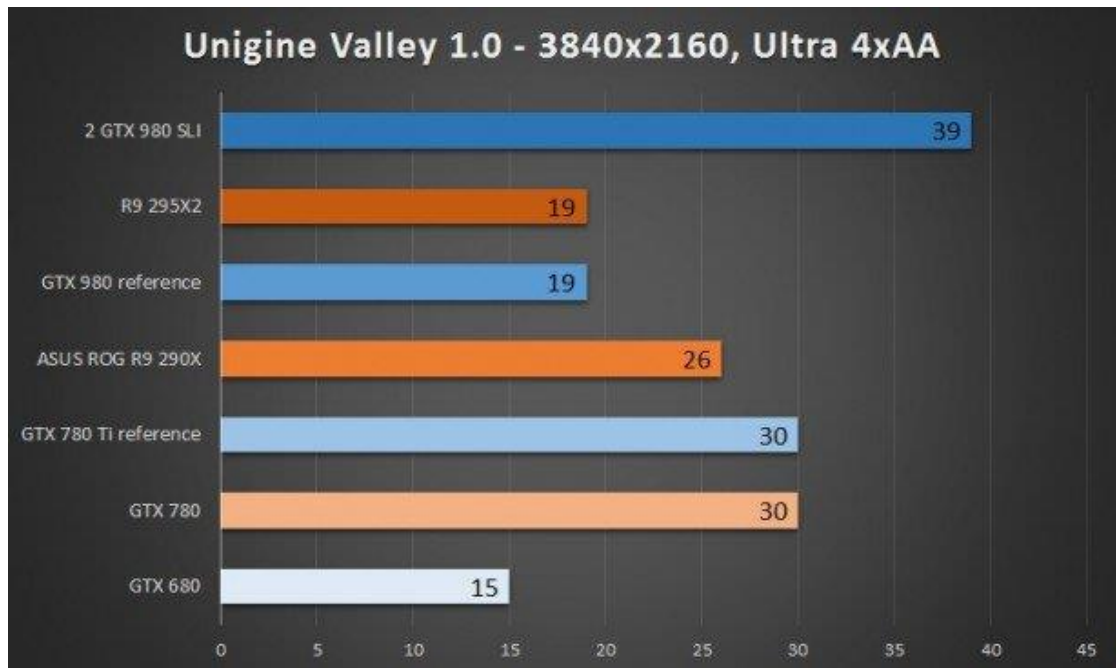
6.2.4 The SLI efficiency of Maxwell Architecture

A new architecture with much better performance/watt is a significant success, and the increased efficiency of the single card also results in SLI efficiency.



Picture 11. The theoretical test about the GTX980 SLI (Flagship card at the moment) (Felon, 2014)

As demonstrated from the graph, in the medium pressure theoretical test, GTX980 SLI has almost double performance than one single GTX980. During the Kepler generation, the SLI efficiency varied around 40-60% under medium pressure due to different models of graphics cards which means the 2-Way SLI will increase the final performance to the level of 1.4 to 1.6 graphics card of the same model. Although this is not a disappointing result, this is one of the reasons that SLI is not widely used during the Kepler generation. The 2-Way SLI of GTX980 under medium pressure has efficiency at 79%, which is a brilliant improvement compared to the Kepler architecture (Hagedoorn, 2014).



Picture 12. SLI theory test result in extreme situation (Felon, 2014)

The SLI graphics cards under high pressure have more efficiency than under medium pressure. Judging from the graph, GTX980 SLI has more than 100% efficiency.



Picture 13. Different number of SLI test result (Shang Fangwen, 2014)

When the graphics card number increases from two to four, judging from the theoretical test result, the efficiency of GTX980 SLI does not decrease.

Furthermore, the structure used by GTX980 is GM204 which is the first version of the Maxwell architecture and it is not known whether it is a full version or not, and the development of GTX980 is not finished, so that it still has much space to grow.

The test results indicate an unexpected gain, and that is the Maxwell architecture provides lower power dissipation while bringing SLI to the next level.

7. NEW DESIGN CONCEPT: QUANTIZATION GRAPHICS

CARD

After analyzing the design of the old graphics card, the fact is that the old design has so many crucial disadvantages though some of them can be solved by NVIDIA or AMD, they choose not to solve them due to competition and market structure reasons. In addition, old architectures do not have enough ability to solve problems like the power dissipation.

The appearance of Maxwell architecture provides more choices in the future graphics card design due to its much more advanced architecture arrangement which is theoretically considered possible to solve the problems.

Based on the Maxwell architecture, a new concept design of the graphics card becomes plausible rather than only a concept. The graphics card designed by this new concept is called quantization graphics card which is simply named as Q-card. In this thesis, the Q-card will be discussed in relation to the Maxwell architecture, but as it is a concept, it may be much better if the next new architecture comes out.

This concept, as its name quantization, causes the disappearance of the various models of graphics cards which means there will be no more than one model for each generation of graphics card. Furthermore, the meaning of flagship graphics card will also disappear since there will be only one model.

The quantization will bring a new order in the dedicated graphics card market. Customers will not be confused when they intend to build their own computers which will lead to much less cheating sales for those customers who do not have enough knowledge about graphics cards.

7.1 Q-card

Unlike the old designs focusing on extreme performance, the Q-card has very limited performance for one single card. To be specific, it has the least performance as a medium-level graphics card nowadays.

The first generation of the Q-card will use Maxwell architecture, for it is the best architecture nowadays which also provides the possibility for the design of the Q-card.

Table 3. Estimate statistics of Q-card

GTX 750 Ti	Q-card
CUDA Cores: 640	CUDA Cores: 256 or 384
Base Clock (MHz): 1020	Base Clock (MHz): estimate 600-700
Boost Clock (MHz): 1085	Boost Clock (MHz): estimate 700-800
Memory Clock: 5.4 Gbps	Memory Clock: 5.4 Gbps
Standard Memory Config: 2048 MB	Standard Memory Config: 512 or 1024 MB
Memory Interface: GDDR5	Memory Interface: GDDR5
Memory Interface Width: 128-bit	Memory Interface Width: 128-bit
Memory Bandwidth (GB/sec): 86.4	Memory Bandwidth (GB/sec): 86.4

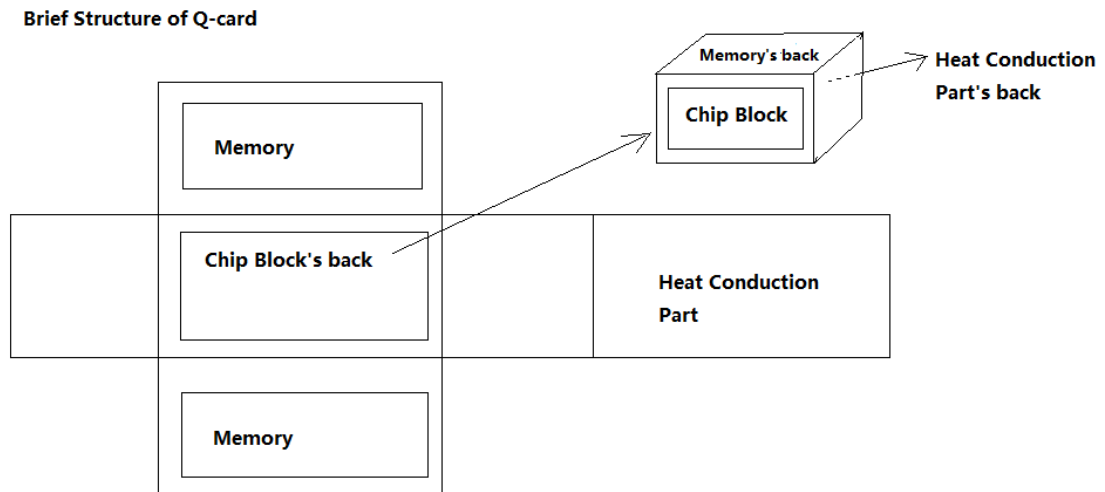
The possible statistics of the Q-card can be illustrated as in the picture above (based on Maxwell architecture) according to the estimated performance the Q-card will have.

In a comparison of statistics between GTX750TI and Q-card, the Q-card will only have 256 or 384 CUBA cores, in other words 2 to 3 control logics, rather than the 5 control logics that GTX750TI has. The clock rate will be reduced to a certain

level in order to get a smaller volume because lower clock rate equals less heat as well as less hardware requirements.

The memory clock and memory interface width stay the same while the standard memory is reduced to 512 or 1024 MB. Reducing the memory while keeping the memory interface width means the efficiency of memory usage receives a huge improvement.

Another point to be stated is that the Q-card is treated as a graphics card, but it does not have ability to work on its own. The Q-card is a module, most likely a cube shape, which only has the chip block, memory module as well as necessary circuits supporting the functioning of it.



Picture 14. Schematic representation of Q-card

The chip block of the Q-card faces outside of the Q-card while the memory and other necessary circuits are totally inside the Q-card. The 6 outside surfaces will be smooth and tough for the protection.

The performance of the Q-card is limited to a low level due to the design, which means that the chip area is reduced to a very small scale, and as the most part of the Q-card is the chip block so that the volume of the Q-card can be limited to a very small cube.

The surface that chip block belongs to will be an interface for output/input data while being a power supplying port as well. On the opposite side of this surface there is the heat dissipation surface having heat conduction components inside the Q-card connecting the back of chip block to the outside surface of heat dissipation. The outside part of the heat dissipation can be added advanced heat dissipation components, for instance, fans or water-cooling systems to achieve better heat dissipation effects.

The interface on the Q-card will connect the components inside the Q-card, like chip block as well as memory with other support components like electric capacity outside the Q-card. It is just like the old design graphics card but, the chip block and memory module can be disassembled freely. The interface also handles the power support for the Q-card. Since GTX750TI does not need extra power supply interfaces, the Q-card with simpler statistics will surely be the same which can work with the circuit directly supporting power for it.

The Q-card is designed to focus on graphics card paralleling, and on the basis of the Maxwell structure it is called SLI. The maximum number of the Q-card SLI increases from 4 to 6-10.

The maximum theoretical performance that comes from 10 Q-cards paralleling will be very impressive with the statistics they have. There will be 2560 or 3840 CUBA cores with 5120 or 10240MB memory providing very high level hardware basis. Meanwhile, because of the type of paralleling of the Q-card, the memory interface width will also add up so that the efficiency of the memory usage will stay at the same level as one single Q-card. The three main features that a graphics card's performance depends on have flawless statistics which make the theoretical maximum performance of 10 Q-card paralleling so brilliant.

The power dissipation of one single Q-card will be less than GTX750TI's 55 Watt, and according to the estimated statistics given in Picture 16, the power dissipation of the Q-card will be approximately 30-40 Watt so that the entire power dissipation of the maximum paralleling graphics cards, namely 10 Q-cards, will be around 300-400 Watt. This power dissipation seems to be pretty

high, but considering the extremely high performance it will bring, this level of power dissipation is totally acceptable and it is only the theoretical maximum power dissipation of the Q-card's potential, which will not be used by most customers.

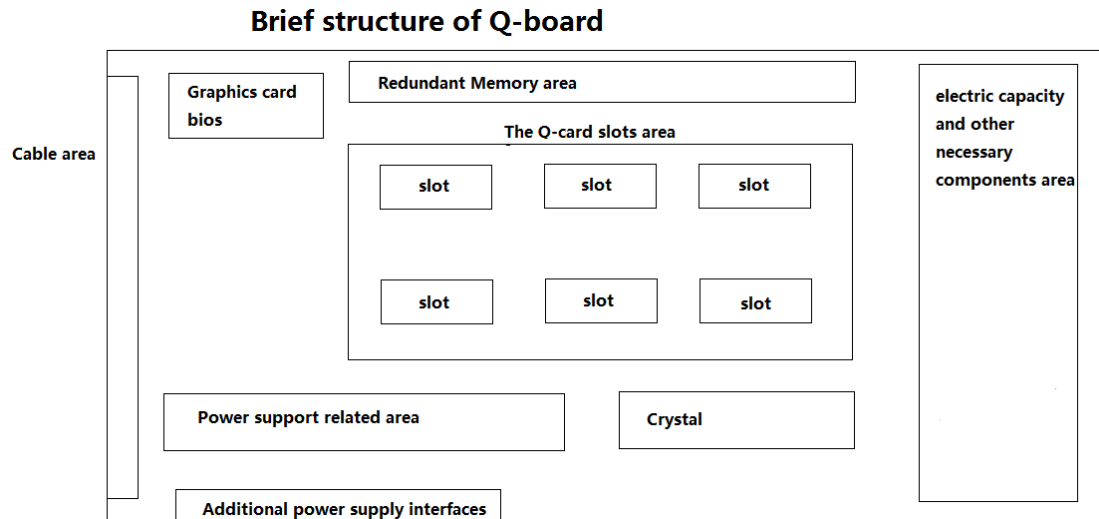
7.2 Q-board

The Q-card does not have anything else so it is not able to work on its own and the interface does not directly fit the traditional PCI-E 3.0 interface on the motherboard. So, a platform is needed to provide other required components for a graphics card, including power source, graphic cards bios, crystal and all other necessary circuits. This platform is called "Q-board".

The existence of Q-board is necessary for the quantization concept. The quantization concept as a brand new concept which not only changes the traditional structure and design of the graphics card, but also has an influence on other related hardware of a computer, especially the mother board. Furthermore, the new design graphics cards will not be able to use the same style of motherboard as the graphics cards with old design, which will still dominate the graphics card market for a long time even if the new design comes out since the evolution of the market from old design to new design will take around 4-5 years.

In addition, there still exists problem that mother boards nowadays are produced by companies other than NVIDIA and AMD so that it is not possible to acquire those companies to change the mother board structure to a brand new style which will not be able to support old design graphics cards anymore.

Conclusively, the Q-board will be a temporary method for the market transition of graphics card from the old design to the new design because the mother board design will also change to fit the new design of graphics card as soon as the market of new design graphics card becomes large enough. The Q-board is like a normal PCB of the graphics card nowadays, but it has differences in many features.



Picture 15. Schematic representation of Q-board

The Q-board has almost the same structure as the normal PCB board although the PCB board has only one fixed chip block while the Q-board has an area providing slots for the Q-card. This means that the Q-board is going to be larger than normal the PCB board. Extra heat radiator components can be added above the slot area and according to different number of the Q-card, different type of heat radiator of different level can be adjusted.

The Q-board has multiple slots for the Q-card, which means that the number of Q-card can be adjusted due to the customers' maximum need. On the Q-board, there is redundant memory in case of additional needs for high level work because memory cannot be all inserted into the Q-card. The Q-board also has additional power supply interfaces to provide power for the Q-cards if the number of Q-card inserted on the Q-board is higher than four.

The Q-cards in the slots are in the SLI status, but the SLI of Q-card is more like a chip block combination rather than two separate graphics cards paralleling. The slot area on the Q-board can be considered to be a large basis for the chip blocks so that whenever a new Q-card is implemented into the slot, its chip blocks will be added to the existed chip block structure of former Q-card. This type of paralleling also results in the memory interface width's directly adding up which will maintain the efficiency of the memory usage.

The installation of the Q-card to the Q-board is simple and flexible: the only procedure before installation or uninstallation is taking off the heat radiator components which are also easy to be taken off and put on. More importantly, Q-card is designed to be hot-swappable which means that it is not necessary to turn off the computer to install it or uninstall it.

Judging from the whole computer, the Q-board and the Q-cards inserted on it are functioning like a complete old design graphics card with slightly different structure. The chip blocks from multiple Q-cards can be seen as a whole chip block of an old graphics card although the scale of feature statistics is adaptable by the user. Only this kind of hardware structure change is apparently not enough to let the Q-card to be called a new concept. There is another crucial feature that Q-card has to declare the innovativeness of Q-card.

7.3 Q-controller

The graphics card control software is not unusual nowadays although most of them are designed for mid-high level graphics cards. This kind of graphics card control software only has very limited function including minor adjustments of the maximum clock rate and the extent of fan operation. The clock rate adjustment is for boosting the maximum clock rate to a slightly higher level to acquire better performance but lowering the clock rate than the normal is not possible. Meanwhile, the extent of fan operation adjustment is designed for users who prefer to exchange small level of temperature increasing for lower noise caused by the fan.

A kind of brand new software named Q-controller is also planned for the Q-board to control many more features of it. The most important feature that Q-controller has is the full control of each Q-card inserted onto the Q-board including the power support, the clock rate setting as well as the paralleling function. From the software, the current task that the computer is processing is assessed and monitored real-time so that when the computer handles tasks that only need a low level of performance, for instance the office tasks or watching a video, the Q-controller will automatically turn off the power supply to the unnecessary

Q-cards completely to maintain the low level necessary performance. This kind of function can significantly increase the average life time of the Q-card, for the Q-cards will not be in use, in other words totally deactivated so that Q-cards not needed for the current task will no longer stay on sleep status. It is also possible for the user to manually turn on the power of extra Q-cards depending on the user's own preference. When a current task changes from a low-level to a high-level one, like high quality games, the Q-controller will assess the needed performance of the new task, and turn on the power to a certain number of Q-cards to fit the need and needless to say, the user can also manually choose more or fewer Q-cards to be activated.

This assessing and monitoring feature will significantly save the power dissipation of the graphics card and increase the average life-time of the Q-card, specifically the chip block, because the former sleep status mode of the old designs still consumes life-time when graphics card is not fully functioning.

The clock rate of each Q-card can also be adjusted by the Q-controller although judging from current SLI technology, all Q-cards activated must use the same clock rate at the same time.

The quantization concept of the graphics cards has the three main components mentioned above: the Q-card, the Q-board, as well as the Q-controller. The hardware and software parts' cooperation will maintain the functioning of the new design graphics card together.

7.4 Upgrades & Adjustments

The Q-card as the only model of each generation totally gives up the market structure of the graphics card used to have. The difference between different level graphics cards becomes the number of Q-card rather than the difference in architecture, which results in the normalization of the graphics card market. In addition, this change to a great extent reduces the complexity in graphics card's architecture design and manufacture in order to make the manufacturing companies of graphics cards to have more focus on improving the architecture

design which turns out to be virtuous cycle.

The new design is very easy to upgrade. A customer with a formerly bought graphics card who intends to upgrade the graphics card to handle higher level tasks now no longer has to buy a brand new graphics card and desert the old one. Instead, he can just buy additional Q-cards to achieve better performance and the whole upgrade procedure adding the new bought Q-cards onto the Q-board, and then the upgrade is completed. Furthermore, the exchange of the broken Q-card is also very simple and it is obvious that the old design graphics card would be immediately deserted if part of its chip block is broken.

The graphics cards with old designs relies on the macrocosm of all components in it, or to be specific all parts on the PCB. Many graphics cards nowadays need a total repair or even are deserted due to only one electric capacity failure, no matter how well its chip block is able to function. The new concept reduces the maintenance pressure very much because just like the broken Q-card, is easy to be abandoned without further changes of other components, the Q-board can also be changed to a new one while Q-cards can still be used on the new Q-board.

7.5 The PCB issue

As a brand new concept, quantization graphics cards design surely faces several technical problems, among which the most important problem is the PCB circuit arrangement.

All components of a graphics card are placed on the PCB board, and the Q-board is also a kind of PCB. According to the functioning of a graphics card, the data has to be transferred to each component at exactly the same time so that the circuits on the PCB have to be arranged precisely to fit the needs of the data transfer (Eurocircuits, 2013).

The graphics card with old designs faces a much simpler situation for the reason that each component is fixed to the certain position which remains unchangeable so that the circuits are easy to calculate as well as implement

because everything is on the plan list.

The modular feature of the Q-card and the Q-board makes the calculation of circuits much more difficult because the number of Q-cards and the position of Q-cards on a Q-board are not defined so that the circuits have to be flexible and competent enough to face different situations.

A temporary solution of this problem is using extra circuit layers for each different situation. The current average layers of old designed graphics cards is 4-6 according to the graphics card's level. To fit the needs of the Q-board, the number of layers may increase to 10-20 because the chip block of each Q-card needs its own circuits connected to the slots. According to the design of Q-card and Q-board, some slots on the Q-board may be not in use in certain situation so that this solution unavoidably has waste of circuits but this is the only plausible method currently.

In the future with the development of the electronic manufacture, this issue may be solved although it is not a very crucial problem for the Q-card and Q-board design currently.

7.6 Future improvements & influence

This thesis has discussed the quantization concept in the Maxwell architecture but the design can be improved with the development of this new architecture in the future so that this concept has a promising future.

This new architecture always has two basic features: more performance and less power dissipation. According to the concept of quantization, these two features are the core points of the Q-card designs which means that this new architecture in the future will lead to a smaller volume, as well as less power dissipation Q-card.

Furthermore, after the first generation of Q-card has taken the most of the graphics card market, which means that most of the computers on this planet are using Q-boards and Q-cards instead of the old design graphics card. The most

important advantage of quantization concept will be obvious.

In hotels or internet bars or offices, the computers will be using Q-board and Q-card and those computers can only provide only one Q-card on its Q-board to cater for the basic needs of a computer while users can bring their own Q-cards to attach on the Q-board according to the tasks or issues they intend to do. For instance, a customer who goes to an internet bar just for news browsing or video watching, he can use the only one Q-card provided by the internet bar while if he intends to play a high quality game, he can bring 2-3 Q-cards of his to attach on the Q-board of the internet bar computer. According to the new concept design, the Q-card has very small volume as well as an easy installation procedure so that it will not cause inconvenience for the customers in the future.

This circumstance reduces the huge waste of hardware of the computers in hotels, internet bars and offices than it used to be because, as discussed, many tasks can be performed with very low performance so that the waste of the medium-high level graphics card like GTX660TI in the internet bar is unavoidable.

In addition, after the Q-card takes most of the market share of graphics cards, it will be very likely to see the cooperation between motherboard manufacturers and graphics card manufacturers to make a new design of motherboard on which Q-cards can be directly attached. This uniform structure will abandon the Q-board transition so that the reliability will be much better than the first generation of Q-cards with Q-board.

In conclusion, the quantization concept will be a very plausible design for the graphics card in the future which will have a huge influence on the hardware usage efficiency as well as the graphics card market.

REFERENCES

Eurocircuits. 2014. Making a PCB - PCB Manufacture step by step, [online]
Available at:

<http://www.eurocircuits.com/Making-a-PCB-PCB-Manufacture-step-by-step>

(Accessed 25 April 2015)

Felon W. Sep 19, 2014. Nvidia GTX980 tested: SLI, 4K, and single-GPU
benchmarks and impressions, [online] Available at:

<http://www.pcgamer.com/nvidia-gtx-980-tested-sli-4k-and-single-gpu-benchmark-s-and-impressions/#page-2>

(Accessed 11 April 2015)

Hagedoorn H. 2014. GeForce GTX 980 2 and 3-way SLI review – Introduction,
[online] Available at:

<http://www.guru3d.com/articles-pages/geforce-gtx-980-sli-review,1.html>

(Accessed 11 April 2015)

Jon Peddie Research. 2015. Add-in board market up in Q1, Nvidia increases
market share lead, [online] Available at:

<http://jonpeddie.com/publications/add-in-board-report/>

(Accessed 10 March 2015)

Lily P. May 6, 2011. NVIDIA Leads in Discrete Desktop GPU Market Share, AMD in Notebook Graphics, [online] Available at:

<http://www.maximumpc.com/nvidia-leads-in-discrete-desktop-gpu-market-share-amd-in-notebook-graphics/>

(Accessed 03 March 2015)

Li Jiasheng. 2014. Maxwell coming: GTX750TI/GTX750 test, [online] Available at:

<http://diy.pconline.com.cn/429/4295160.html>

(Accessed 15 April 2015)

Nvidia.com. 2014. Company information, [online] Available at:

<http://www.nvidia.com/object/about-nvidia.html>

(Accessed 10 March 2015)

NVIDIA. 2014. *NVIDIA GeForce GTX 750 TI, whitepaper, NVIDIA*, pp 6-8

NVIDIA. 2013. *NVIDIA's next generation CUDA compute architecture: GK110, whitepaper, NVIDIA*, pp 6

NVIDIA. 2012. *NVIDIA GeForce GTX680, whitepaper, NVIDIA*, pp 5-6

NVIDIA. 2014. *NVIDIA GeForce GTX 980, whitepaper, NVIDIA*, pp 5-7

NVIDIA. 2011. Introduction to SLI technology, [online] Available at:

<http://www.geforce.com/whats-new/guides/introduction-to-sli-technology-guide#2>

(Accessed 04 April 2015)

Shang FW. 2014. GTX980 4-way SLI test, mydrivers.com, [online] Available at:

http://news.mydrivers.com/1/322/322679_all.htm

(Accessed 01 April 2015)

Weiss S. 2012. GPUs and GPU programming, [online] Available at:

http://compsci.hunter.cuny.edu/~sweiss/course_materials/csci360/lecture_notes/gpus.pdf

(Accessed 05 March 2015)

VideoCarz. 2014. NVIDIA GeForce GTX 750 TI Specifications, [online] Available at:

<http://videocardz.com/nvidia/geforce-700/geforce-gtx-750-ti>

Wikipedia. The 'Dedicated graphics card VS integrated graphics card' part, [online] Available at: http://en.wikipedia.org/wiki/Video_card

(Accessed 02 March 2015)

Zhang Wei. 2013. GTX780 iChill test, [online] Available at:

http://www.pcpop.com/doc/0/910/910592_all.shtml

(Accessed 15 March 2015)

Zhang Qingle. 2012. Top players' gear: ASUS GTX680 practical test, [online]

Available at: http://diy.pconline.com.cn/graphics/reviews/1204/2735323_1.html

(Accessed 01 April 2015)