XGI TECHNOLOGY WHITEPAPER

BitFluent^{TM} Architecture

XGI Technology Inc.
Product Marketing
2003/08/12

BitFluent^{TM}
Architecture

Introduction

In the pass few years, the GPU size has grown so rapidly that it has raise to a historic high. There are a few problems that GPU venders will be facing in the very near future. For example, the yield has dropped to a historic low, the average die cost has risen to a historic high and heat dissipation might affect the fan sink design and also the life expectancy of the GPU. As we are looking further into the future, there shall only be a few products that will drive the fabrication technology, the CPU and the GPU. Research reports have also shown that the progress in fabrication technology will slow down as it runs into nanometer age. This information is critical at the beginning stage of the product development. As we have seen on the competitor that the cost, risk and challenges with adopting new process on large GPU might have big impact to the operation of the company. XGI has adopted the idea of dual GPU architecture which will provide greater performance with less production risk, and more flexibility with product planning.

The All new Dual GPU configuration

BitFluent Architecture Configuration

1. 32-bit data bus width
2. 133MHz operation frequency
3. Maximum bandwidth: 133*4*4 = 2.13GB/s
4. Low pin count: total 38 pins
5. 1.5V low voltage swing I/O. 3.3V compatible.
6. Source synchronous signal strobes
7. Multi-channel architecture: can transfer PCI (including config registers, standard VGA registers, enhanced registers, MMIO and memory access), AGP (include texture read, vertex read and command read),CRT, secondary display and 2D engine Bitblt data.
8. AGP and PCI Express like Split transaction and pipeline command support
9. Pipeline execution in each virtual channel
10. Parallel execution among each channel
11. Flexible arbitration scheme and variable bandwidth of each channel
12. Credit-based flow control
13. Merge requests and completions to increase bandwidth utilization rate
14. Support PCI snoop memory write function
15. Support 4X/2X/1X flexible operation modes
16. Support DBI (dynamic bus inversion)
17. Support dynamic change mode
18. Embedded auto-test function

Data flow through BitFluent Bridge

The BitFluent Bridge has been designed with the concept of not having any bottleneck at the data path so that both Volari V8 can perform at its peak. As shown on the following figure, the Volari V8 on the left is called the primary processor and the one on the right is called the secondary processor. The primary processor is design to be connected with the AGP8X bus with the secondary processor connected to the primary processor. The connection between the two GPU are called 'BitFluent Bridge' and is having a high speed interface of up to AGP8X. This is designed so that the performance will not be bounded by interface.

High Speed interface up to AGP8X

The interface will also support AGP2X and AGP4X. *(Figure 1)*

Primary Chip CRT display data path

At the final stage of the 3D processing, GPU will store pixel data in the local frame buffer and the display data will be fetched from local frame buffer to GPU then finally to the CRT display. *(Figure 2)*

Secondary Chip CRT display data path

Similar to the primary chip, secondary GPU will store pixel data in its local frame buffer and will firstly fetch the data to the secondary GPU then to the Primary GPU and finally to the CRT display. *(Figure 3)*

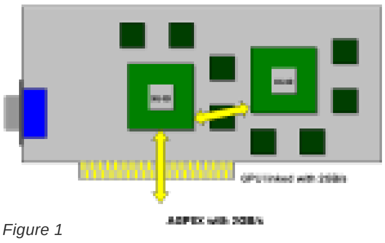


Figure 1

256-bit DDR/DDR2 DRAM Bandwidth

The maximum bandwidth will reach astonished 32GB/s with 500MHz DDR2. The 8 channels, 256-bit combination will have a smaller DRAM access overhead than the 4 channels, 256-bit organization. This shall ensure higher utilization rate and higher performance. *(Figure 4)*

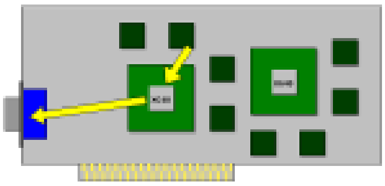


Figure 2

Volari V8 has built-in with eight 3D pixel pipelines

Volari Duo V8 Ultra with BitFluent Bridge will provide a total of sixteen 3D pixel pipelines configuration in a single AGP board. The VGA board shall reach astonished performance than every graphics board with only eight 3D pixel pipelines.

Parallel processing for each primary and secondary GPU

In order to achieve the highest performance with the dual GPU configuration, the driver must also maintain two command queues for each chip. The two GPU work as parallel processors and one GPU must not wait for the other one. Both Volari V8 can assert pipelined commands to AGP bus simultaneously and wait for data to come back from AGP bus.

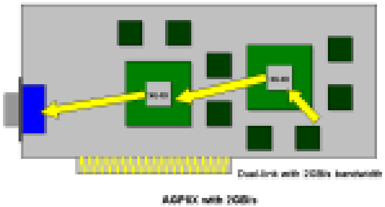


Figure 3

Dual AGP command queues for each chip

The driver handles the command dispatch and make working load balanced. *(Figure 5)*

PCI snoop memory write aperture for automated duplicating data for both chips.

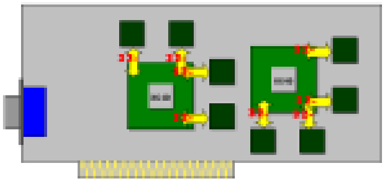


Figure 4

Virtual channel architecture to optimize the data transaction latency and bus utilization.

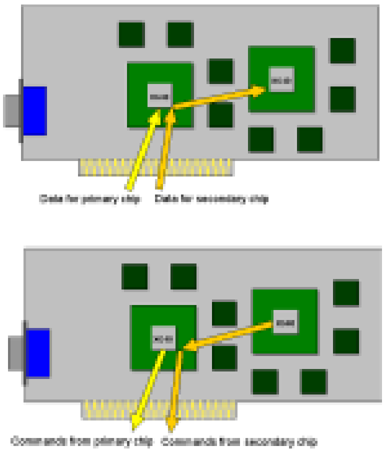
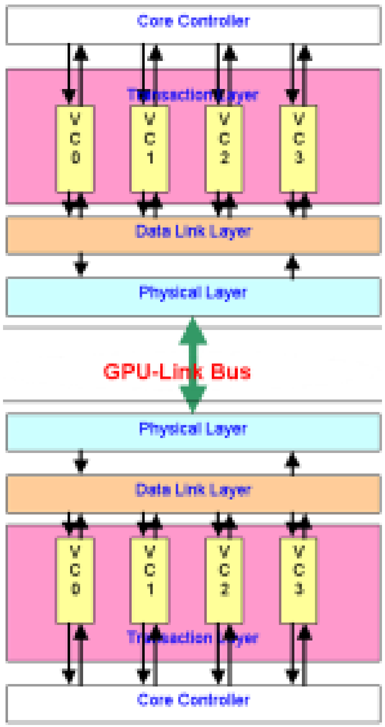


Figure 5



Summary

As we have observed over the years, GPU is the IC device with the highest growing rate of transistor counts. This also means that the die size and cost for average cost has grown dramatically. We believe there must be some way we could redefine the growing trend without pushing the cost into the sky. Dual GPU is believe to be the most logical and efficient way. In this kind of configuration, highest performance could be reached with minimal production risk and maximal product planning flexibility.