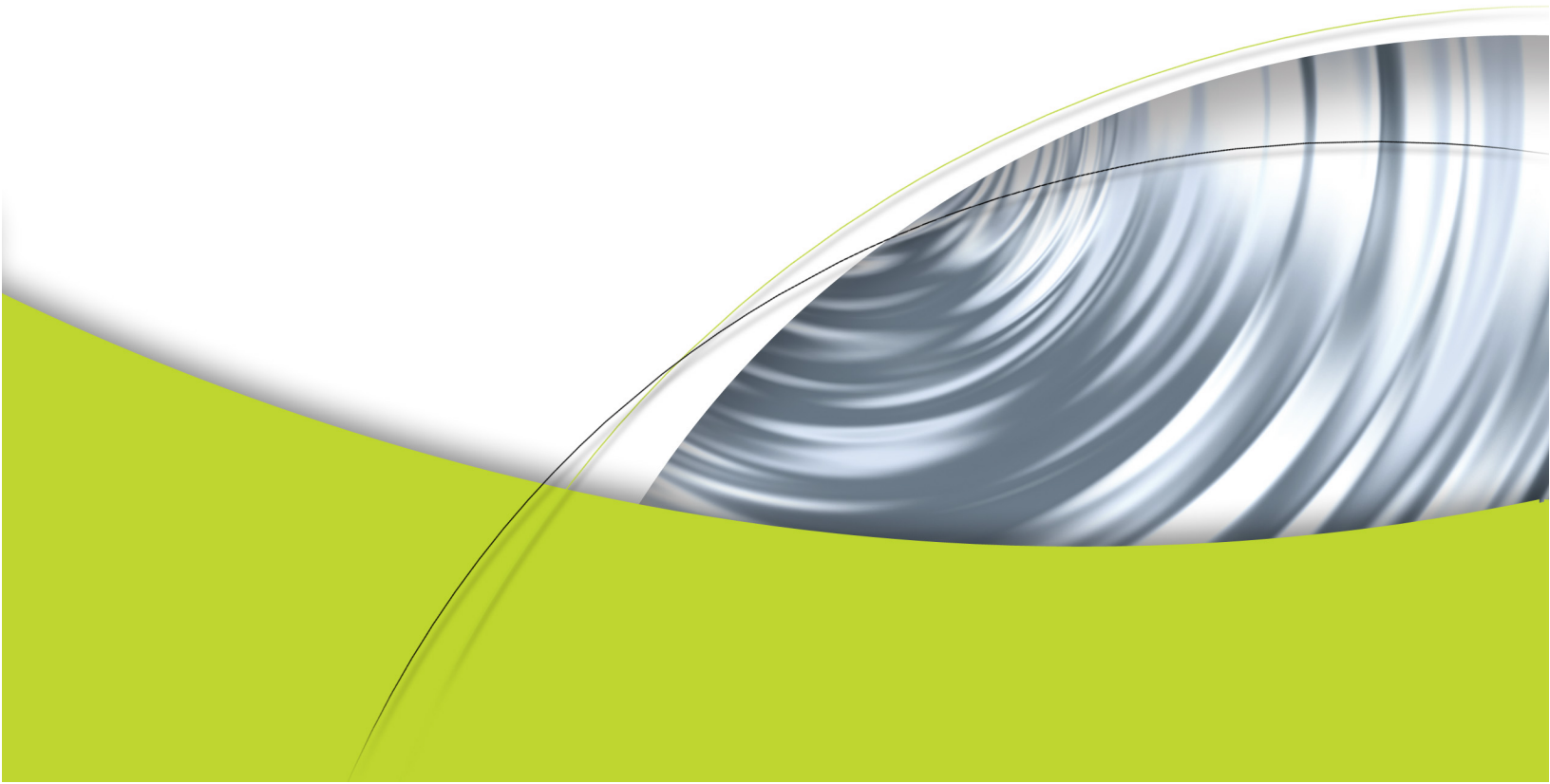
Technical Brief

AGP 8X

Evolving the Graphics Interface



Increasing Graphics Bandwidth

No one needs to be convinced that the overall PC experience is increasingly dependent on the efficient processing of graphics information. Web sites, games, and virtually every type of application employs graphics to deliver more intuitive, visually compelling views into the world of information.

Platforms have evolved to better accommodate this explosion of visual information. At its inception, the Accelerated Graphics Port (AGP) specification defined a more optimized path than PCI for moving data between the graphics subsystem, the central processing unit (CPU) and memory. This interface continues to play a vital role. But, many forces are stressing the AGP 4X bus today:

- **Content:** Graphic scenes contain increasing amounts of complex geometry and texture data.
- **Precision:** Developers are clamoring for higher-precision data. The next-generation NVIDIA graphics processing units (GPUs) offer true 128-bit color, the same level of precision used in the film industry today and the basis for achieving spectacular cinematic effects in real time. The wider data formats increase the bandwidth requirements throughout the system.
- **Interactivity:** Real-time scene changes are stressing the AGP bus with the dynamic loading of games environments as well as streaming special effects through the graphics subsystem.
- **System advances:** The capabilities of both host platforms and graphics subsystems continue to evolve with increases in processor speeds, memory capacity and bandwidth, and multiprocessing functionality. The speed of the bridge between them must evolve for software programs to be able to benefit from these advances.

The latest release of the AGP specification—version 3.0—represents a defining moment in the PC industry and introduces the AGP 8X interface to address each of these trends. AGP 8X doubles the graphics bus bandwidth, dramatically improving the overall throughput for today's graphics-intensive applications. In order to take full advantage of current and emerging graphics applications, NVIDIA has introduced the industry's first and only top-to-bottom family of AGP 8X-ready GPUs and core logic products.

This paper describes the benefits of the latest AGP enhancements and details the NVIDIA plans for adoption of the new standard.

The AGP Specification

The AGP standard originated to define high-performance interconnects for enhancing 3D graphics performance. The dedicated high-speed port connects the core-logic chipset and graphics controller, creating a direct path for transferring graphics textures in and out of memory when the local frame buffer space has been exceeded or a new scene must be loaded. This design offers several benefits:

- AGP can transfer texture data at gigabytes per second (GB/sec.) rates that exceed that of the PCI bus (2.1GB/sec. instead of 132MB/sec. over PCI), and can support execution of texture maps from system memory rather than forcing all texture data to be preloaded into local graphics memory.
- The specification includes a sideband-addressing mode that allows the GPU to issue new addresses and requests before the previous request is finished.
- The PCI bus becomes much less congested, maximizing the performance of the devices restricted to that bus (disk controllers, LAN cards, video capture systems, etc.).

The AGP 3.0 Release

Since its introduction in 1996, the AGP interface has been updated in an evolutionary fashion. Originally, the specification arose to address shortcomings in the popular PCI bus, and to define an interface tailored to the demands of graphics operations and data movement. Revisions to the specification have focused on scaling bandwidth:

- AGP 1X and AGP 2X bandwidth levels were introduced simultaneously in version 1.0 of the AGP 1.0 interface (or AGP 1.0) specification. AGP 1.0 allowed for these two interface speeds, where AGP 2X was theoretically twice as fast as AGP 1X. (See Table 1 in the performance section for details on the various transfer rates.)
- AGP 4X, defined in the AGP 2.0 specification, was introduced two years later.
- Today, AGP 3.0 brings AGP 8X bandwidth to the industry and introduces isochronous operation and AGP texturing abilities (described in detail in following sections).

Version 3.0 of the AGP interface (or AGP 3.0) doubles the theoretical performance over the bus. This latest version of the specification also incorporates some new features, and removes some unused features to simplify the interface. NVIDIA is supporting AGP 3.0 beginning with GPU and core logic offerings scheduled for release in Fall 2002.

Optimizing Graphics Operations and Texture Storage

System designs and graphics subsystems in particular have advanced at a rapid pace since AGP 2.0 emerged in 1998. Today, AGP 2.0 and its associated AGP 4X bandwidth has become a bottleneck in the overall flow of graphics-related data (see Figure 1).

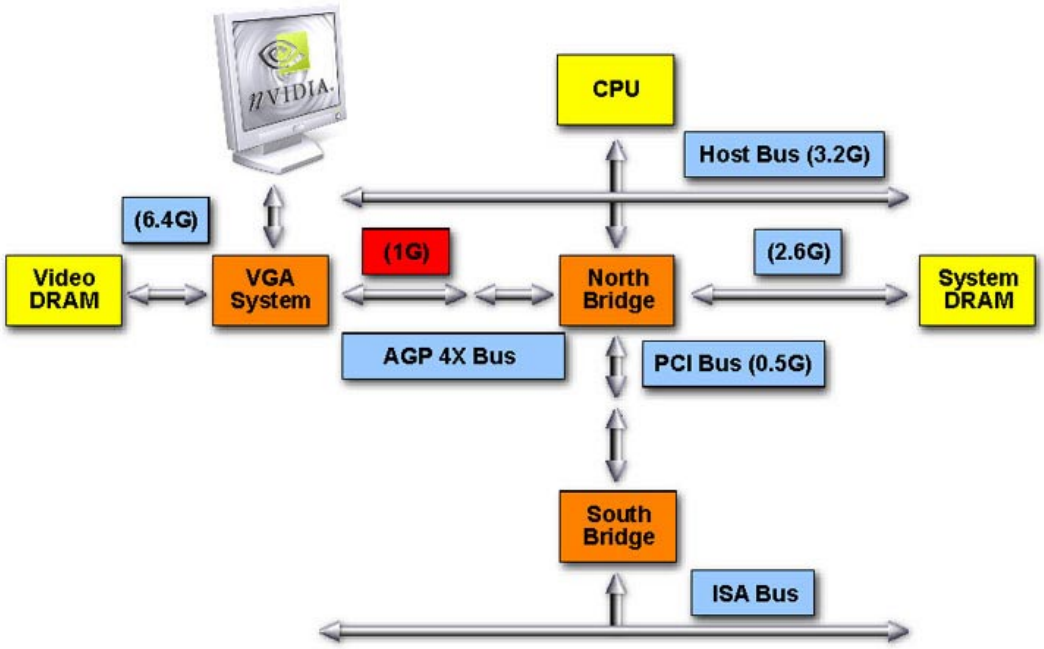


Figure 1. AGP 4X bandwidth falls short of the rest of the system. With AGP 8X, the graphics bandwidth will be balanced with the other system and memory throughput rates.

With a theoretical bandwidth of 2.1GB/sec., AGP 8X also allows developers to more efficiently manage scenes with complex geometries, and dynamically switch to new scenes in real time. By lowering the overhead associated with storing and retrieving complex textures or streaming data from memory, AGP 8X increases the overall throughput for the AGP bus. This in turn will yield significant performance improvements for visualizations that involve complex textures and geometry, and will inspire more realistic experiences for the user.

Isochronous Operation For Streaming

The overall bandwidth improvements of AGP 3.0 benefit all graphics operations involving complex geometries, textures, or streaming. One new AGP 3.0 feature—isochronous mode operation—specifically improves graphics operations that require predictable, uninterrupted data flow. Previous versions of the AGP interface cannot guarantee the required bandwidth for latency-sensitive transfers. This "best effort" arrangement works well to achieve low average latency and high average throughput, but does not protect against an occasional arbitrarily long delay now and then, and therefore can result in lost data. Streaming applications—applications

that involve real-time flows of digital information for video broadcasts, network downloads, and similar tasks—do not tolerate data loss and require predictable transfers. Further, low-cost designs require a manner to support isochronous transfers without increasing the required amounts of expensive data buffering. With isochronous mode operation, AGP 3.0 addresses these application requirements in a cost-effective manner.

Compatibility

The AGP 3.0 specification provides a smooth upgrade path to AGP 8X. The mechanical bus specification remains the same. AGP 8X speeds and capabilities are achieved by taking advantage of some previously unused pins, but in a manner that facilitates the support of AGP 8X cards in existing AGP 2X and 4X systems, as well as new systems that fully leverage the 8X interface. NVIDIA AGP 8X graphics solutions will be able to detect the AGP level of the host system, and automatically configure the AGP interface to run in 3.0 mode (at 4X or 8X speeds), or in 2.0 mode (at 2X or 4X speeds). Therefore, a new NVIDIA graphics solution will be fully capable of 8X speeds, and will be completely compatible with 2X, 4X, and 8X systems. The NVIDIA-based cards will automatically deliver the maximum speed supported by the host system.

AGP 8X Performance

The AGP 8X bandwidth is double that of AGP 4X. The impact of the AGP 8X bandwidth on overall application performance will vary with the type of applications:

- **Static worlds:** Applications that run within a small virtual environment—in essence, the entire "world" is loaded into graphics memory at all times—will see little, if any, improvement in performance due to the AGP 8X rate of transfer.
- **Complex worlds:** Today's fast-paced "fly-through" applications and games will see significant improvements in overall performance due to the doubled AGP 8X speed. These applications and games must predict and load geometries and textures into the frame buffer and benefit from the improved throughput between main memory and the graphics subsystem. Applications with high-precision data and large textures will also benefit, since these also require more transfers to and from main memory.

Table 1. Comparison of AGP 4X and AGP 8X.

	AGP 4X	AGP 8X
Bytes per transfer	4 (32 bits)	4 (32 bits)
Clock rate	266.67MHz	533.33MHz
Bus bandwidth	1.1GB/sec.	2.1GB/sec.

Conclusion

Graphics applications stress every part of the system and require constant advances to maintain a balanced environment and avoid bottlenecks. AGP 3.0, by delivering AGP 8X bandwidth, takes an revolutionary step forward in the bandwidth continuum. The increased bandwidth and the improved bus design complement the emerging graphics hardware and will inspire more creative and practical ways to use system memory resources for supporting complex textures and visualizations. While maintaining compatibility with existing AGP 2X and 4X systems, new AGP 8X solutions can support a painless path to:

- Real-time cinematic graphics and better use of main memory for complex geometries, textures, and higher-precision data.
- Higher-performance applications that involve dynamic "world" loading or streaming.
- Support for the isochronous transfers that characterize many digital media applications such as streaming from digital devices and networks.
- Balanced system performances where graphics operations and data flow are intelligently offloaded from the CPU.

AGP 8X provides immediate performance improvements for many games and applications that employ complex textures and scenes. This newest release of the AGP interface takes the bandwidth pressure off of the graphics subsystem, and provides the headroom that will be required to handle applications for years to come.

New NVIDIA GPUs and PC platforms will incorporate AGP 8X capabilities and take full advantage of the AGP 3.0 specification. As always, NVIDIA solutions incorporate new technologies and advances without compromising the overall stability and quality of the system, and the NVIDIA unified driver architecture (UDA) makes it painless to take advantage of new capabilities as they are introduced.

Notice

NVIDIA Corporation
2701 San Tomas Expressway
Santa Clara, CA 95050
www.nvidia.com