Technical Brief

NVIDIA nForce IGP
TwinBank Memory Architecture

I. Memory Bandwidth and Capacity—There's Never Enough

With the recent advances in PC technologies, including high-speed processors, large broadband pipelines, and realistic 3D graphics and 3D positional audio, we all continue to rely on the PC for many of our daily tasks. E-mail, stock quotes, 3D gaming, news broadcasts, digital audio and video—our reliance on the PC is balanced only by the boundless limitations it has to offer. However, advanced as we might think we are, there's still one problem that affects our daily technology experience: memory. In fact, you would be hard pressed not to find someone who hasn't experienced an "out of memory error", or a fatal crash, timed coincidentally to occur right before they had a chance to save their work in progress. The solution to this problem lies in the PC design, whose underlying technology and memory infrastructures struggle mightily to keep up with our list of daily demands.

NVIDIA®'s patent-pending TwinBank™ Memory Architecture, an innovative 128-bit memory controller supporting DDR-266MHz system memory technologies, was designed to support such demands, achieve optimal system and graphics performance, and provide the highest memory bandwidth possible in the process. Fully scalable, with support for a wide variety of memory configurations, TwinBank's dual-independent, crossbar memory controller, a key innovation in NVIDIA's Platform Processing Architecture's nForce™ Integrated Graphics Processor (IGP), grants the CPU, GPU and Media and Communications Processor (MCP) simultaneous access to the system's 4.2GB/sec. of memory bandwidth, guaranteeing continuous access for all applications, every time.

II. Driving the Need for More Bandwidth

There are four key elements in determining system and graphics performance in an SMA system:

- Graphics processor
- Available system memory bandwidth
- CPU memory read latency
- Context-switching overhead and arbitration efficiency

The advanced integrated GPU with its high-performance dual pixel pipeline architecture, second-generation transform and lighting engine, and 256-bit 3D/2D engines enable end-users to run 3D applications with more complex graphics objects and more realistic 3D environments at higher resolutions, higher color depths (32-bit color), higher frame rates and higher refresh rates. Each of these features consumes valuable system memory bandwidth and can easily push the current PC-133 (1.05GB/s) and PC-800 RDRAM (1.6GB/s) memory technologies beyond their limits. New CPUs such as AMD®'s Athlon™ or Duron™ employ a DDR front-side bus running up to 133MHz (266MHz effective). With the CPU core frequency running at 1GHz and beyond, the front-side bus utilization will be extremely high, consuming a theoretical peak of 2.1GB/sec. of the system memory bandwidth. Figure 1 shows how the memory bandwidth requirement increases with higher resolution, higher color depth, and other 3D

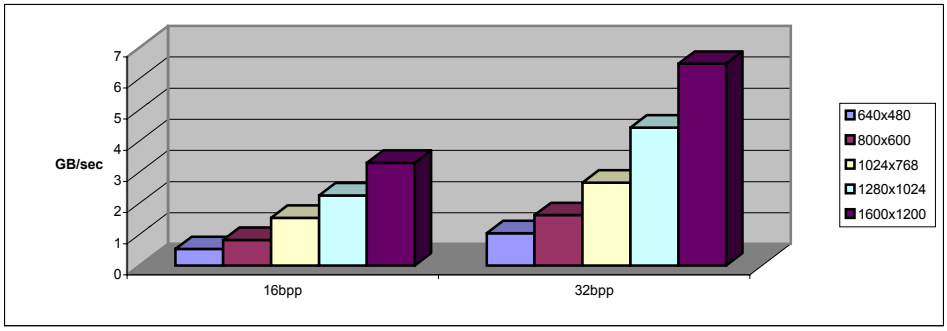

Figure 1: Memory bandwidth requirements for a simple 3D application

features. To complicate matters, the system memory bandwidth in traditional shared memory architectures (SMAs) is shared between the activities of the integrated GPU, the CPU, and the many "Southbridge" peripheral devices, some of which are isochronous (time-dependent) in nature. Therefore, with traditional 128-bit memory architectures, it's very difficult to keep the high-performance CPU satisfied while servicing the GPU and the other, many real-time "Southbridge" peripherals. Since the average read latency of the CPU is greatly increased, system performance is negatively impacted. Also, SMA-based "core-logic chipsets" are typically perceived as low-cost, low-performance designs for the basic and value PC market segments. This perception exists for good reason: The majority of PC OEMS have to make cost/performance tradeoffs in order to meet price targets for various PC market segments, a task accomplished only by using low-performance graphics cores and memory subsystem architectures that are largely ineffective. Although the price is sometimes right, end-users generally dislike such solutions,

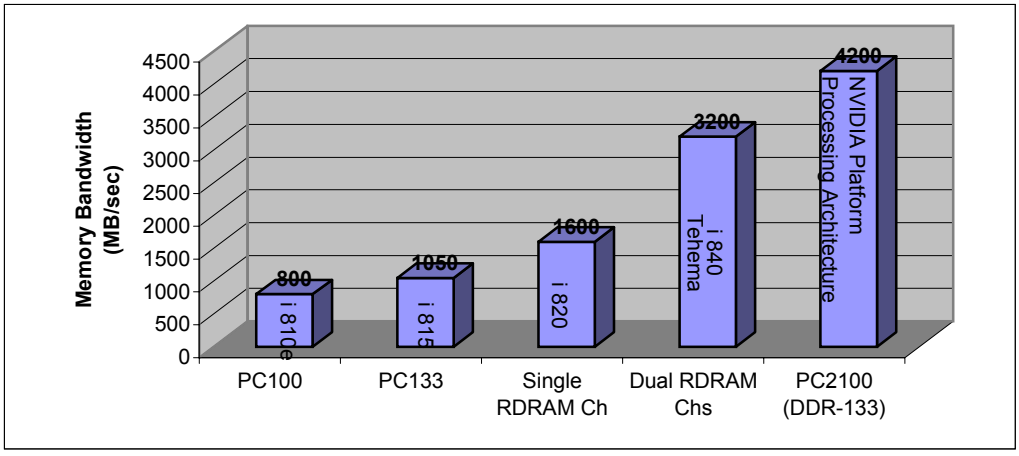

Figure 2: 128-bit DDR delivers 30% more bandwidth than dual RDRAM channels

especially when they realize they've traded in a lower price for extremely limited performance. Clearly, there is a need for high-performance, cost-effective alternatives.

To achieve optimal system and graphics performance, TwinBank avoids the SMA approach and adopts an innovative dual-independent, 64-bit memory controller architecture to support 128-bit DDR-

133MHz (PC2100-266MHz) system memory, delivering a peak bandwidth of 4.2GB/sec. By employing DDR memory technologies, TwinBank provides a much more cost-effective solution compared to high-cost RDRAMs, (which do not scale well with increasing CPU speeds, and have worse latency than DDR). TwinBank delivers 30% more bandwidth than other dual-channel RDRAM chipsets available on the market such as Intel®'s i840 or the upcoming i850 (Tehama) for the Pentium® 4 CPU (as shown in Figure 2). In comparison to current 64-bit PC-133 architectures, TwinBank quadruples the available memory bandwidth.

III. The TwinBank Architecture

Crossbar Memory Controller

TwinBank consists of two independent 64-bit DDR-266 memory controllers (MC0 and MC1) to deliver a whopping 4.2GB/s peak memory bandwidth. This is four times the memory bandwidth of PC133 SDR memory and over two-and-a-half times of single RAC 800MHz DRDRAM. The radical crossbar memory controller enables CPU and GPU to concurrently access the two 64-bit memory banks and is optimized for 64-bit CPU and GPU accesses to ensure near perfect bandwidth utilization. The two memory controllers are interleaved so that consecutive CPU memory requests can be started before the previous one is completed, reducing CPU read latency. The TwinBank architecture allows the two independent 64-bit memory controllers to access 128-bits of data on each clock cycle using DDR memory, effectively fetching

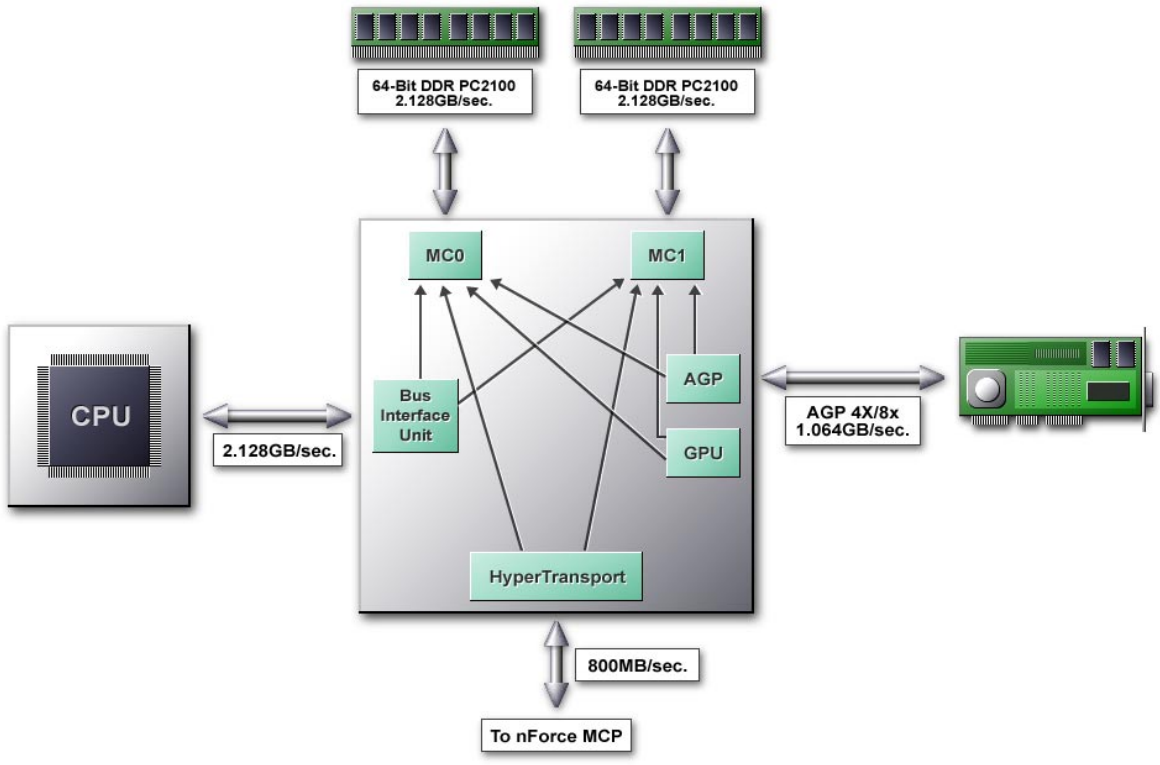

Figure 3: Conceptual TwinBank Concurrent Accesses

256-bits of data total on each clock cycle. Since the high-performance CPU and GPU data types are optimized for 64-bit access, both can access the two memory banks simultaneously *and* independently, fully utilizing available memory bandwidth. The average read latency of the CPU is now greatly reduced, which increases both graphics and system performance. Without this type of architecture, there would be tremendous bottlenecks in the system with the high-performance CPU and GPU both struggling for access to valuable system memory bandwidth.

Imagine a scenario where the GPU is fetching AGP high color depth 3D texture data from the system memory to render on the CRT while refreshing the high-resolution/color depth/refresh rate display. The CPU has to wait until the AGP and CRT refresh transfers are completed before it can fetch new data, stalling the CPU pipeline. Furthermore, if the AGP access closes the DRAM pages that the CPU is going to access, it will create context-switching overhead that introducing additional CPU latency. With TwinBank, the CPU and GPU/AGP can both access the two banks simultaneously, increasing both system and graphics performance through concurrency, higher available bandwidth and lower context-switching overhead. As TwinBank architecture increases parallelism in graphics and CPU accesses, it benefits both the integrated GPU as well as the external AGP add-in card in terms of textures and other graphics/video data accesses, further enhancing graphics and system performance.

Single-Step Memory Arbitration

The NVIDIA Platform Processing Architecture enables a typical user to run high-performance 3D applications, capture video using a USB camera, conduct video conferencing on the Internet, convert MP3 music files and create CDs simultaneously. This creates a very complex multi-threaded, multi-tasking environment that necessitates the need for a concurrent memory arbitration architecture that can facilitate the various high-bandwidth, low-latency, and isochronous real-time data streams and device requirements within the PC.

TwinBank's single-step memory arbitration unit is designed around the intimate knowledge of the memory traffic of the GPU, as well as the other masters within the chipset, instead of just adding black box units to the chipset. This allows TwinBank's dual independent memory controller architecture, with its highly efficient arbitration logic, to enable multiple data streams, such as CPU, integrated GPU or AGP 4X add-in card, and the various MCP functions (multiple PCI, dual IDE ATA-100, multiple USB, Fast Ethernet, audio/modem, and more), to access system memory simultaneously, substantially minimizing system latency and increasing performance. The highly efficient arbitration logic can access the 64-bit memory banks independently with interleaved data within the two 64-bit memory banks. Each of the independent 64-bit controllers can take full advantage of the DDR memory's available bandwidth and access 128-bits per clock. It can also open and efficiently manage 24 pages with all three DIMMs installed. This allows high degrees of concurrency, which again increases performance. In comparison, systems utilizing a multiple arbitration process introduce increased latency, reducing overall system performance.

Flexibility, Scalability, and Upgradeability

The TwinBank architecture is designed to provide a high degree of flexibility, scalability and upgradeability. Its technical benefits include:

- Both 64-bit and 128-bit operations. In 64-bit mode, the DIMM can be located on either MC1 or MC2. In 128-bit mode, both MC1 (DIMM0) and MC2 (DIMM1/DIMM2) are utilized.
- Both controllers are functionally identical with all control and timing parameters independently programmable. This allows asymmetric DIMMs with different memory organization, size, and speed to be used on MC1 and MC2 and still provide the full performance benefits of the 128-bit memory system.
- Support for both 3.3V standard SDRAM or 2.5V DDR SDRAM memory technologies.
- Support for 133/100MHz DDR (266/200MHz) SDRAM or 133/100MHz standard SDRAM clock frequencies.
- Support for 1-3 unbuffered, non-ECC DIMMs.
- Support for 64, 128, 256, or 512Mbit x8 or x16 memory configurations.
- Support for 64MB to 1.5GB of system memory.
- Support for odd total memory size; eg. 64MB + 128MB = 192MB, while still taking advantage of the 128-bit TwinBank architecture.

IV. Benefits of the TwinBank Architecture

The obvious key benefit to end-users is significantly increased graphics and system performance. This increase in performance is illustrated in the four performance benchmark graphs below. It truly allows the end user to experience full 3D graphics environments on a low-cost, yet powerful PC.

The end-user also has an easy and low-cost upgrade option. By simply installing an additional 64-bit DIMM to a default 64-bit system, not only will they experience an increase in Microsoft® Windows® performance, but the resulting TwinBank performance will substantially increase system and graphics performance due to higher bandwidth and increased concurrency. This automatically provides additional open memory pages to utilize thereby reducing context-switching overhead and CPU read latency that

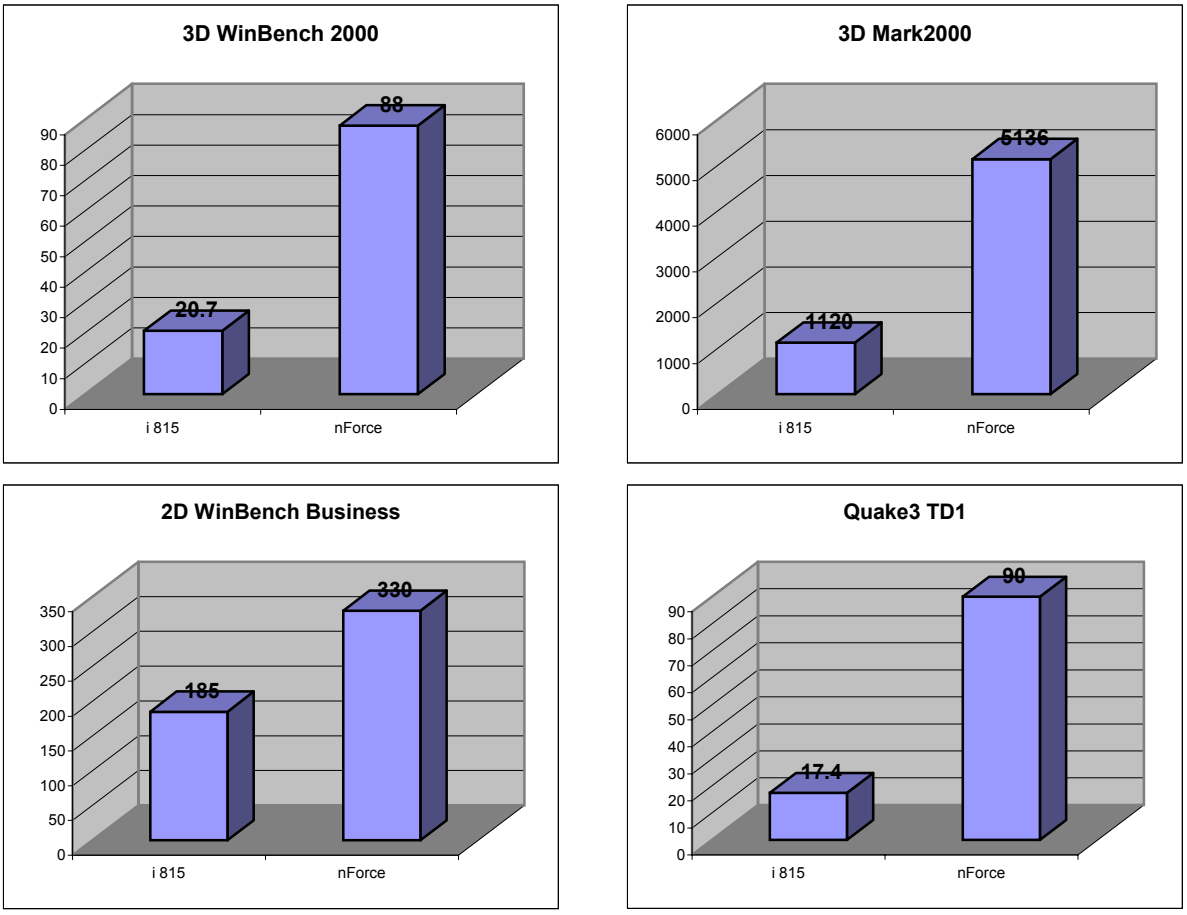

Figure 4: nForce's TwinBank performance compared to the Intel i815 core-logic chipset

again improves system performance. Finally, the end-user has the option of using an even more powerful external AGP GPU, such as NVIDIA's GeForce3, which also takes full advantage of the TwinBank dual independent 64-bit memory controller architecture for dramatic increases in performance.

V. Conclusion

Quite simply, TwinBank provides a no-compromise PC graphics memory solution, allowing everyone to take advantage of today's most intensive applications. Not only can you run multiple applications concurrently, you can do so without fear that you'll bring your system to a screeching halt. From 3D gaming to Web browsing, TwinBank provides the power and the performance to allow high-performing GPUs and the most powerful CPUs to run to their fullest capabilities.

As part of the NVIDIA nForce Platform Processing Architecture, TwinBank delivers the most cost-effective, highest performing system memory architecture in history. By quadrupling the bandwidth of current 64-bit PC-133 designs from 1.05GB/sec. to 4.2GB/sec. while also delivering a 30% increase in bandwidth as compared to the very expensive, complex, and high-latency dual-channel RDRAM designs present in much more expensive PC workstations, TwinBank is the only memory solution suitable for the next-generation of PC architectures, such as NVIDIA's nForce Platform Processing Architecture, and radically redefines the baselines on which traditional SMA systems should be compared.

Appendix A – Glossary

DDR SDRAM: Double Data Rate SDRAM.

DIMM: Dual In-Line Memory Module.

DVI: Digital Video Interface. A new interface to connect to digital monitors such as flat panel displays, digital CRTs, and flat panel projectors.

GB/sec.: Gigabytes per second.

GPU: Graphics Processing Unit. The IGP integrates an NVIDIA GeForce2™ GPU on-chip. This white paper will use GPU and graphics processor interchangeably.

Northbridge: One half of a PC core logic chipset that interfaces to the CPU, GPU, memory, AGP and Southbridge.

PC-100: 100MHz standard SDRAM 64-bit DIMM system memory.

PC-133: 133MHz standard SDRAM 64-bit DIMM system memory.

PC-800: 800MHz Rambus DRAM RIMM system memory.

PC-2000 (PC-200): 100MHz Double Data Rate SDRAM 64-bit DIMM system memory.

PC-2100 (PC-266): 133MHz Double Data Rate SDRAM 64-bit DIMM system memory.

RDRAM: Rambus DRAM.

RIMM: Rambus In-Line Memory Module.

SDRAM: Synchronous DRAM.

SMA: Shared Memory Architecture. The total installed system memory is shared between the system operating system (Windows) and the graphics frame buffer (2D, 3D, video, texture).

Southbridge: One half of a PC core logic chipset that interfaces to the Northbridge and various peripherals (PCI, IDE ATA-100, USB, Fast Ethernet, audio/modem, etc.).