

Technical Brief

Technische Details

**NVIDIA nForce IGP
TwinBank Speicherarchitektur**

I. Speicherbandbreite und Kapazität – es gibt niemals genug davon

Aufgrund der aktuellen Fortschritte im Bereich der PC-Technologien, zu denen Hochgeschwindigkeits-Prozessoren, große Breitband-Pipelines, realistische 3D-Grafiken und 3D-Positional Audio gehören, verlassen wir uns alle, was die Erledigung unserer täglichen Aufgaben betrifft, auf den PC. E-Mail, Börsenkurse, 3D-Spiele, Nachrichten, digitales Audio und Video – unser Vertrauen in den PC wird nur durch seine Beschränkungen begrenzt. Egal, für wie fortschrittlich wir uns halten mögen, gibt es jedoch ein Problem, das unsere Technologieerfahrungen beeinträchtigt: Speicher. Sie werden sicherlich kaum jemanden finden, der keinen „Speicherfehler“ oder einen völligen Systemabsturz erlebt hat, natürlich bevor überhaupt die Möglichkeit gegeben war, die Arbeit zu sichern. Die Lösung für dieses Problem liegt im PC-Design, dessen untergelegte Architektur und Speicher-Infrastrukturen damit zu kämpfen haben, mit unseren täglichen Anforderungen Schritt zu halten.

NVIDIAS zum Patent angemeldete TwinBank™ Speicherarchitektur, ein innovativer 128-Bit Speichercontroller, der DDR-266MHz Speichertechnologien unterstützt, wurde darauf ausgerichtet, optimale System- und Speicherleistung und die größtmögliche Speicherbandbreite zu erzielen. Vollständig skalierbar und mit Support für eine Vielzahl von Speicherkonfigurationen, garantiert der TwinBank dual-independent Crossbar-Speichercontroller, eine der Schlüsselinnovationen im Integrierten Grafikprozessor (IGP) der NVIDIA Plattformprozessor-Architektur, dem CPU, dem GPU und dem Media- und Kommunikationsprozessor (MCP) simultanen Zugang zur Speicherbandbreite von 4,2GB/Sek – garantierter Zugang für alle Applikationen zu jeder Zeit.

II. Der Bedarf an höherer Bandbreite

Es gibt vier Schlüsselemente zur Ermittlung der System- und Grafikleistung in einem SMA-System:

- den Grafikprozessor
- die verfügbare System-Speicherbandbreite
- die CPU-Speicher Leselatenz
- die kontextbezogene Overhead- und Arbitration-Effektivität

Der fortschrittliche integrierte GPU mit seiner Hochleistungs-Dual Pixel Pipeline-Architektur, der

Transform und Lighting Engine der zweiten Generation und 256-Bit 3D/2D Engines, ermöglicht den Endverbrauchern, 3D-Applikationen mit komplexeren grafischen Objekten und realistischeren 3D-Umgebungen bei höheren Auflösungen, einer höheren Farbtiefe (32-Bit Color), höheren Frame Rates und höheren Refresh Rates auszuführen. Jedes dieser Leistungsmerkmale belegt wertvolle System-Speicherbandbreite und bringt die aktuellen PC-133 (1,05GB/Sek) und PC-800 RDRAM (1,6GB/Sek) Speichertechnologien leicht an ihre Grenzen. Neue CPUs wie der AMD® Athlon™ oder Duron™ verfügen über einen DDR Front Side Bus bei 133 MHz (effektiv 266MHz). Wenn die CPU-Kernfrequenz bei 1GHz und darüber läuft, wird die Ausnutzung des Front Side Bus extrem hoch sein., sie verbraucht theoretisch maximal 2,1GB/Sek der System-Speicherbandbreite. Abbildung 1 zeigt, wie die Anforderung der System-Speicherbandbreite bei höherer Auflösung, einer höheren Farbtiefe und anderen 3D-Leistungsmerkmalen steigt.

GB/sec	GB/Sek
--------	--------

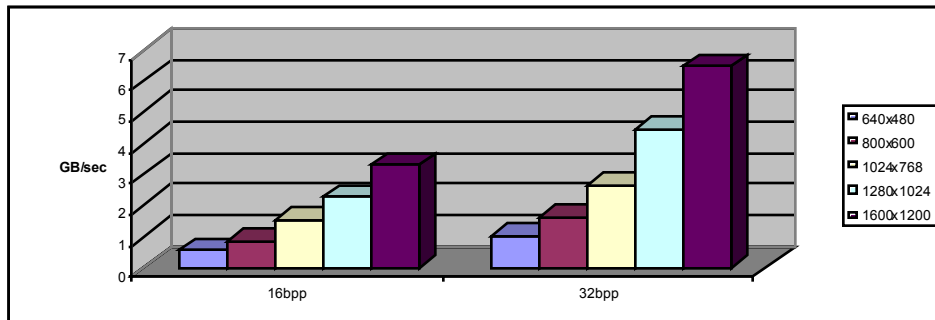


Abbildung 1: Anforderungen an die Speicherbandbreite in einer einfachen 3D-Applikation

Was es noch komplizierter macht ist die Tatsache, dass die System-Speicherbandbreite in traditionellen SMAs zwischen den Aktivitäten des integrierten GPU, dem CPU und den vielen „Southbridge“-Peripheriegeräte geteilt wird, von denen einige von ihrer Natur her isochron (zeitabhängig) sind. Es ist aus diesem Grunde sehr schwierig, mit traditionellen 128-Bit Speicherarchitekturen einen Hochleistungs-CPU und gleichzeitig den GPU und die vielen anderen Echtzeit-„Southbridge“ Peripheriegeräte zu versorgen. Da die durchschnittliche Leselatenz des CPU stark angestiegen ist, wird die Systemleistung negativ beeinflusst. Darüber hinaus werden „kernlogische Chipsätze“ auf SMA-Basis häufig als günstige Designs mit geringer Leistung für die Basis- und preiswerten PC-Marktsegmente verstanden. Das hat seinen guten Grund: Die meisten PC

OEMs müssen Kosten-/Leistungs-Preisabgleiche durchführen, um die Preisziele für verschiedene PC-Marksegmente zu erzielen – eine Aufgabe, die nur bei Verwendung von Niedrigleistungs-Kerngrafik und von Speichersubsystem-Architekturen durchgeführt werden kann, die zum Großteil ineffektiv agieren. Obwohl der Preis manchmal stimmt, mögen die Endverbraucher derartige Lösungen nicht,

Memory Bandwidth (MB/sec)	Speicherbandbreite (MB/Sek)
NVIDIA Plattform Processing Architecture	NVIDIA Plattformprozessor-Architektur

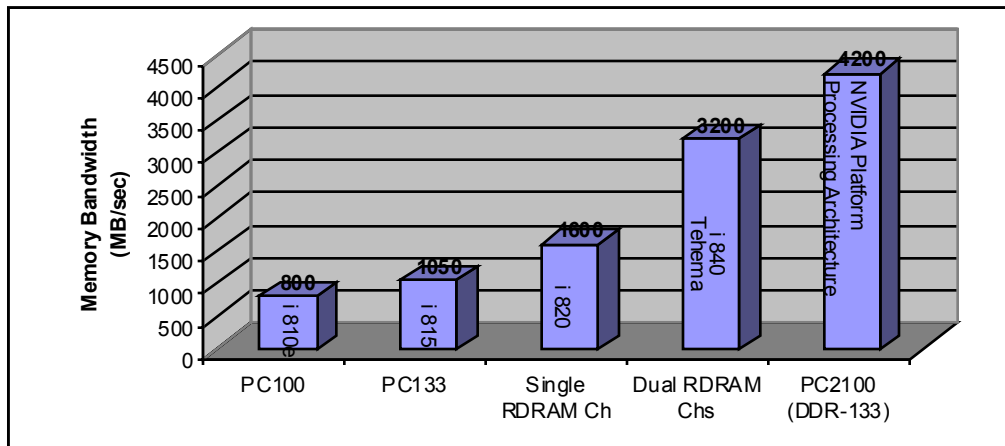


Abbildung 2: 128-Bit DDR liefert 30% mehr Bandbreite als duale RDRAM-Kanäle

besonders wenn sie merken, dass sie einen niedrigeren Preis gegen eine extrem eingeschränkte Leistung eingetauscht haben. Ganz klar ausgedrückt: Es besteht ein hoher Bedarf an leistungsstarken, preiseffektiven Alternativen.

Zur Erreichung einer optimalen System- und Grafikleistung umgeht TwinBank SMA und verwendet eine innovativen, dual-independent, 64-Bit Speichercontroller-Architektur zur Unterstützung von 128-Bit DDR 133 MHz (PC2100-266MHZ) Systemspeicher, der eine Spitzenbandbreite von 4,2GB pro Sekunde liefert. Durch die Verwendung von DDR-Speichertechnologien ermöglicht TwinBank eine kosteneffektivere Lösung im Vergleich zu hochpreisigen RDRAMs, (die sich bei steigenden CPU-Geschwindigkeiten nicht gut machen und eine schlechtere Latenz als DDR haben). TwinBank liefert eine um 30% höhere Bandbreite als andere Zweikanal RDRAM-Chipsätze, die auf dem Markt erhältlich sind, wie Intel@s i840 oder der

bevorstehende i850 (Tehama) für den CPU Pentium 4 (wie in Abbildung 2 ersichtlich). Im Vergleich zu aktuellen 64-Bit PC-133 Architekturen, vervierfacht TwinBank die zur Verfügung stehende Speicherbandbreite.

III. Die TwinBank Architektur

Crossbar-Speichercontroller

TwinBank besteht aus zwei unabhängigen 64-Bit DDR 266 Speichercontrollern (MC0 und MC1), die eine hervorragende Spitzenbandbreite von 4,2GB/Sek erbringen. Das ist die vierfache Speicherbandbreite eines PC133 SDR Speichers und mehr als zweieinhalb mal so viel wie die Bandbreite eines einzelnen RAC 800MHz DRDRAM. Der radikale Crossbar Speichercontroller ermöglicht dem CPU und GPU den gleichzeitigen Zugang zu den zwei 64-Bit Speicherbanken, und ist auf 64-Bit CPU- und GPU-Zugriffe spezialisiert, um eine nahezu perfekte Bandbreitenausnutzung zu erzielen. Die zwei Speichercontroller sind versetzt angeordnet, so dass nachfolgende CPU-Anfragen begonnen werden können, bevor der vorherige Vorgang beendet ist, wodurch die Leselatenz des CPU reduziert wird. Die TwinBank Architektur erlaubt zwei unabhängigen 64-Bit Speichercontrollern auf 128 Datenbits in jedem Taktschritt unter Verwendung von DDR-Speicher zuzugreifen,

CPU	CPU
2.128GB/sec	2,128GB/Sek
64 Bit DDR PC 2100 2.128GB/sec	64-Bit DDR PC 2100 2,128GB/Sek
MC0	MC0
MC1	MC1
Bus Interface Unit	Busschnittstellen-Einheit
AGP	AGP
GPU	GPU
AGP4X/8X 1.064GB/sec	AGP4X/8X 1,064GB/Sek
HyperTransport	HyperTransport
800MB/sec	800MB/Sek
To nForce MCP	zum nForce MCP

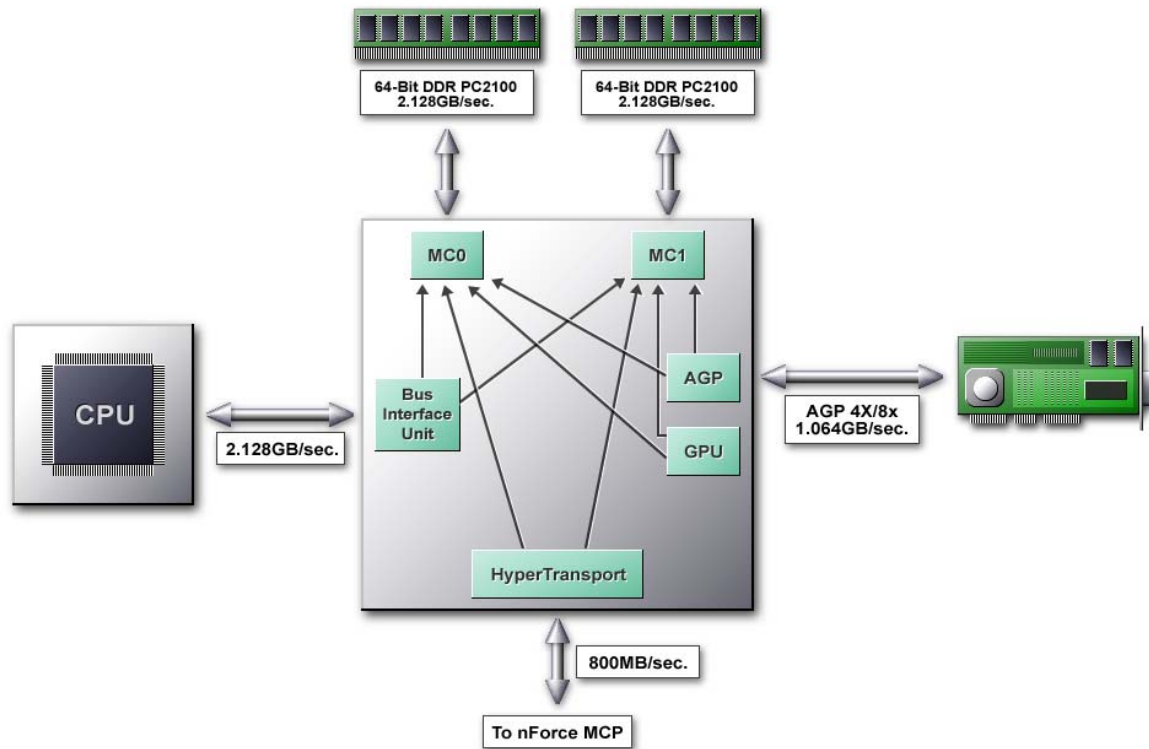


Abbildung 3: Logische Gleichzeitige TwinBank-Zugriffe

wobei effektive 256 Datenbits bei jedem Taktschritt erreicht werden. Da die Hochleistungs-Datentypen des CPU und GPU für 64-Bit Zugang optimiert sind, können beide simultan *und* unabhängig voneinander auf die zwei Speicherbanken zugreifen, und dabei die verfügbare Speicherbandbreite vollständig ausnutzen. Die durchschnittliche Leselatenz des CPU wird so in hohem Maße reduziert, sowohl die Grafik- als auch die Systemleistung wird gesteigert. Ohne diesen Architekturtyp gäbe es enorme Engpässe im System, da CPU- und GPU beide um wertvolle Systembandbreite kämpfen müssten.

Stellen Sie sich ein Szenario vor, in dem der GPU AGP High Color 3D-Texturdaten vom Systemspeicher zum Rendern auf dem CRT abzieht, während das High-Resolution/Color Depth Refresh Rate Display erneuert wird. Der CPU muss so lange warten, bis die AGP- und CRT- Refresh-Übertragungen abgeschlossen sind, bevor er auf neue Daten zugreifen kann, die CPU Pipeline wird somit blockiert. Wenn der AGP darüber hinaus die DRAM-Seiten, auf die der CPU zugreifen will schließt, wird er einen kontextbezogenen Overhead erstellen, der für zusätzliche CPU-Latenz sorgt. Mit TwinBank können der CPU und GPU/AGP beide gleichzeitig auf beide Bänke zugreifen, durch

die Gleichzeitigkeit wird sowohl die System- als auch die Grafikleistung erhöht, die verfügbare Bandbreite gesteigert und der kontextbezogene Overhead ist geringer. Da die TwinBank-Architektur parallel ablaufende Vorgänge bei den Grafik- und CPU-Zugriffen zulässt, profitiert sowohl der integrierte GPU als auch die externe AGP Einsteckkarte in Bezug auf Texturen und andere Grafik-/Videodaten-Zugriffe; die Grafik- und Systemleistung wird noch weiter gesteigert.

Einzelschritt Speicher-Arbitration

Die NVIDIA Plattformprozessor-Architektur ermöglicht einem typischen Benutzer, Hochleistungs-3D-Applikationen zu nutzen, Videos mit einer USB-Kamera aufzunehmen, Videokonferenzen über das Internet durchzuführen, MP3-Musikdateien zu konvertieren und gleichzeitig CDs zu erstellen. Damit wird eine sehr komplexe multi-threaded, multi-tasking Umgebung erzeugt, durch die der Bedarf an einer gleichzeitigen Speicherarbitrations-Architektur erzeugt wird, die die unterschiedlichen Hochbandbreiten-, Niedriglatenz- und isochrone Echtzeit-Datenströme und die Geräteanforderungen innerhalb des PCs erleichtern kann.

TwinBanks Einzelschritt-Speicherarbitration weiss um den GPU-Speicherverkehr und die anderen Master im Chipsatz, es werden nicht nur einfach dem Chipsatz Black Box-Einheiten zugefügt. Diese Tatsache ermöglicht der TwinBank dualen unabhängigen Speichercontroller-Architektur mit ihrer hocheffizienten Arbitration-Logik, mehrere Datenströme, wie die des CPU, des integrierten GPU oder der AGP 4X Einsteckkarte und die verschiedenen MCP-Funktionen (mehrere PCI, duales IDE ATA-100, Fast Ethernet, Audio/Modem und mehr), gleichzeitig auf den Systempeicher zuzugreifen, dabei die Systemlatenz zu minimieren und die Leistung zu steigern. Die hocheffiziente Arbitration-Logik kann unabhängig auf die 64-Bit Speicherbanken mit versetzten Daten innerhalb der zwei 64-Bit Speicherbanken zugreifen. Jeder der unabhängigen 64-Bit Controller kann vollständig von der dem DDR Speicher zur Verfügung stehenden Bandbreite profitieren und auf 128-Bit pro Takt zugreifen. Er kann, sobald alle drei DIMMs installiert sind, ebenfalls 24 Seiten öffnen und effizient verwalten. Dadurch wird ein hoher Grad der Parallelausführung erreicht, die noch einmal die Leistung steigert. Im Vergleich dazu haben Systeme mit mehreren Arbitration-Prozessen den Nachteil, eine erhöhte Latenz zu erzeugen, die in einer Verringerung der Gesamt-Systemleistung resultiert.

Flexibilität, Skalierbarkeit und Upgrade-Möglichkeit

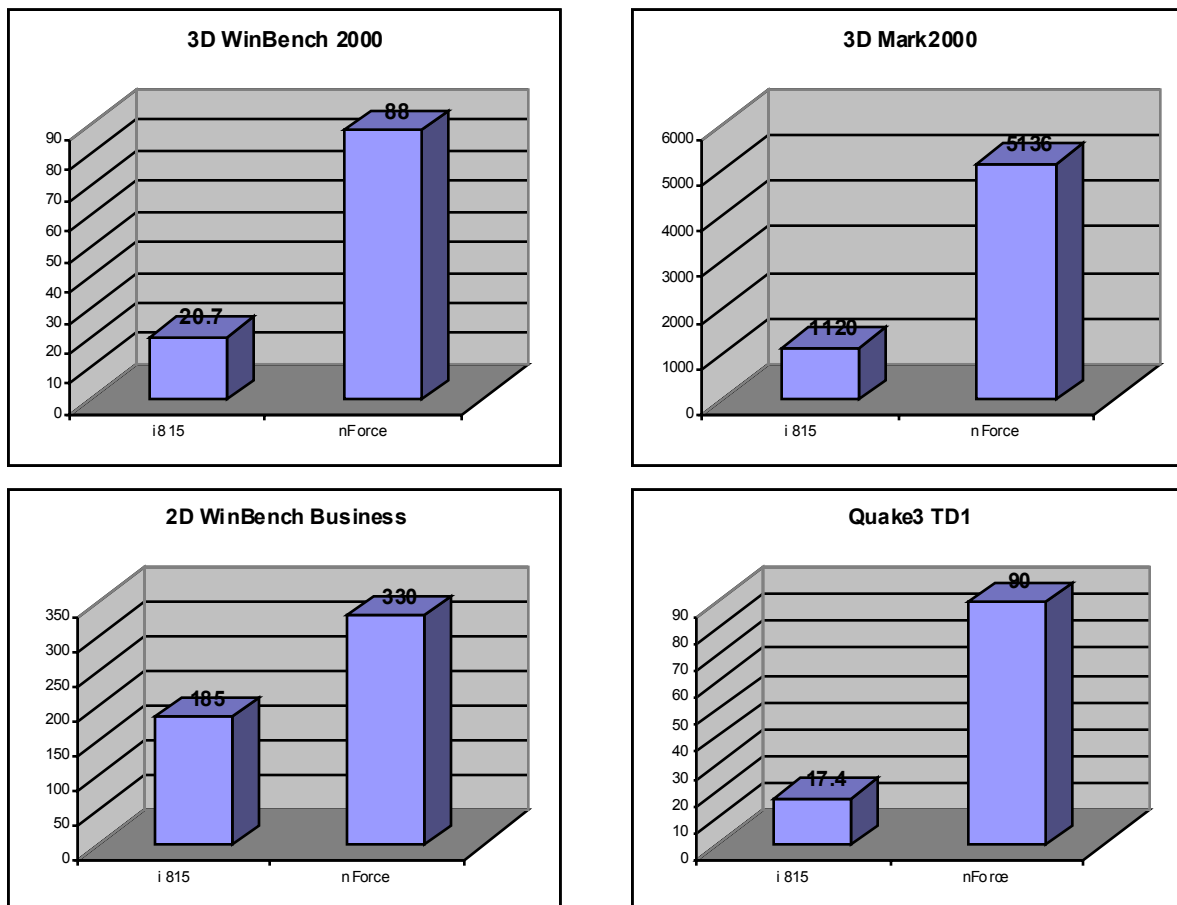
Die TwinBank-Architektur ist darauf ausgerichtet, einen hohen Grad an Flexibilität, Skalierbarkeit und der Upgrade-Möglichkeit zu ermöglichen. Zu den technischen Leistungsmerkmalen gehören:

- Sowohl 64-Bit als auch 128-Bit Operationen. Im 64-Bit Modus kann der DIMM entweder auf dem MC1 oder dem MC2 eingesetzt werden. Im 128-Bit Modus werden sowohl MC1 (DIMM0) und MC2 (DIMM1/DIMM2) verwendet.
- Beide Controller sind von der Funktion her identisch, die Kontroll- und Timing-Parameter sind unabhängig voneinander programmierbar. Dies ermöglicht den Einsatz asymmetrischer DIMMs mit verschiedenen Speicherorganisationen, Größe und Geschwindigkeit auf dem MC1 und MC2 bei vollständiger Nutzung der Leistungsvorteile eines 128-Bit Speichersystems.
- Support für 3,3V Standard SDRAM oder 2,5V DDR SDRAM Speichertechnologien.
- Support für 133/100MHz DDR (266/200MHz) SDRAM oder 133/100MHz Standard SDRAM Taktfrequenzen.
- Support für 1-3 ungepufferte, nicht ECC DIMMs.
- Support für 64, 128, 256, oder 512Mbit x8 or x16 Speicherkonfigurationen[^].
- Support für 64MB bis 1,5GB Systemspeicher.
- Support ungerader Speichergrößen, z.B. 64MB + 128MB = 192MB bei gleichzeitiger Nutzung der 128-Bit TwinBank Speicherarchitektur.

IV. Vorteile der TwinBank Architektur

Der offensichtlich größte Vorteil für den Endbenutzer liegt in der bedeutend erhöhten Grafik- und Systemleistung. Der Leistungsanstieg wird in den vier Leistungs-Benchmarkgrafiken weiter unten dargestellt. Der Endverbraucher bekommt wirklich die Möglichkeit, vollständige 3D-Grafikumgebungen in einem preisgünstigen PC zu erfahren, der aber trotzdem über Leistung verfügt.

Der Endbenutzer hat auch eine einfache und preisgünstige Upgrade-Option. Durch das einfache Installieren eines zusätzlichen 64-Bit DIMMs in ein Standard 64-Bit-System wird nicht nur eine verbesserte Leistung von Microsoft® Windows® erzielt, sondern die TwinBank Leistung wird die System- und Grafikleistung aufgrund der höheren Bandbreite und gestiegenem Parallelsatz steigern. Dadurch werden automatisch zusätzlichen offenen Speicherseiten frei, die genutzt



werden können, der kontextbezogene Overhead und die CPU-Latenz werden reduziert, wodurch

Abbildung 4: Die Leistung der nForce TwinBank im Vergleich zu einem Intel i815 Kernlogik-Chipsatz

die Systemleistung noch einmal erhöht wird. Letztendlich hat der Benutzer die Option, einen noch leistungsfähigeren externen AGP-GPU wie den NVIDIA GeForce3 einzusetzen, der ebenfalls vollen Nutzen aus der dual independent 64-Bit Speicherarchitektur TwinBank zieht, um einen drastischen Leistungsanstieg zu erzielen.

V. Schlussfolgerung

TwinBank liefert ganz einfach eine PC Grafik-Speicherlösung ohne Kompromisse, die es jedem erlaubt, Vorteil aus den aufwendigsten aktuellen Applikationen zu ziehen. Sie können nicht nur mehrere Applikationen gleichzeitig ausführen, sondern dies auch ohne Angst vor einem Stocken des Systems tun. Angefangen bei 3D-Gaming bis hin zum Browsen im Web – TwinBank gibt die Power und die Leistung, Hochleistungs-GPUs und den stärksten CPUs die Möglichkeit zu geben, ihre Kapazitäten vollständig einzusetzen.

Als Teil der NVIDIA nForce Plattformprozessor-Architektur liefert TwinBank die kosteneffektivste Systemspeicherarchitektur mit der höchsten Leistung, die bisher erzielt wurde. Durch das Vervierfachen der Bandbreite aktueller 64-bit PC-133 Designs von 1,05GB/Sek bis 4,2GB/Sek und das Ermöglichen eines 30%igen Anstiegs der Bandbreite im Vergleich zu den sehr teuren, komplexen und Hochlatenz-Zweikanal RDRAM-Designs, die in vielen teuren PC-Workstations eingesetzt werden, ist TwinBank die einzige Speicherlösung, die für die nächste Generation der PC-Architekturen wie der NVIDIA nForce Plattformprozessor-Architektur geeignet ist, und definiert die Basislinien auf denen traditionelle SMA-Systeme verglichen werden sollten, völlig neu.

Anhang A - Glossar

DDR SDRAM: Double Data Rate SDRAM

DIMM: Dual In-Line Memory Module – Dual In-Line Speichermodul

DVI: Digital Video Interface. Eine neue Schnittstelle zur Verbindung mit digitalen Monitoren wie Flachbildschirmen, digitalen CRTs und Flachbildprojektoren.

GB/Sek: Gigabyte pro Sekunde

GPU: Grafikprozessor-Einheit Der IGP integriert einen NVIDIA GeForce2™ GPU On-Chip In diesem White Paper sind die Begriffe GPU und Grafikprozessor austauschbar.

Northbridge: Eine Hälfte des kernlogischen PC-Chipsatzes, die eine Schnittstelle zum CPU, GPU, Speicher, AGP und der Southbridge darstellt.

PC-100: 100MHz Standard SDRAM 64-Bit DIMM Systemspeicher.

PC-133: 133MHz Standard SDRAM 64-Bit DIMM Systemspeicher.

PC-800: 800MHz Rambus DRAM RIMM Systemspeicher.

PC-2000 (PC-200): 100MHz Double Data Rate SDRAM 64-Bit DIMM Systemspeicher.

PC-2100 (PC-266): 133MHz Double Data Rate SDRAM 64-Bit DIMM Systemspeicher.

RDRAM: Rambus DRAM.

RIMM: Rambus In-Line Speichermodul.

SDRAM: synchroner DRAM.

SMA: Shared Memory Architecture. Der gesamte installierte Systemspeicher wird zwischen dem Betriebssystem (Windows) des Systems und dem Grafikpuffer (2D, 3D, Video, Textur) geteilt.

Southbridge: Ein Teil des kernlogischen PC-Chipsatzes, der eine Schnittstelle zur Northbridge und verschiedenen Peripheriegeräten darstellt (PCI, IDE ATA-100, USB, Fast Ethernet, Audio/Modem, etc.).

© 2001 NVIDIA Corporation

NVIDIA, das NVIDIA-Logo, TwinBank, nForce und GeForce2 sind eingetragene Warenzeichen oder Warenzeichen der NVIDIA Corporation. Andere Unternehmens- und Produktnamen können

Warenzeichen oder eingetragene Warenzeichen der innehabenden Unternehmen sein.