



Technical Brief

**NVIDIA nForce[®] 790i
SLI[®] Chipsets**

**Reducing Latencies &
Bandwidth Utilization**

Introduction

The NVIDIA nForce 790i SLI chipset features an improved communication protocol which reduces latencies and optimizes bandwidth utilization for CPU-to-GPU and GPU-to-GPU messages. This improved logic consists of two different optimizations that come together to enhance the graphics experience and overall system performance: direct GPU-to-GPU communication and broadcast support.

GPU-to-GPU Direct Link (PWShort)

Typically, if a GPU needs to communicate with another GPU it has to first relay the message through the PCIe controller which forwards it to the memory controller. The memory controller parses the message and sends it back to the PCIe controller, which finally forwards it to the appropriate GPU (see Figure 1). This uses unnecessary bandwidth in the memory-to-PCIe controller link as well as creating additional latency for messages traveling between GPUs.

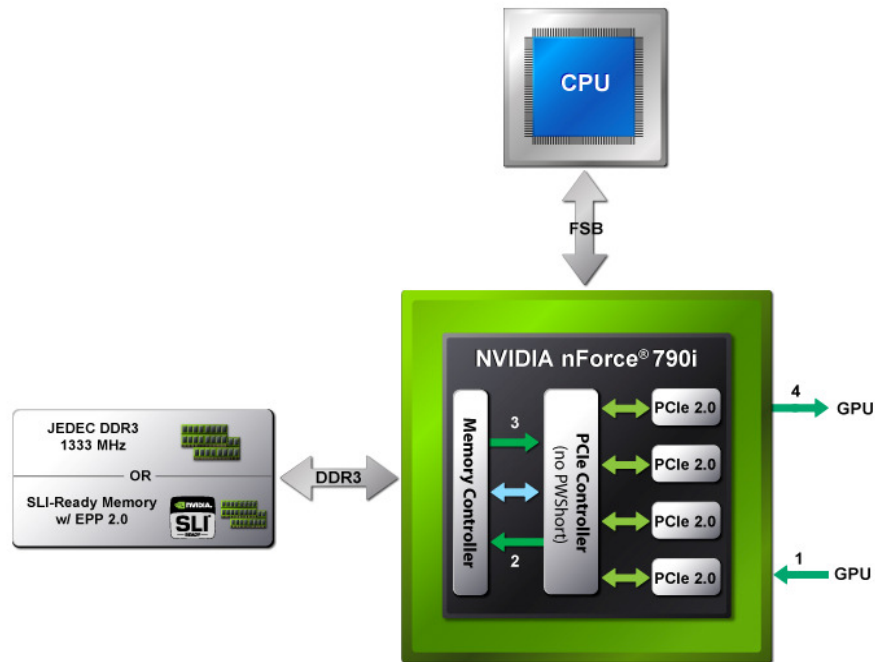


Figure 1 – GPU-GPU Indirect Path

The PCIe controller inside the nForce 790i SLI chipset now has the ability to forward a message from a GPU directly to its destination, a technology we call Posted-Write Shortcut (PWShort). Any nForce 790i-based system running in SLI mode will benefit from this as GPUs often need to send updates to other GPUs to keep their frame buffers synchronized. This point-to-point scheme greatly reduces the latency for traffic between GPUs and alleviates congestion on the memory-to-PCIe controller link (see Figure 2).

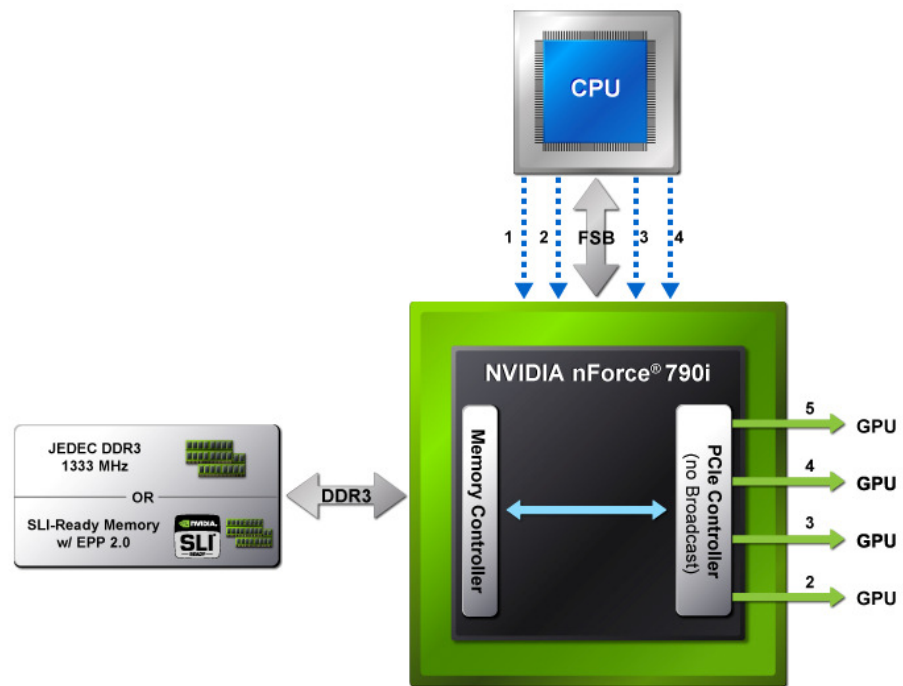


Figure 1 – GPU-GPU Direct Path - PWShort

Broadcast

In systems with multiple GPUs, a CPU often has to send the same data to all the GPUs. For example, all GPUs need to receive the same geometry, texture, and other rendering data from the CPU. In addition to the GPU-GPU direct link, the nForce 790i SLI now has the ability to broadcast CPU commands and data to all GPUs. Instead of serially sending the same data and commands to each GPU (Figure 3), only one message is sent across the frontside bus to the chipset, which replicates it in parallel to all GPUs (Figure 4). This optimization greatly reduces congestion across the frontside bus and improves latencies for CPU-to-GPU broadcast messages.

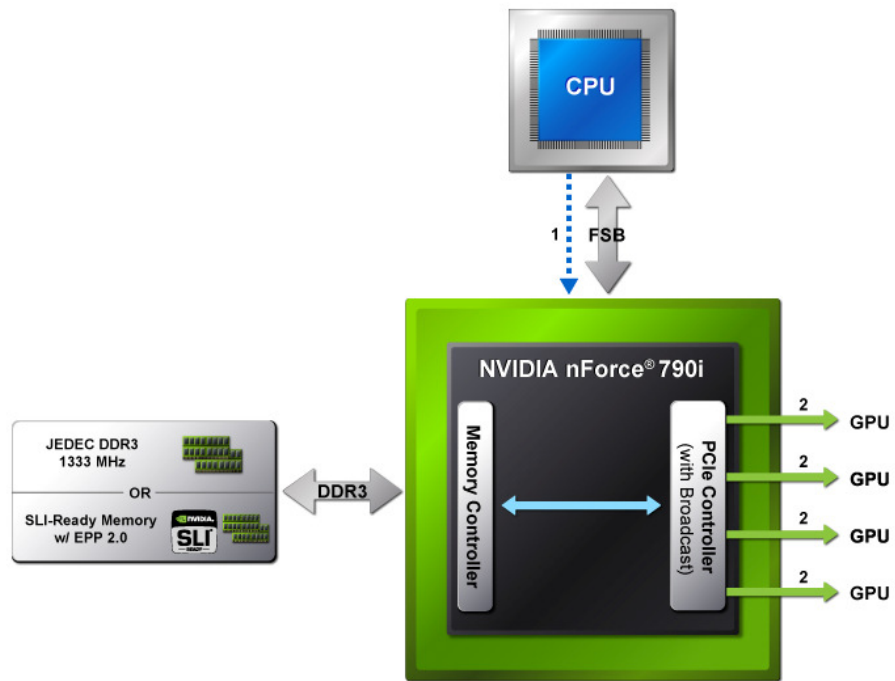


Figure 3 - CPU-to-GPUs without Broadcast

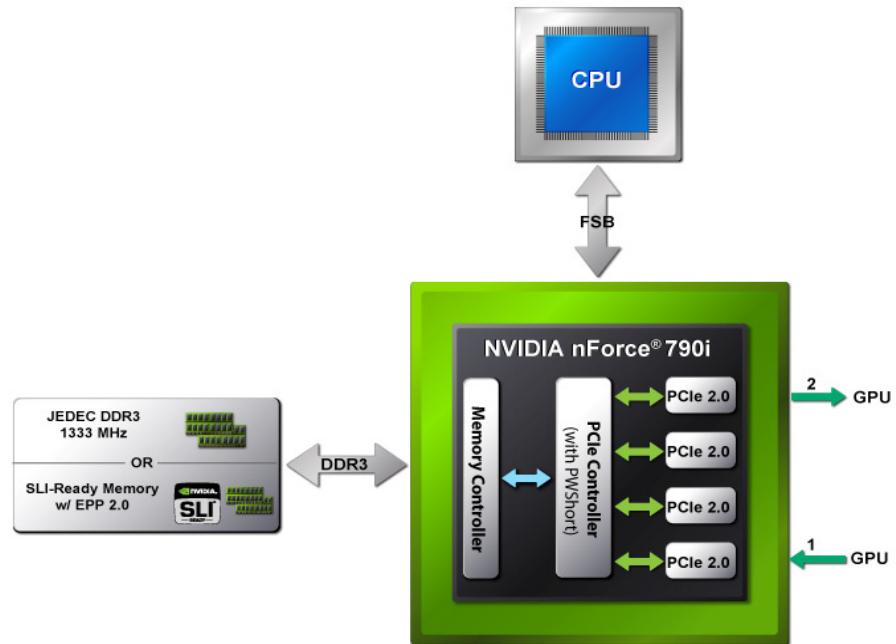


Figure 4 - CPU-to-GPUs with Broadcast

Notice

ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE.

Information furnished is believed to be accurate and reliable. However, NVIDIA Corporation assumes no responsibility for the consequences of use of such information or for any infringement of patents or other rights of third parties that may result from its use. No license is granted by implication or otherwise under any patent or patent rights of NVIDIA Corporation. Specifications mentioned in this publication are subject to change without notice. This publication supersedes and replaces all information previously supplied. NVIDIA Corporation products are not authorized for use as critical components in life support devices or systems without express written approval of NVIDIA Corporation.

Trademarks

NVIDIA, the NVIDIA logo, nForce, and SLI are trademarks or registered trademarks of NVIDIA Corporation in the United States and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

© 2008 NVIDIA Corporation. All rights reserved.



NVIDIA.