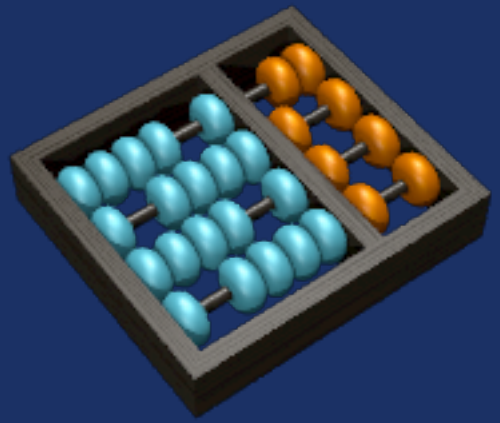
Computer Systems Laboratory



IC-UNICAMP

Introdução à Arquitetura de GPUs

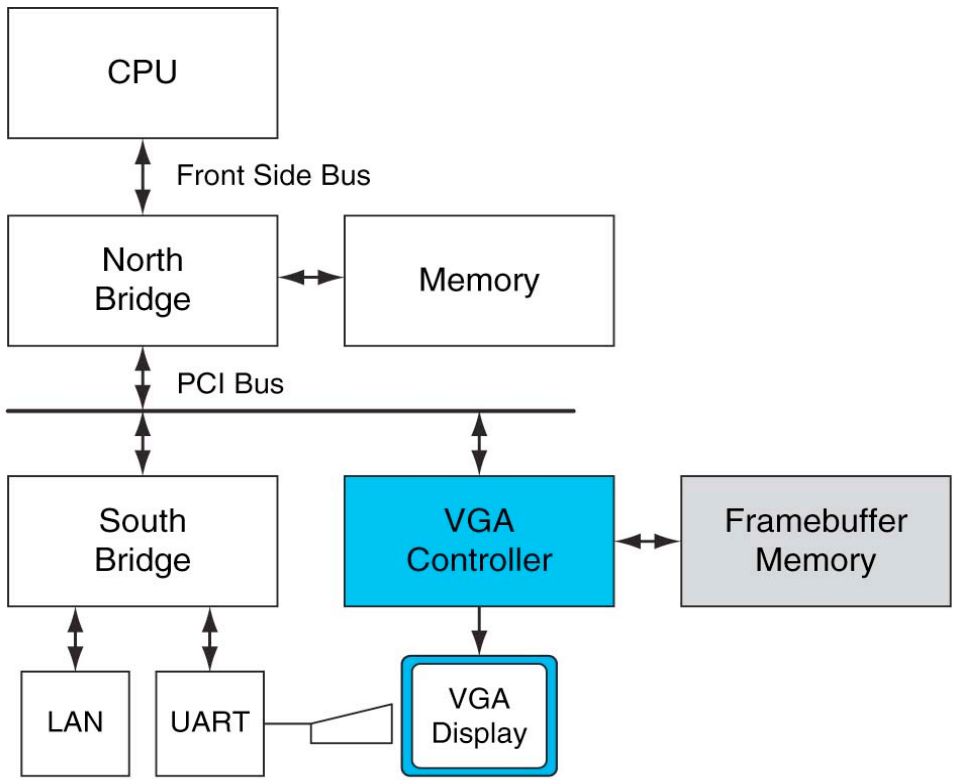
Prof. Sandro Rigo.

UNICAMP

Appendix A

Evolução das GPUs

- 15 anos atrás
  - Gráficos eram gerados por um controlador VGA



- Com o passar dos anos ...
  - Aceleração 3D
  - Rasterização
  - Mapeamentos
- Por volta do ano 2000
  - Aparece o termo GPU
  - O dispositivo gráfico se torna um processador

Hoje em dia ...

- GPU se tornaram programáveis
  - processadores programáveis substituiram lógica dedicada;
  - mais precisão
    - double precision FP
  - massivamente paralelos
    - centenas de cores, milhares de threads
    - é o processador mais poderoso do PC
- GPU+CPU
  - Sistema multiprocessador heterogêneo

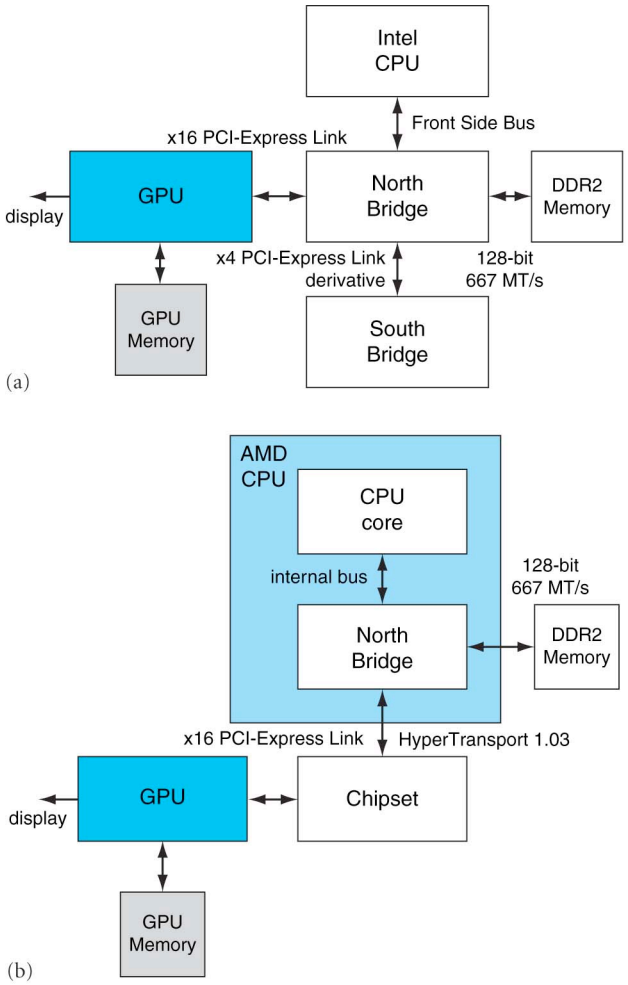


FIGURE A.2.2 Contemporary PCs with Intel and AMD CPUs. See Chapter 6 for an explanation of the components and interconnects in this figure. Copyright © 2009 Elsevier, Inc. All rights reserved.

Arquitetura interna

- Antes:
  - Um pipeline de elementos diferentes especializados
  - Programados por API específica
- Hoje:
  - Elementos especializados substituídos por processadores programáveis
  - GeForce 8-series: processamento de geometria, vértices e pixels executados no mesmo tipo de processador

Arquitetura interna

- Escalabilidade
  - Throughput
  - Balanceamento
- Versatilidade
  - Permite execução de aplicações de propósito geral
    - GPU Computing

Interfaces e Drivers

- Conexão com a CPU
  - PCI, AGP
  - Hoje em dia: PCI-Express
- Aplicações
  - Usam GPU como coprocessador
  - APIs: OpenGL ou DirectX (Direct3D)
  - APIs usam device drivers para enviar dados, programas e comandos

Pipeline Lógico

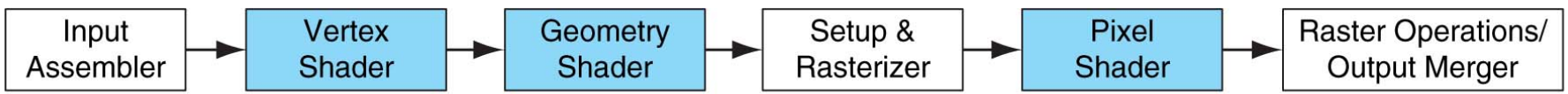


FIGURE A.2.3 Graphics logical pipeline. Programmable graphics shader stages are blue, and fixed-function blocks are white.

Mapeamento Físico

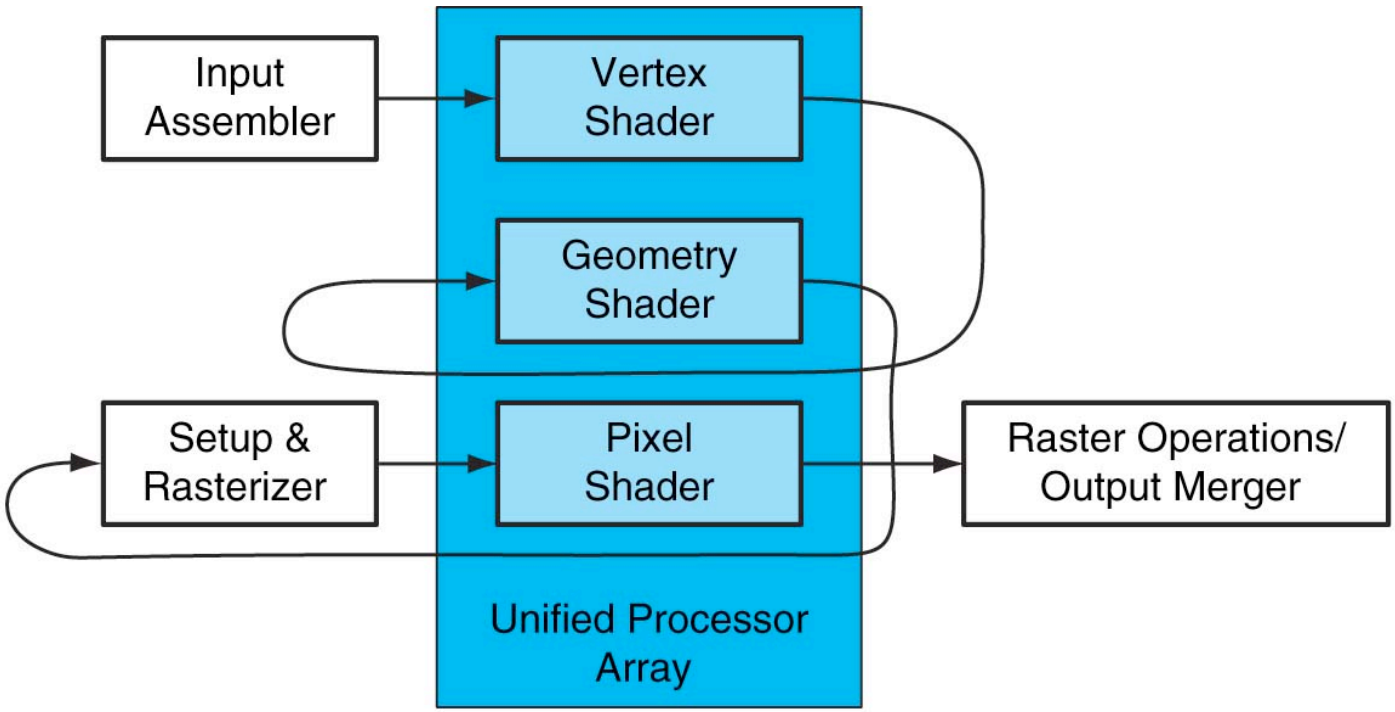


FIGURE A.2.4 Logical pipeline mapped to physical processors. The programmable shader stages execute on the array of unified processors, and the logical graphics pipeline dataflow recirculates through the processors.

Unified GPU Processor Array

- Muitos cores (many cores)
  - organizados como multithreaded multiprocessors
- Exemplo (GForce 8800 (Tesla)):
  - 112 streaming processor cores (SP)
    - Formam 14 Streaming multiprocessors (SM)
  - Cada core SP gerencia até 96 threads em hardware
  - 4 64-bit DRAM
  - SM: 8 SPs, 2 special function units (SFU), I-Cache, C-Cache, MT Instr. unit, Shared Mem

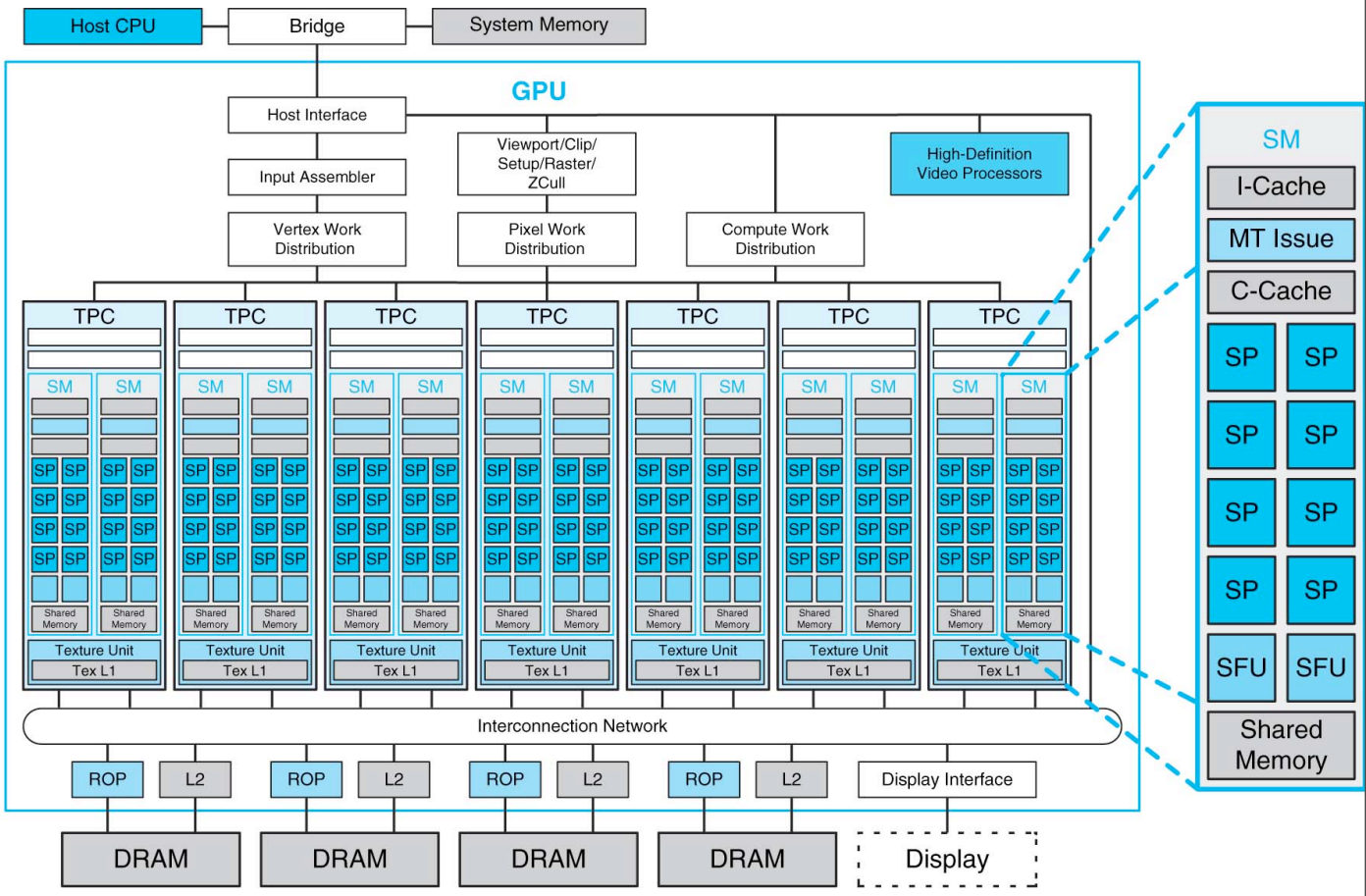


FIGURE A.2.5 Basic unified GPU architecture. Example GPU with 112 streaming processor (SP) cores organized in 14 streaming multiprocessors (SMs); the cores are highly multithreaded. It has the basic Tesla architecture of an NVIDIA GeForce 8800. The processors connect with four 64-bit-wide DRAM partitions via an interconnection network. Each SM has eight SP cores, two special function units (SFUs), instruction and constant caches, a multithreaded instruction unit, and a shared memory.

Unified GPU Processor Array

- 7 clusters (TPC) de 2 SMs
  - 1 Texture unit
  - 1 texture L1 cache
- Conexões do array
  - ROP: raster operation processors
  - L2 texture caches
  - external DRAM
- Número de processadores e memórias pode ser balanceado de acordo com mercado alvo
  - Varia o número de TPCs

Referências

- Computer Organization & Design, Patterson e Hennessy. 4a ed. Apêndice A. Morgan Kaufmann.

- Lindholm, E.; Nickolls, J.; Oberman, S.; Montrym, J.; , "NVIDIA Tesla: A Unified Graphics and Computing Architecture," Micro, IEEE , vol.28, no.2, pp.39-55, March-April 2008
  - doi: 10.1109/MM.2008.31