МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(национальный исследовательский университет)

Факультет прикладной математики и физики
Кафедра вычислительной математики и программирования

Лекция 2

«Основная терминология курса: шейдер, SM, ROP, TPC, SP. Типы параллельных архитектур: SISD, MISD, SIMD, MIMD, DSP»

Выполнил: Семенов С.А.
Руководитель: Ревизников Д.Л.

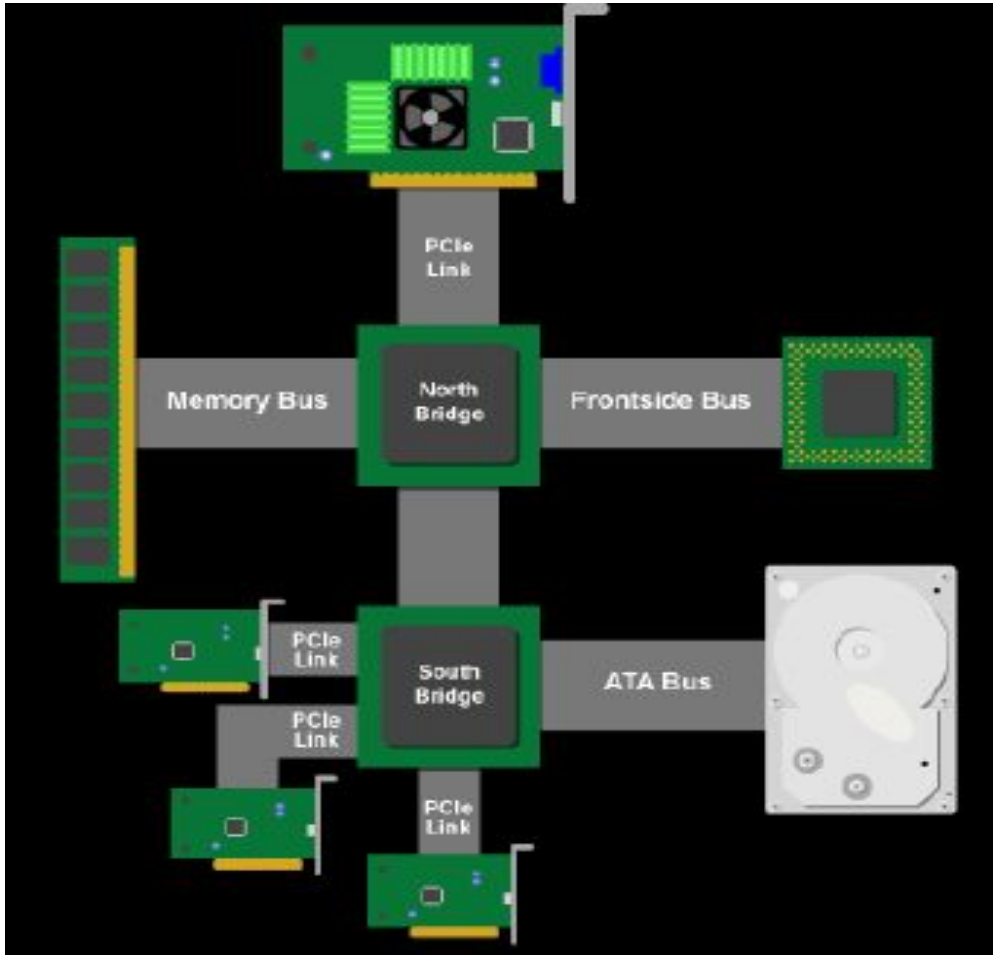
Введение

- Схематическое изображение графического адаптера
- Классификация вычислительных систем по Флинну
- Схематическое устройство SMP
- Multithreading
- Bottleneck

Графическая плата NVIDIA

Средства обмена данными в компьютере

- Обмен данными – важнейшая составляющая компьютера
  Примеры: многопроцессорные системы, FPGA etc.
- По традиции отдельные устройства имеют разные возможности (уровни и способы) обмена данными
- Традиционная архитектура ориентирована на одно, центральное счётное устройство

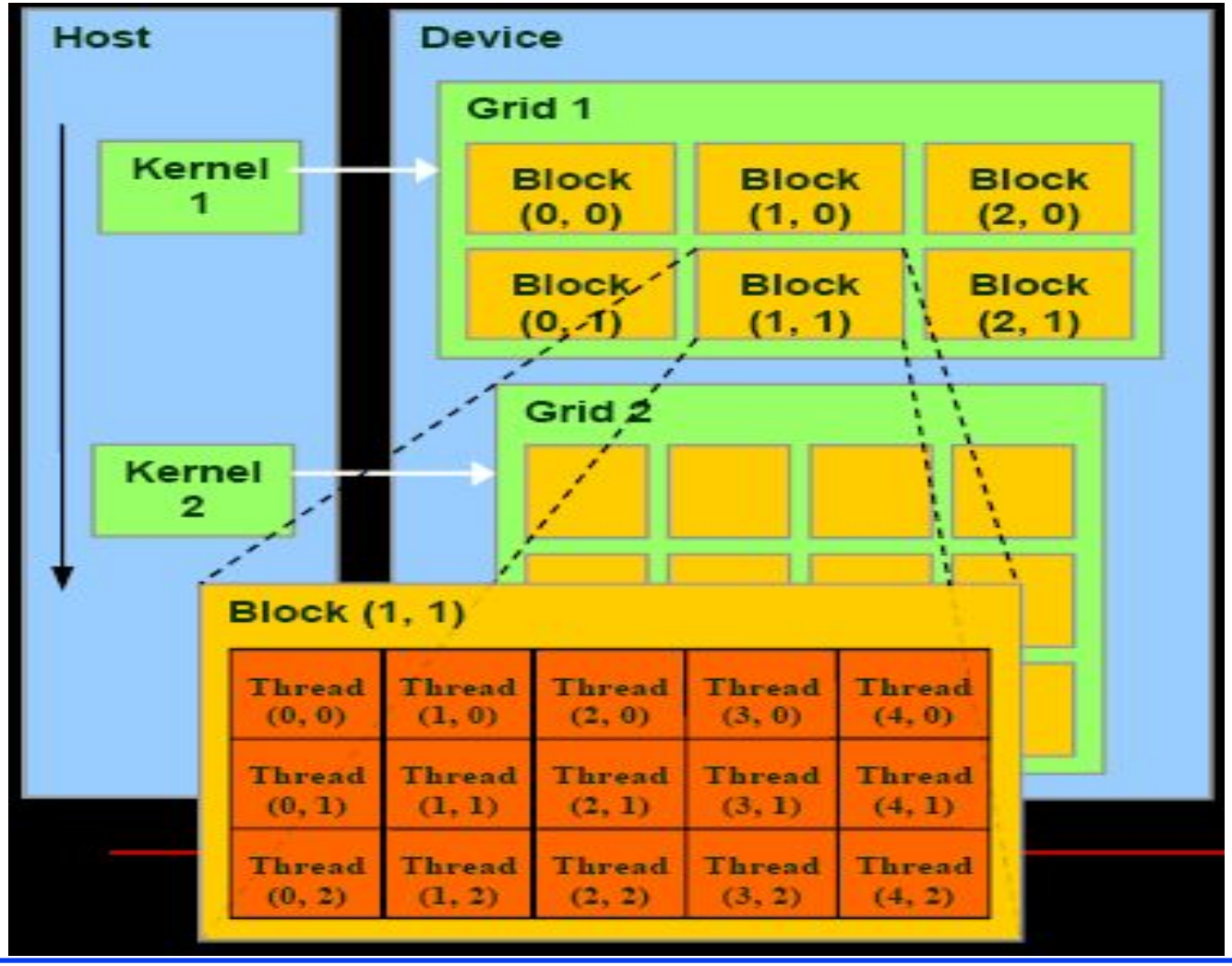


Программная часть технологии CUDA

Введем основные термины и отношения между ними [CUDA C Best Practices, 2010].

- Хост (Host) — центральный процессор, управляющий выполнением программы.
- Устройство (Device) — видеоадаптер, выступающий в роли сопроцессора центрального
- процессора.
- Грид (Grid) — объединение блоков, которые выполняются на одном устройстве.
- Блок (Block) — объединение тредов, которое выполняется целиком на одном SM. Имеет
- свой уникальный идентификатор внутри грида.
- Тред (Thread, поток) — единица выполнения программы. Имеет свой уникальный
- идентификатор внутри блока.
- Варп (Warp) — 32 последовательно идущих треда, выполняется физически одновременно.
- Ядро (Kernel) — параллельная часть алгоритма, выполняется на гриде.

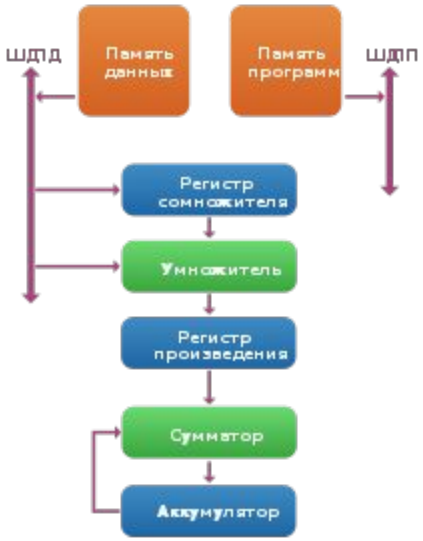
Схематическое изображение графического адаптера

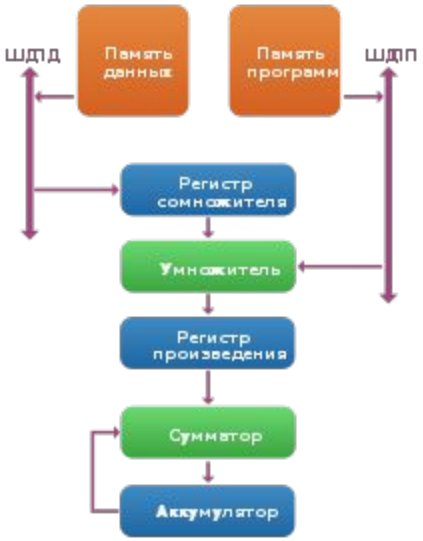


DSP

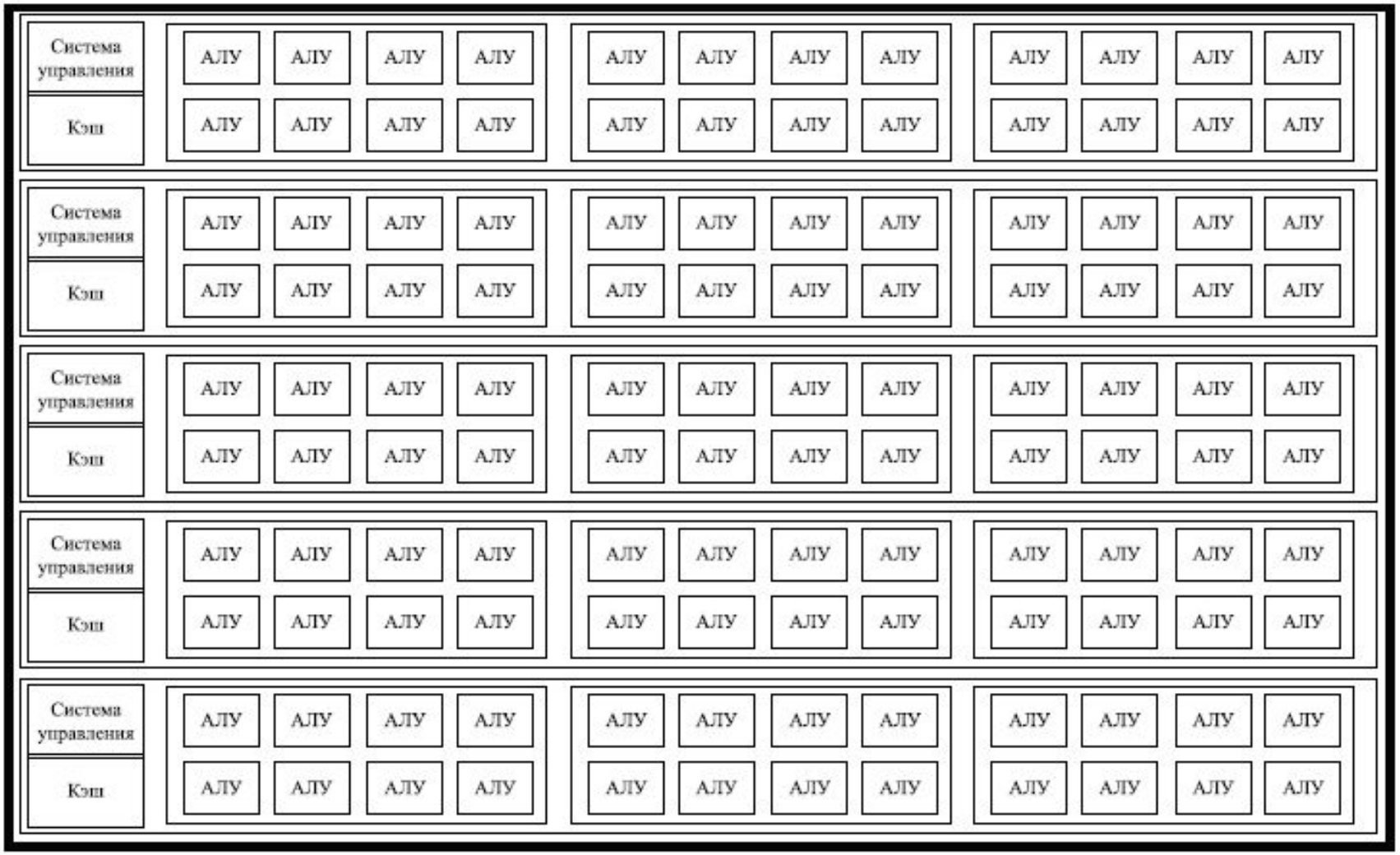
Цифровой сигнальный процессор (англ. *Digital signal processor*, *DSP*; сигнальный микропроцессор, СМП; процессор цифровых сигналов, ПЦС) — специализированный микропроцессор, предназначенный для цифровой обработки сигналов (обычно в реальном масштабе времени).

ЦСП строятся на основе Гарвардской архитектуры

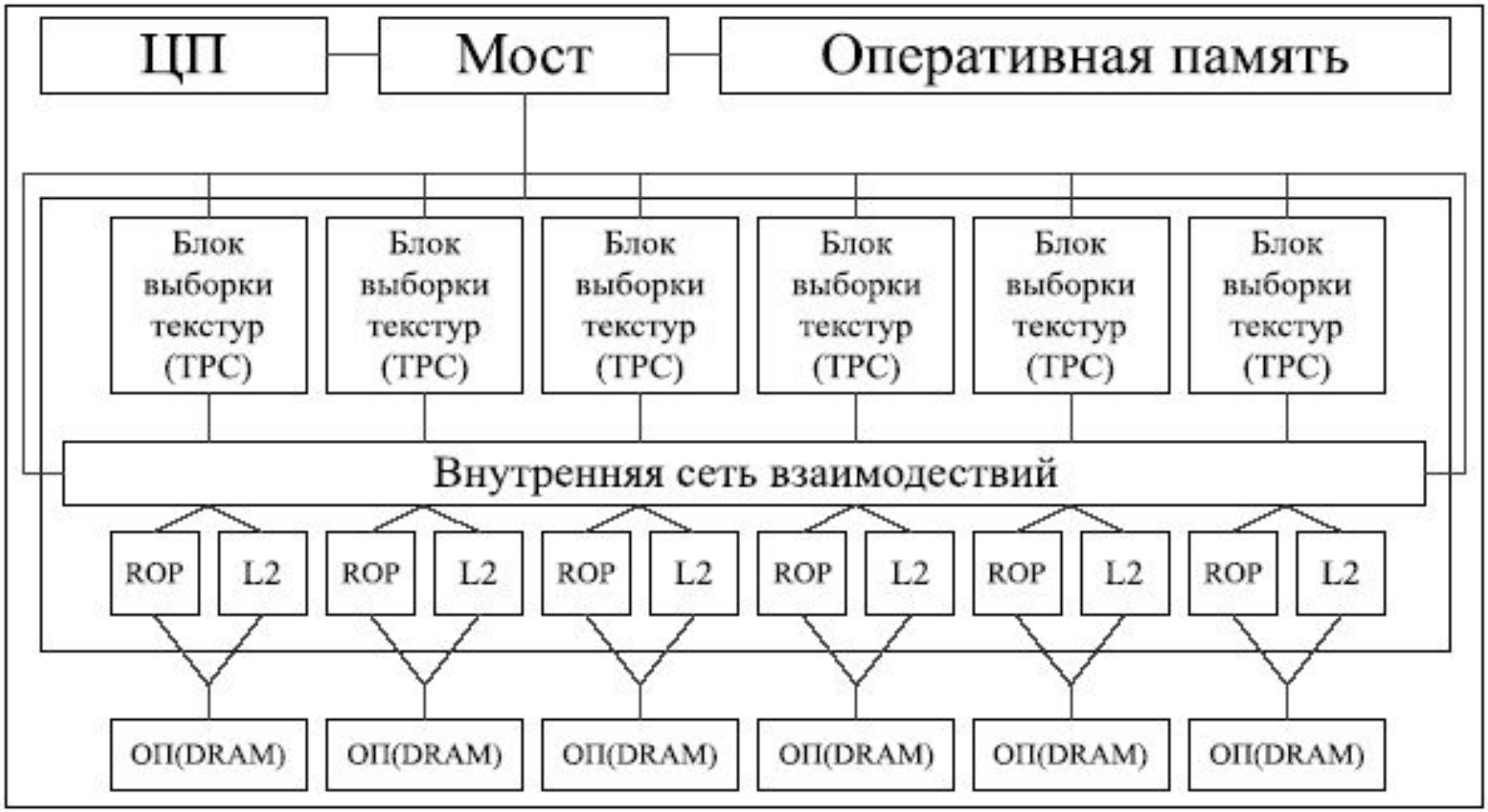




Стандартные ЦСП

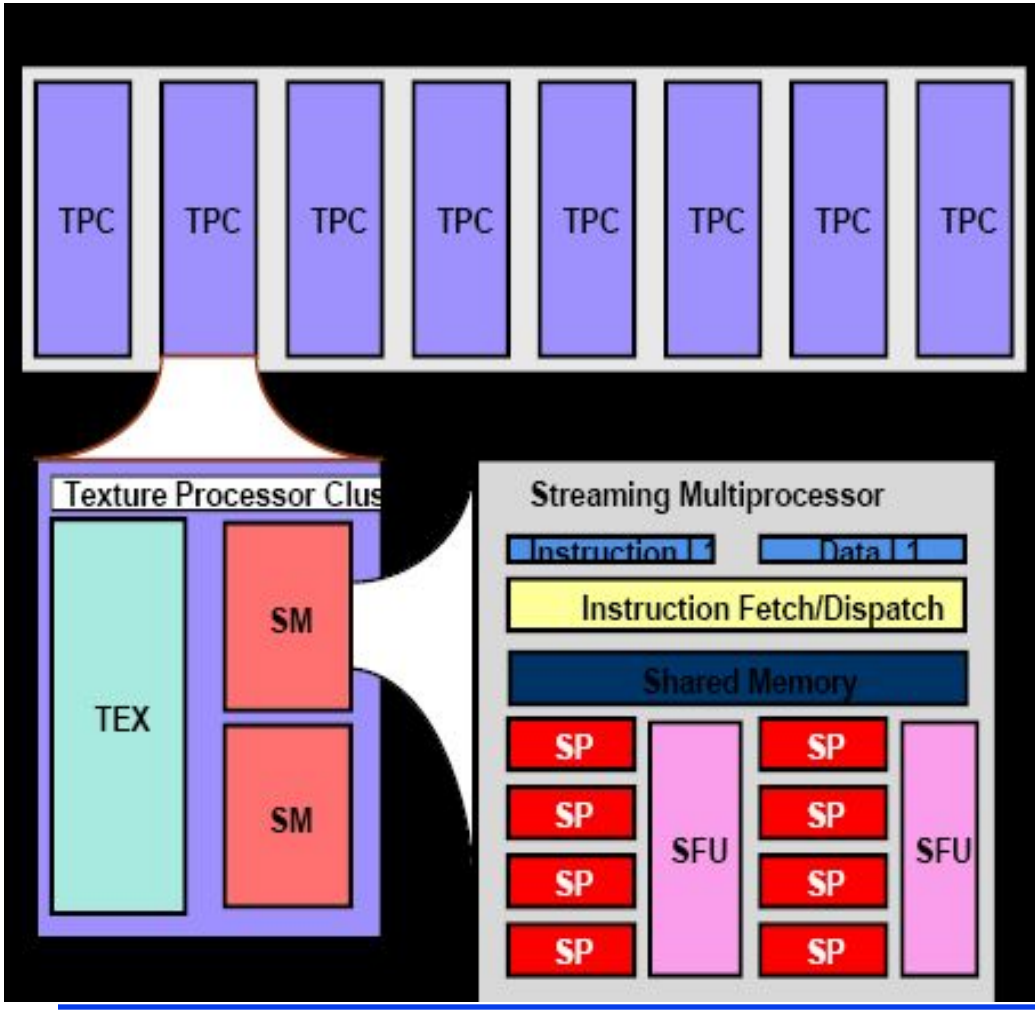

Система управления
Кэш
АЛУ
Оперативная память

Схематическое изображение устройства графического адаптера



Схематические особенности видеочипа

Схематическое расположение блоков GPU



- SPA - Streaming Processor Array
- TPC - Texture Processor Cluster
- SM - Streaming Multiprocessor
  - Multi-threaded processor core
  - Fundamental processing unit for CUDA thread block
- SP - Streaming Processor
  - Scalar ALU for a single CUDA thread

Графический адаптер на «аппаратном» уровне

- TPC (Texture process cluster)
- ROP — Raster Operations Pipeline
- SP (Streaming Processor)
- SM (Streaming Multiprocessor)
- SFU (Super Function Unit)
- WS — Warp Scheduler

Классификация вычислительных систем по Флинну

	Single Instruction (Одна инструкция)	Multiple Instruction (Много инструкций)
Single Data (Одиночные данные)	SISD	MISD
Multiple Data (Множественные данные)	SIMD	MIMD

Классы систем

Классификация систем

- CPU – SISD
  - Multithreading: позволяет запускать множество потоков – параллелизм на уровне задач (MIMD) или данных (SIMD)
  - SSE: набор 128 битных регистров ЦПУ
    - можно запаковать 4 32битных скаляра и проводить над ними операции одновременно (SIMD)
- GPU – SIMD*

Звездочка стоит для того, чтобы вы обратили внимание.

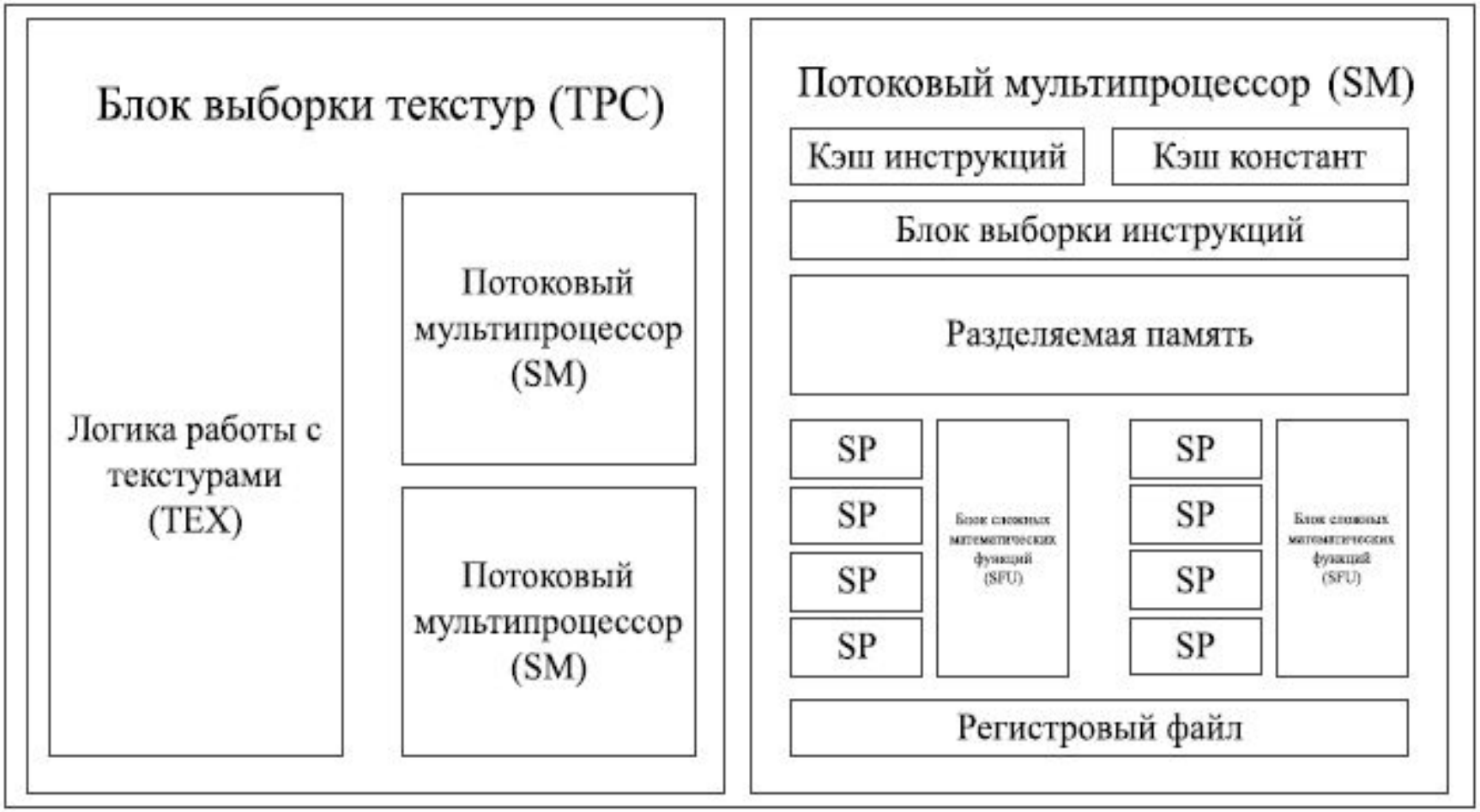
На следующих лекциях вы увидите, что GPU не совсем SIMD архитектура а скорее SIMT (simultaneous multithreading):

* разные блоки могут выполнять разный код (без потери производительности)
* внутри одного блока можно выполнять разный код (с потерей производиетльности)

SIMT (Single instruction, multiple threads)

- Параллельно на каждом SM выполняется большое число отдельных нитей (threads)
- Нити подряд разбиваются на warp (по 32 нити) и SM управляет выполнением warp
- Нити в пределах одного warp выполняются физически параллельно
- Большое число warp покрывает латентность

Схематическое изображение устройства TPC и SM

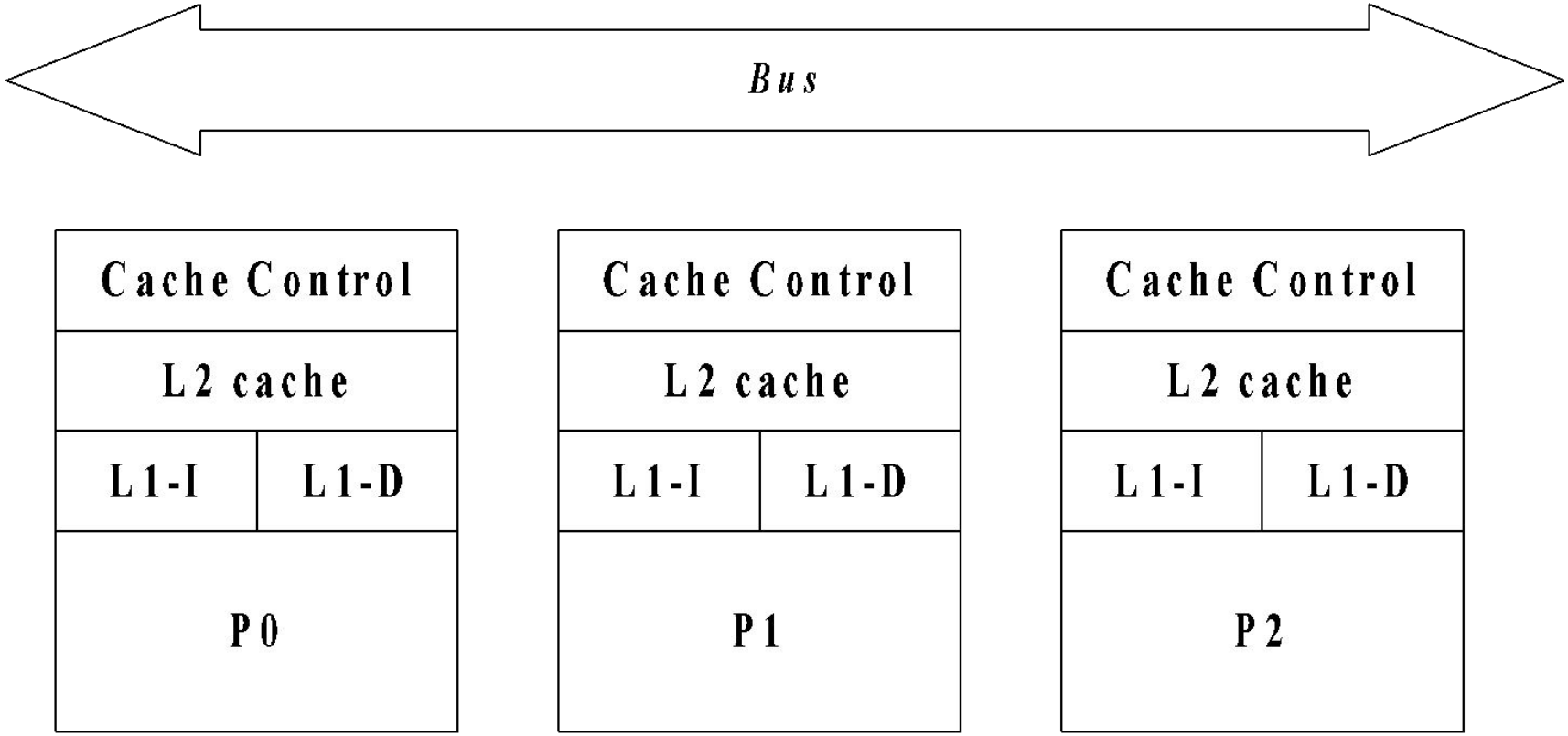


Symmetric Multiprocessor Architecture (SMP)

Каждый процессор

- ✔ имеет свои L1 и L2 кэши
- ✔ подсоединен к общей шине
- ✔ **отслеживает доступ других процессоров к памяти** для обеспечения единого образа памяти (например, один процессор хочет изменить данные, кэшированные другим процессором)

Symmetric Multiprocessor Architecture (SMP)

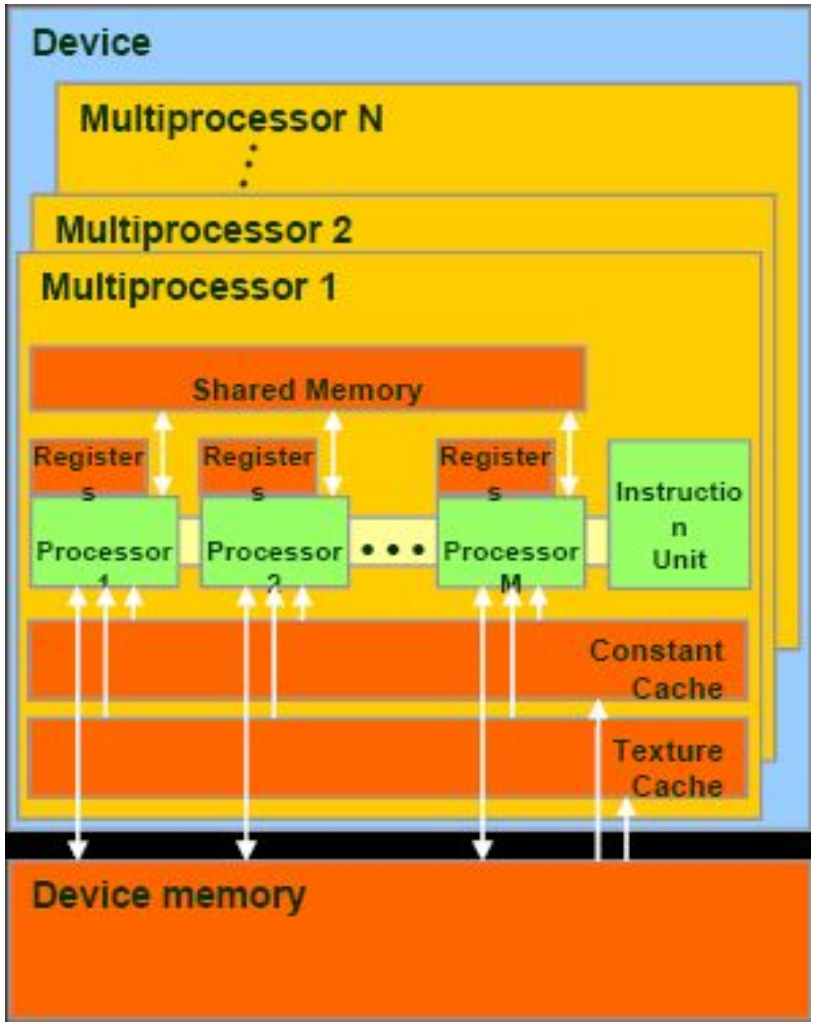


Программная модель CUDA

- Параллельная часть кода выполняется как большое количество нитей (*threads*)
- Нити группируются в блоки (*blocks*) фиксированного размера
- Блоки объединяются в сеть блоков (*grid*)
- Ядро выполняется на сетке из блоков
- Каждая нить и блок имеют свой уникальный идентификатор

Что такое ВОРП (WARP)?

- Device делает 1 grid в любой момент
- SM обрабатывает 1 или более blocks
- Каждый Block разделён на SIMD группы, внутри которых одни и те же инструкциивыполняются реально одновременно над различными данными (warps) warp size=16/32
- Связывание в ворпы детерминировано в порядке нарастания threadID
- threadID=TIDX.x+TIDX.y*Dx+TIDX.z*Dx*Dy
- **Полуворп** – первая или вторая половина ворпа



Итоги лекции

В результате лекции Вы должны :

- **Понимать возможности использования GPU для расчётов с точки зрения пропускной способности системы обмена данными компьютера**
- **Иметь понятие об организации разработки приложений**
- **Достаточные знания для начала самостоятельной работы**