Терминология

- Параллельное программирование – способы организации одновременного выполнения нескольких вычислительных процессов

- Связанные термины: распределенные вычисления, асинхронные вычисления, конкурентные вычисления, HPC

Мотивация

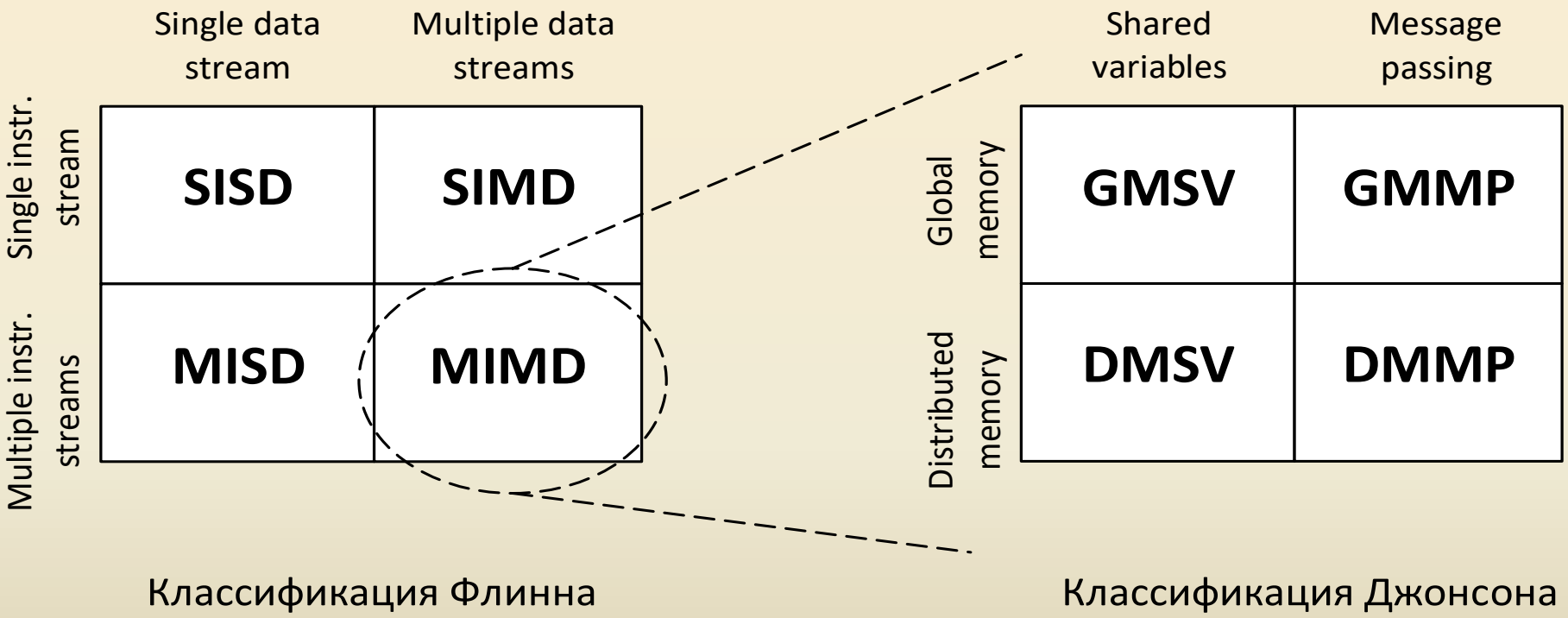
- Повышение эффективности обработки вычислительно-емких задач
- Улучшение «отзывчивости» приложения
- Организация одновременного доступа к ресурсам

Актуальность параллельных вычислений

- Развитие многопроцессорной многоядерной архитектуры

  Ограничения увеличения частоты процессоров из-за энергопотребления
- Развитие сетевых технологий, обеспечивающих распределенную работу приложений
- Развитие технологий виртуализации и облачных вычислений
- Развитие архитектуры массивно-параллельных систем (GPU)

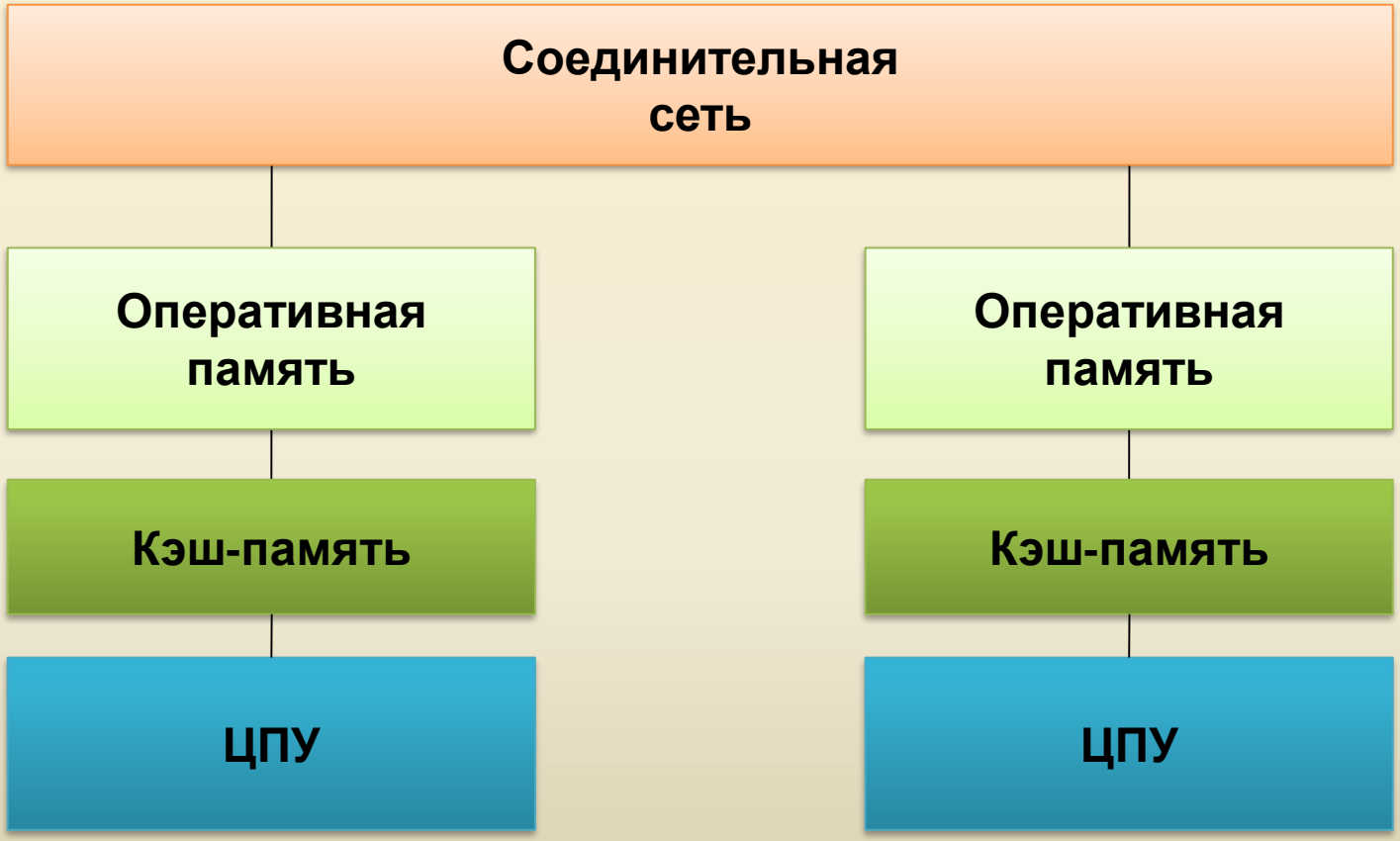
Классификация вычислительных систем



Классификация Флинна

Классификация Джонсона

Распределенные системы



Многопроцессорные системы

Системы с общей памятью

Достоинства: простота программирования

Недостатки: плохая масштабируемость

Системы с общей памятью

Классификация

- Симметричные мультипроцессоры и ассиметричные мультипроцессоры
- С поддержкой гиперпоточности и без поддержки гиперпоточности
- С равномерным доступом к памяти и неравномерным доступом к памяти

Симметричные мультипроцессоры

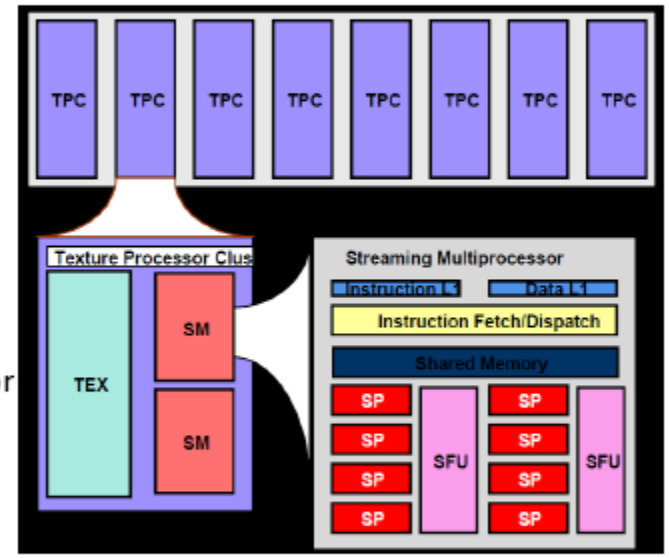
- Symmetric Multi-processor (SMP)
- Включает несколько равнозначных процессоров или ядер процессора

Ассиметричные мультипроцессоры

- Включает одно главное управляющее ядро и несколько специализированных ядер

Графические процессоры

- Архитектура G80
  - SPA – Массив потоковых процессоров (8 x TPC)
  - TPC – Кластер процессоров текстур (2x SM + TEX)
  - SM – Потоковый мультипроцессор (8 x SP)
    - Многопоточное процессорное ядро
    - Fundamental processing unit for CUDA thread block
  - SP – Простой потоковый процессор
    - Скалярное АЛУ для одного потока CUDA
  - SFU – специальный процессор для сложных функций



NUMA-системы

- Non-uniform multiprocessor architecture
- Мультипроцессорные системы с неравномерным доступом к памяти
- Как правило, состоят из большого числа вычислительных устройств (>=256)
- Распределенная память рассматривается ВУ как единое адресное пространство

Shared-systems

- Единое физическое адресное пространство
- Каждый процессор (вычислительное устройство) обладает своей кэшируемой памятью, которая обеспечивает быстрый доступ к данным
- Кэш-память может содержать частные и разделяемые данные
  - Частные данные (private) – используются только одним процессором
  - Разделяемые данные (shared) – используются несколькими процессорами
- Данные перемещаются в блоках (кэш-линии)
- Кэш-контроллер проверяет содержатся ли необходимые данные в кэше (cashe-hit) или нет (cashe-miss)

Проблема когерентности кеш-памяти

- Проблема согласованности (когерентности) кэш-памяти возникает, когда значения переменной в оперативной памяти, в кэш-памяти разных процессоров различаются
- В современных системах когерентность кэш-памяти поддерживается автоматически (неявно для программиста)

Time	Shared memory	Caches				Comment
		C_0	C_1	C_2	C_3	
0	b	b	—	—	—	Block b is loaded in C_0.
1	b	b	b	—	b	Block b is loaded in C_1 and C_3.
2	b	b	b	—	b_3	Processor P_3 modifies its copy of b. Now the system is noncoherent.
3	b_3	b	b	—	b_3	Processor P_3 performs a write-through. The system is noncoherent since C_0 and C_1 have different copies.
4	b_3	b_3	b_3	—	b_3	Shared memory controller updates C_0 and C_1. Now the system is coherent.

Проблема ложного разделения кэш-памяти

- false-sharing
- Проблема возникает, когда процессоры работают с разными данными, расположенными физически близко

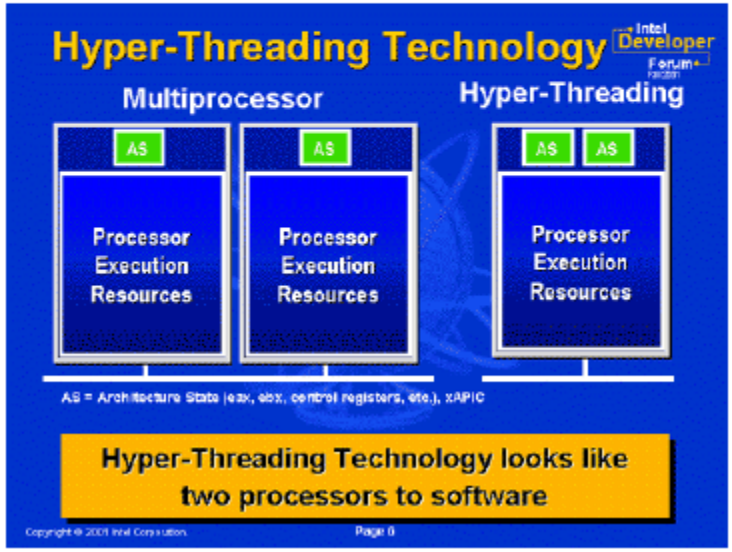
Simultaneous multithreading

- Одновременная многопоточность (также hypethreading) – аппаратная поддержка возможности выполнения нескольких потоков (как правило двух) на одном процессоре или ядре процессора.
- Физический процессор (ядро) состоит из двух логических ядер.
- Каждое логическое ядро включает собственный набор регистров и контроллер прерываний
- Выигрыш от гиперпоточности – 5 – 30% в зависимости от особенностей организации потоков

Гиперпоточность

Процессоры с поддержкой гиперпоточности

- Поддержка:
  - Intel Pentium 4
  - Intel Core i3/i5/i7
  - Intel Xeon



Уровни параллелизма

1. Уровень битов (bit-level parallelism) Разрядность процессора: 8-, 16-, 32-, 64-,128-

2. Уровень инструкций (instruction level parallelism).

3. Уровень потоков

4. Уровень процессов