МультиТредовые архитектуры

Пути увеличения производительности

- Наращивание тактовой частоты
- Реализация параллелизма на уровне команд (ILP)
- Реализация параллелизма на уровне нитей (TLP)

Многопоточность

- **Многопоточность** – это свойство платформы или программы, позволяющее процессу состоять из нескольких потоков команд, исполняющихся параллельно, без предписанного порядка во времени.
- **Цель** – более эффективное использование ресурсов компьютера

Многопоточное программирование

- Все потоки находятся в едином адресном пространстве процесса, т.е. имеют общую память
- Создание потока и переключение между ними менее трудоемко, чем аналогичные операции над процессами
- Исключение некорректной работы с памятью обеспечивается использованием средств синхронизации

Средства многопоточной синхронизации

- **Mutex** (Mutual Exclusion Lock) – объект, который может быть захвачен одновременно только одним потоком. После выполнения необходимых действий мьютекс освобождается.
- **Семафор** – объект, который может быть захвачен одновременно ограниченным числом потоков.
- **Событие** – объект, сообщающий ожидающему потоку (или потокам) о возникновении какого-либо события

Пример использования нитей

```
void mul(int start, int end)
{
    for (int i=start; i<end; i++)
    c[i] = a[i] + b[i];
}
```

```
mul(0, N);
```

```
CreateThread(mul, 0, N/2);
CreateThread(mul, N/2, N);
```

Аппаратная поддержка многопоточности

- Coarse-grained MT (крупнозернистая многопоточность)
- Fine-grained MT (мелкозернистая многопоточность)
- SMT (Одновременная многопоточность)

Крупнозернистая многопоточность (Coarse-grained MT)

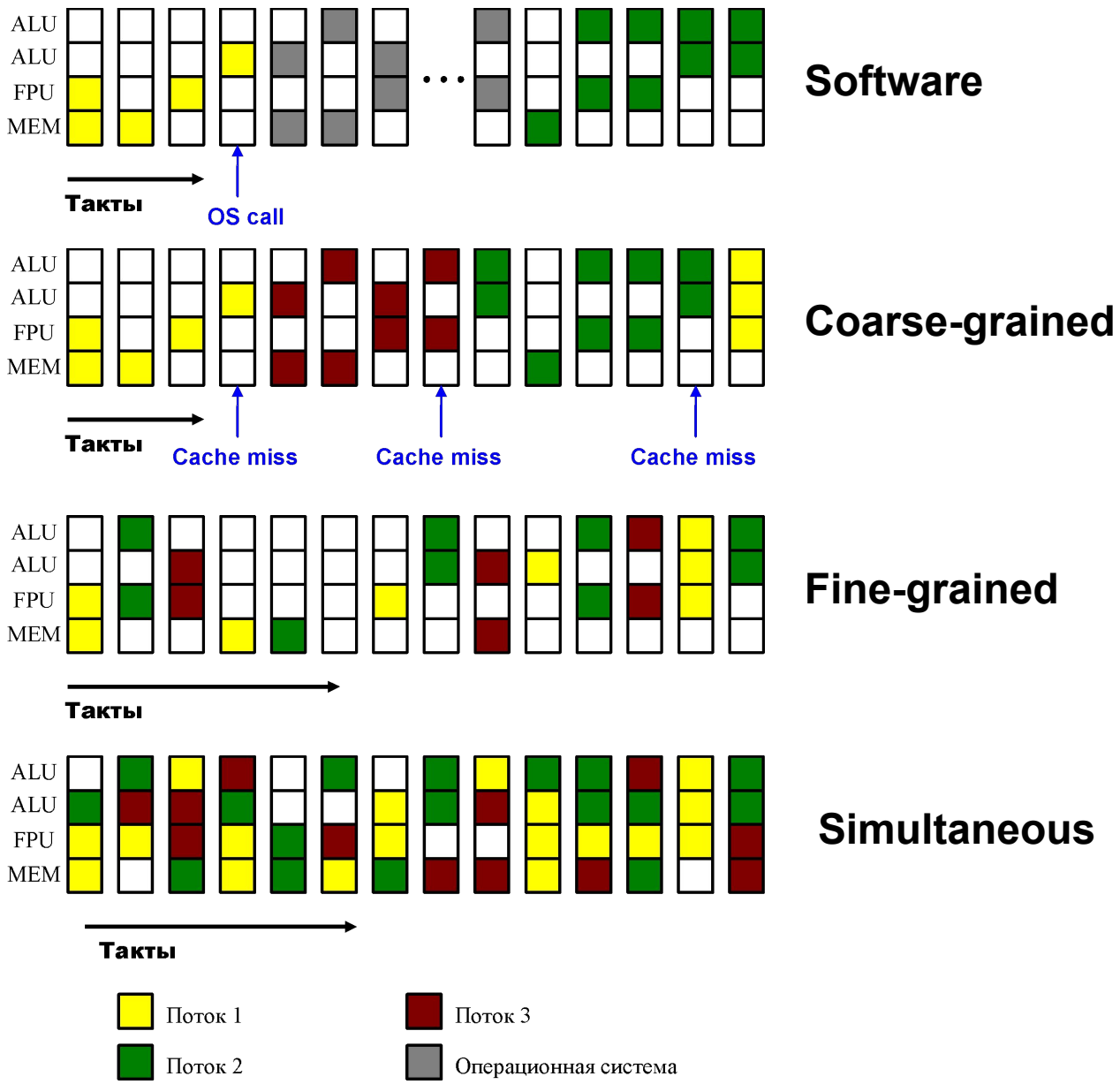
- 2 и более аппаратных контекстов
  - Регистры общего назначения
  - Счетчик команд
  - Буфер выборки инструкций
- Одновременно не более 1 нити
- Аппаратное переключение контекстов при прерывании

Мелкозернистая многопоточность (Fine-grained MT)

- 2 и более аппаратных контекстов
  - Регистры общего назначения
  - Счетчик команд
- Одновременно не более 1 нити
- Аппаратное переключение на каждом такте

Одновременная многопоточность (Simultaneous MT)

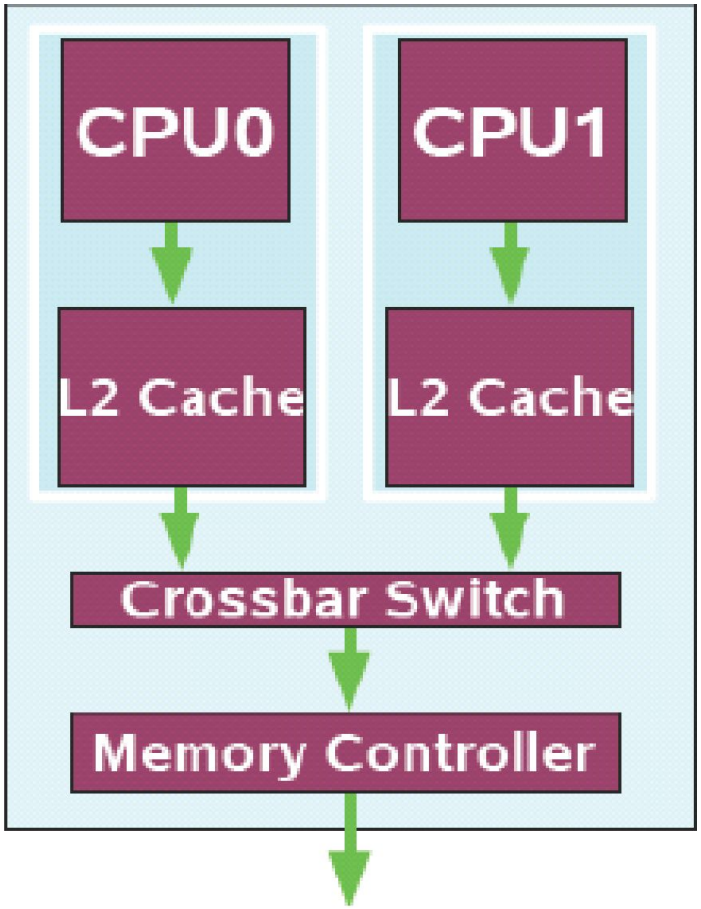
- 2 и более аппаратных контекстов
  - Регистры общего назначения
  - Буфер выборки инструкций
  - Буфер переупорядочивания
  - Стек возврата
- Привязка команд и нитей
- До нескольких нитей одновременно
- Контексты активны, переключения не происходит

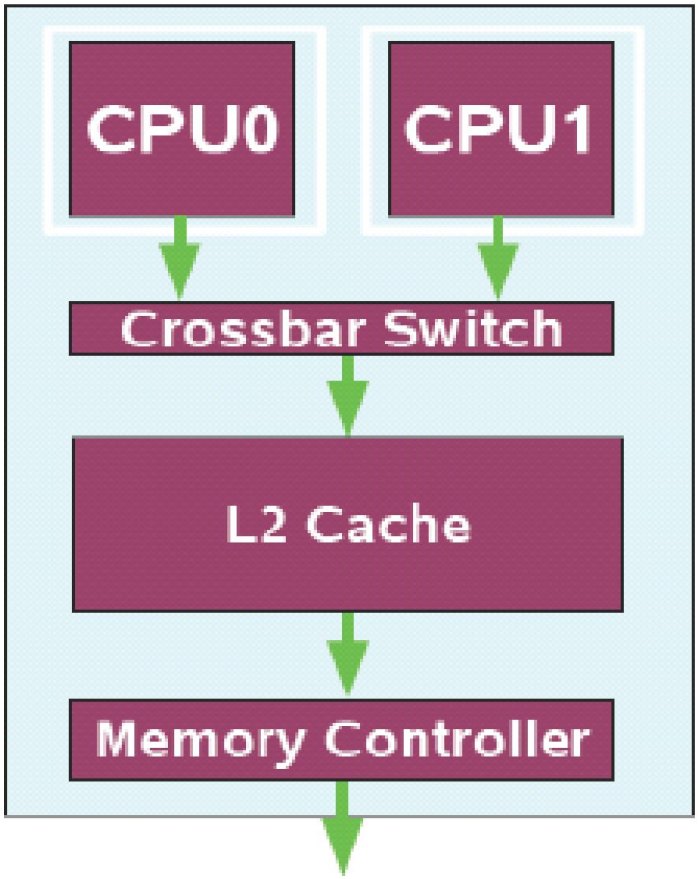

ALU
ALU
FPU
MEM
Software
Такты
OS call
Coarse-grained
Cache miss
Cache miss
Cache miss
Fine-grained
Simultaneous
Поток 1
Поток 2
Поток 3
Операционная система

Многоядерность vs многопроцессорность (CMP – Chip Multiprocessors)

- Меньше стоимость при той же производительности
- Выше скорость обмена между ядрами
- Меньше места, меньше выделяемого тепла, меньше потребляемая мощность

Структура CMP-процессора





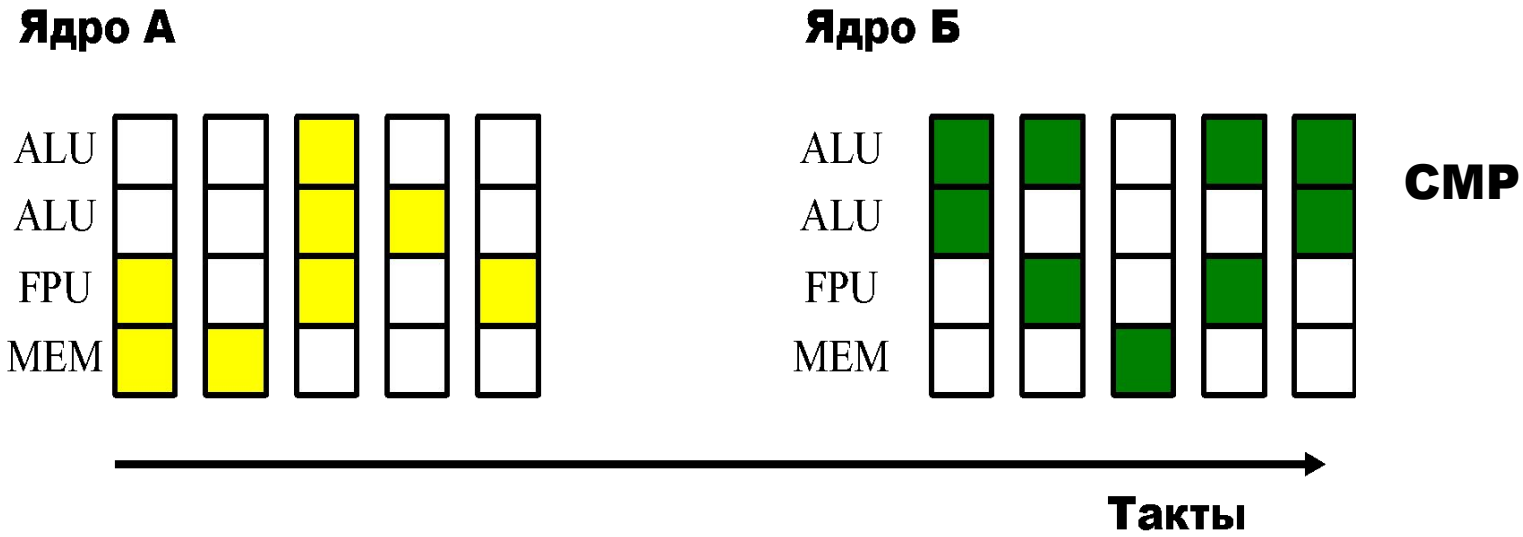
Многопоточность на основе CMP

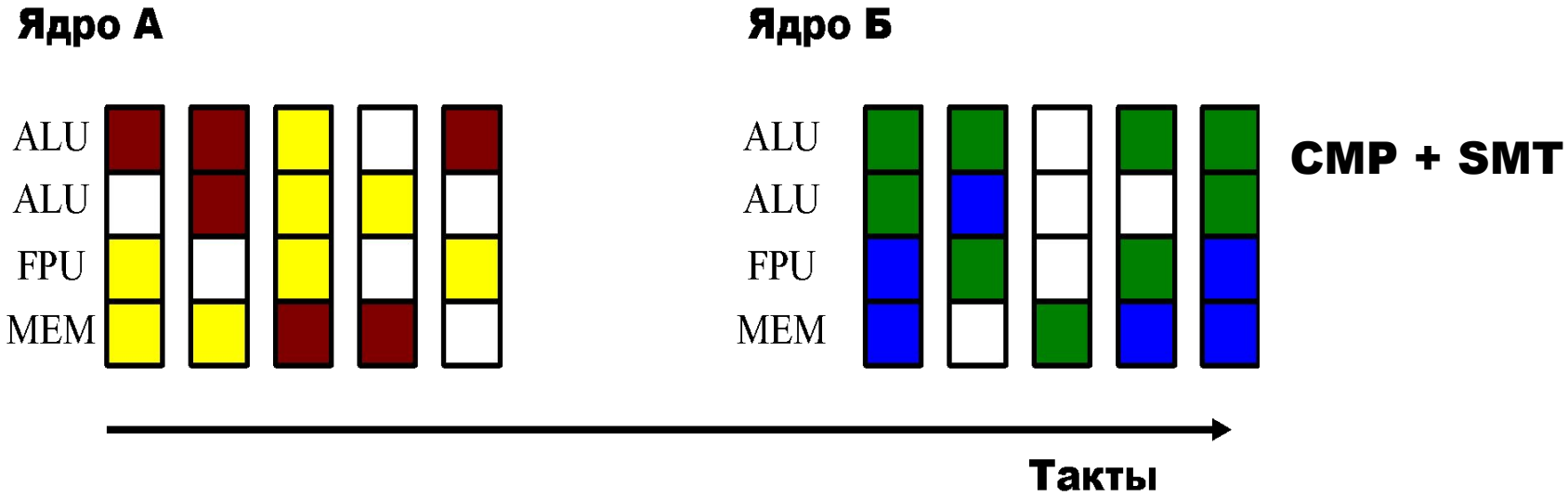
- Простая логика, один поток на ядро
- Масштабируемость за счет локальности
- Условная многопоточность (Speculative MultiThreading) для ускорения последовательного кода

Распределение ресурсов

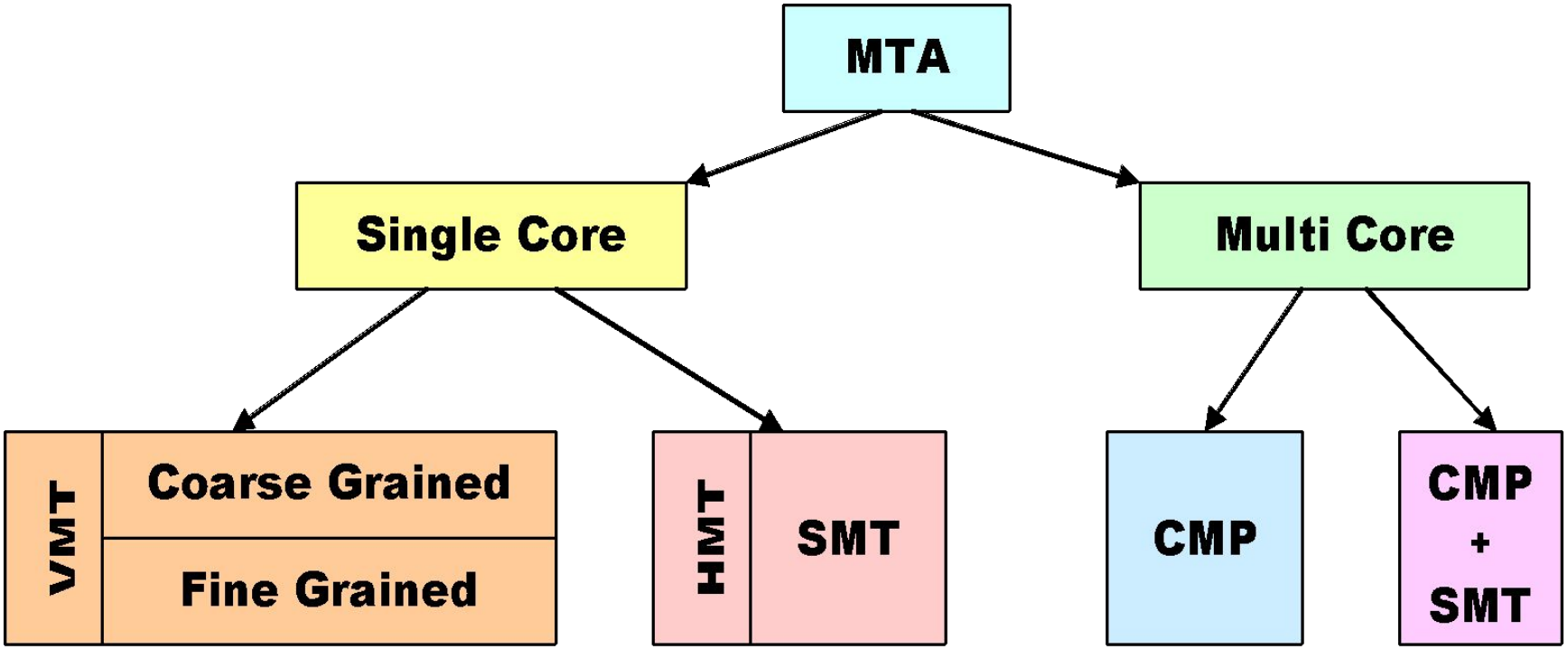
MT	**Общие ресурсы**	**Переключение контекстов**
нет	Все	Явное переключение ОС
Fine-grained	Все кроме регистров и управляющей логики	На каждом такте
Coarse-grained	Все кроме регистров, управляющей логики и буферов выборки инструкций	При остановке конвейера
SMT	Все кроме регистров, управляющей логики, буферов выборки инструкций, стека адресов возврата, буфера переупорядочивания, очереди записи и т.п.	Все контексты активны, переключения нет
CMP	Кэш второго уровня, системное межсоединение ядер	Все контексты активны, переключения нет

CMP+SMT


Ядро А
Ядро Б
ALU
ALU
FPU
MEM
ALU
ALU
FPU
MEM
CMP
Такты


Ядро А
Ядро Б
ALU
ALU
FPU
MEM
ALU
ALU
FPU
MEM
CMP + SMT
Такты

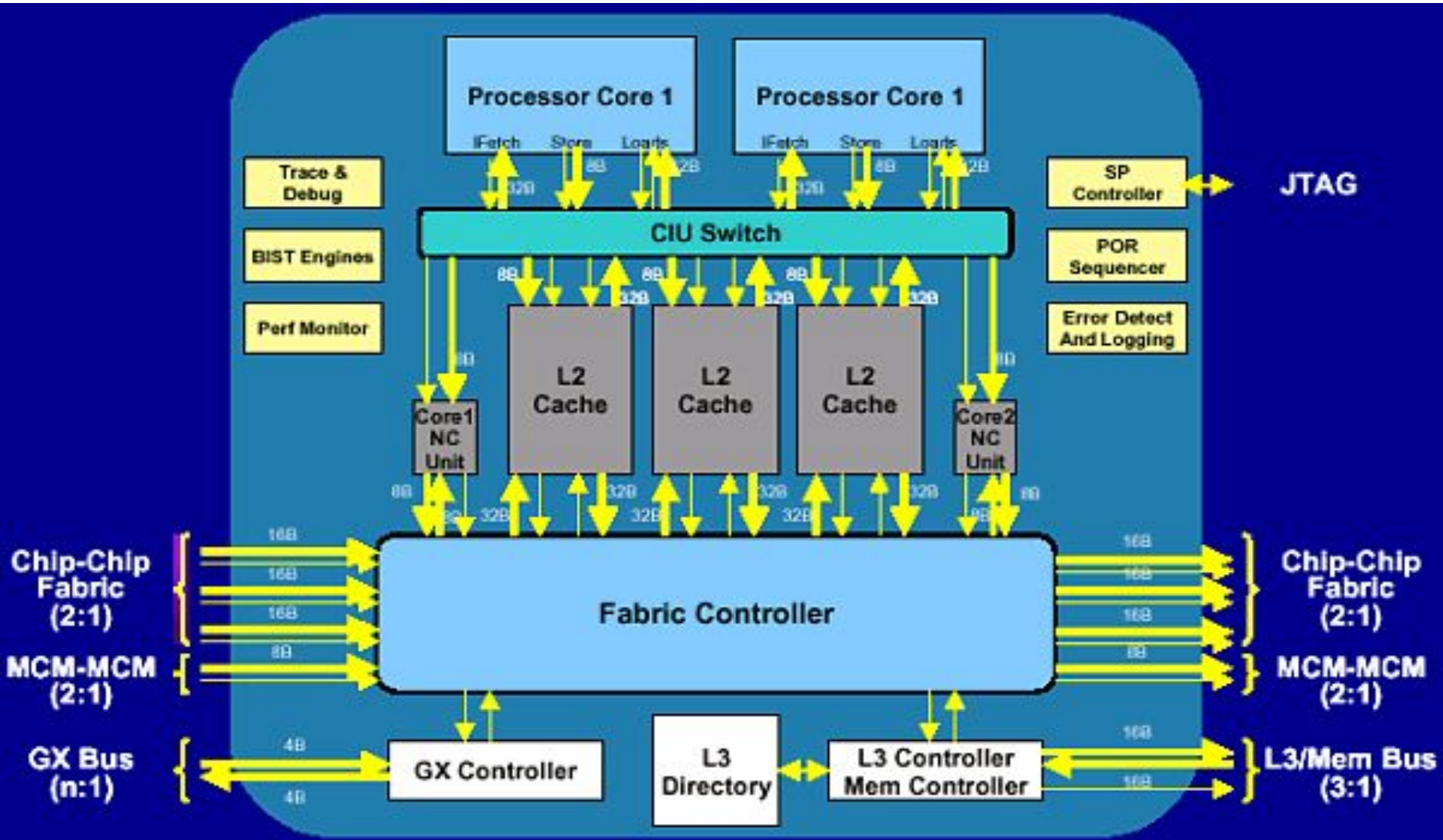
Многопоточные архитектуры



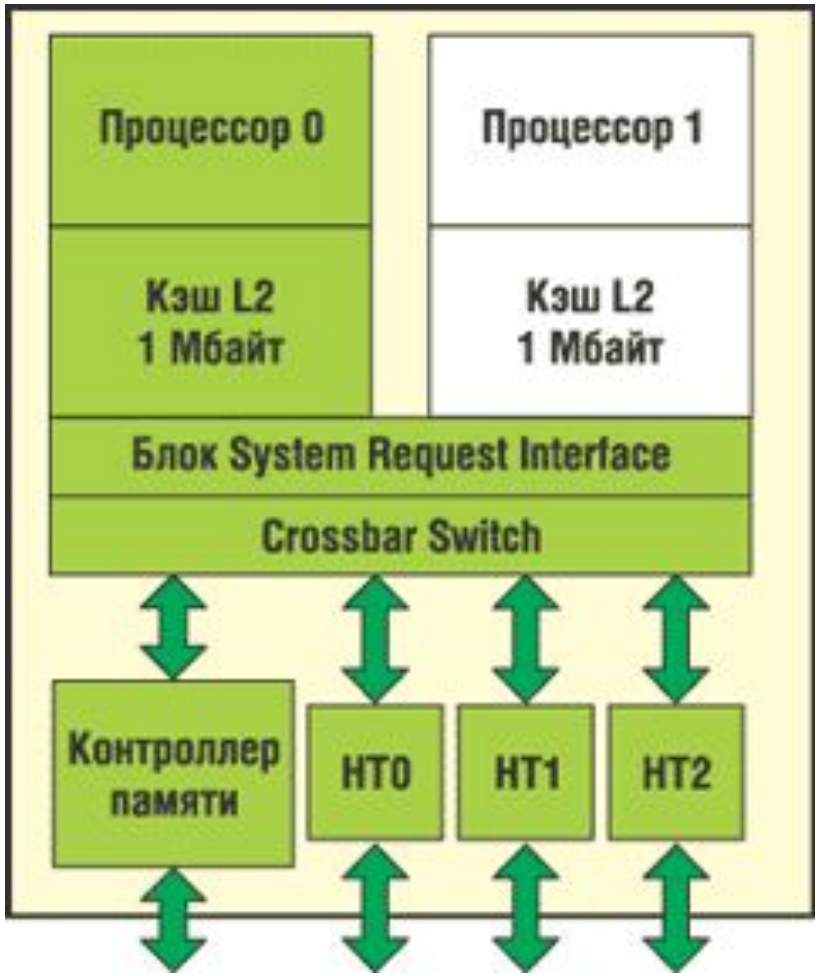
Примеры многопоточных архитектур

- **CMP**
  - POWER 4
  - Barcelona
  - Clovertown
  - Opteron
- **SMT**
  - Alpha 21464
  - Pentium 4
  - Itanium 2
- **CMP + SMT**
  - POWER 5
  - UltraSPARC T2
  - Keifer

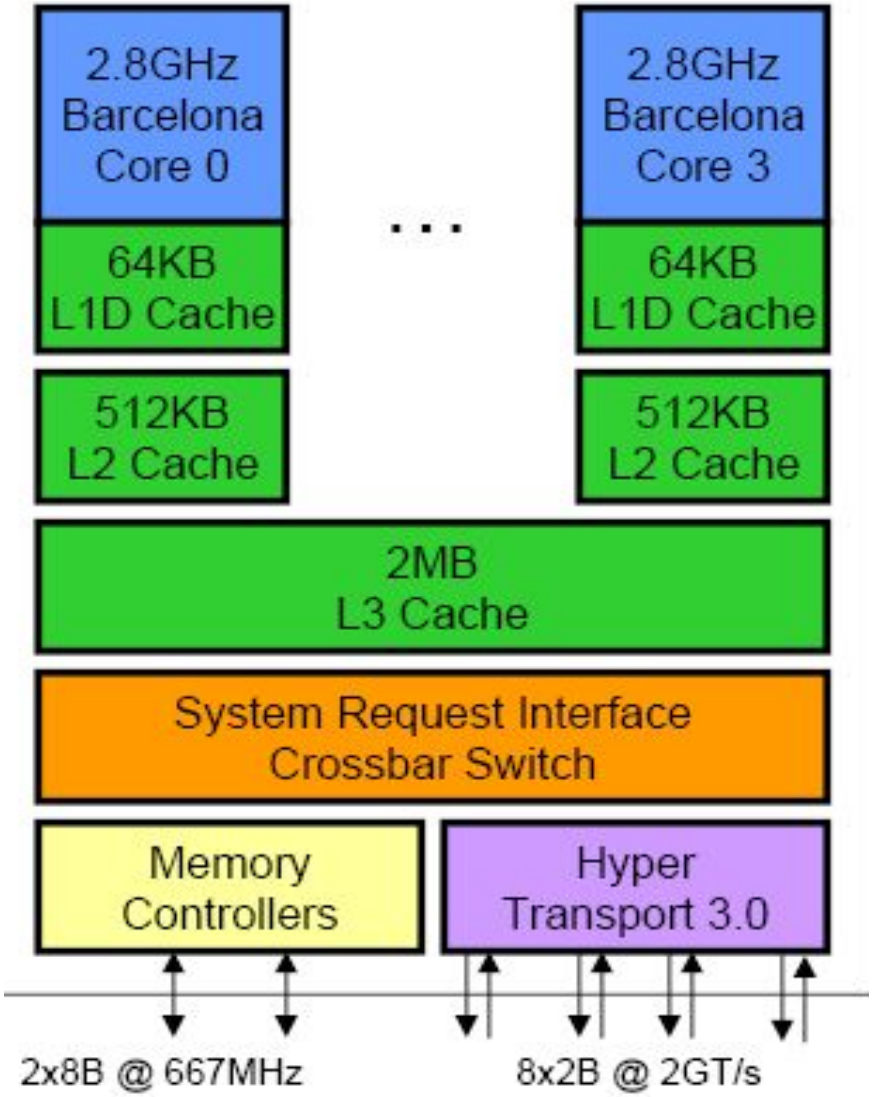
POWER 4
(2 ядра)



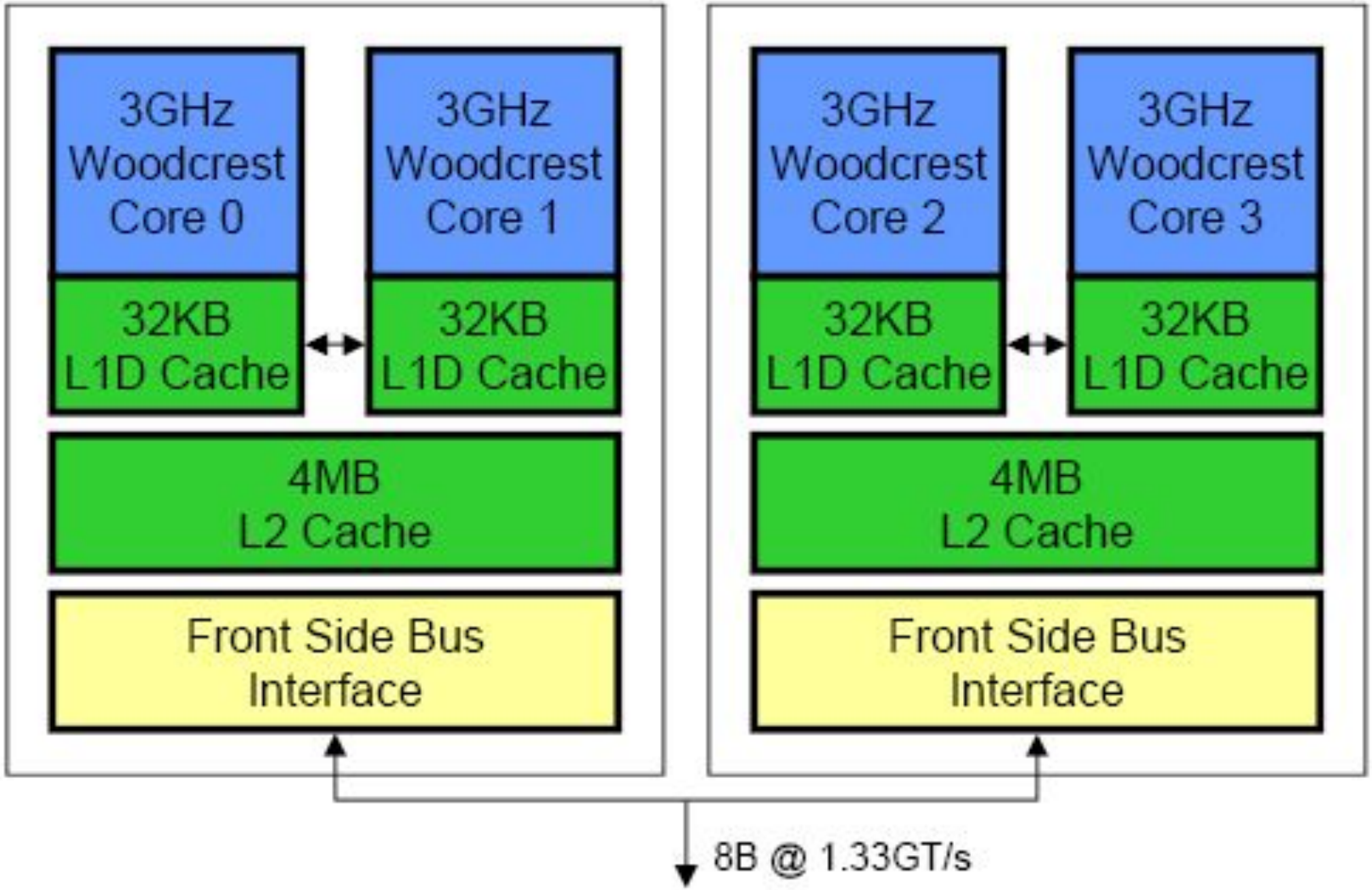
Dual core Opteron (2 ядра)



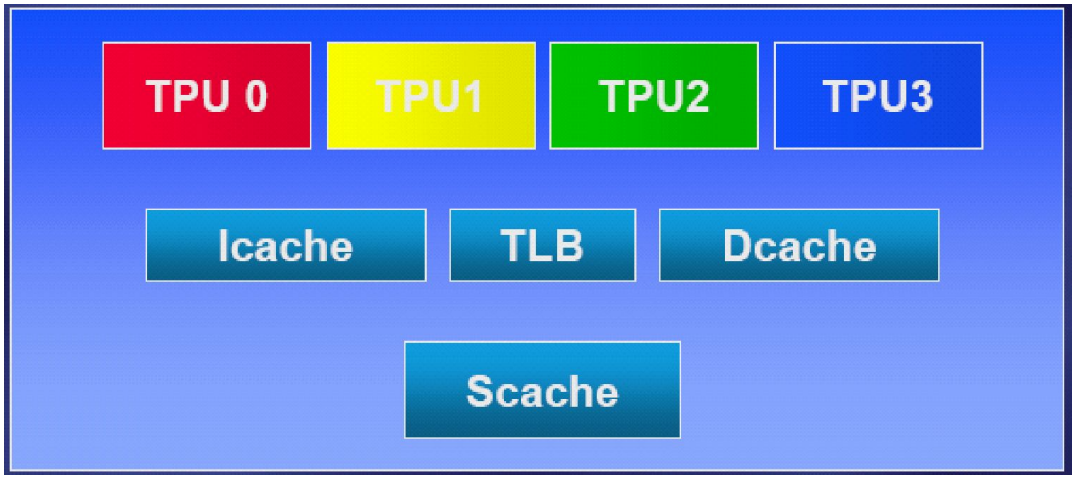
Barcelona
(4 ядра)

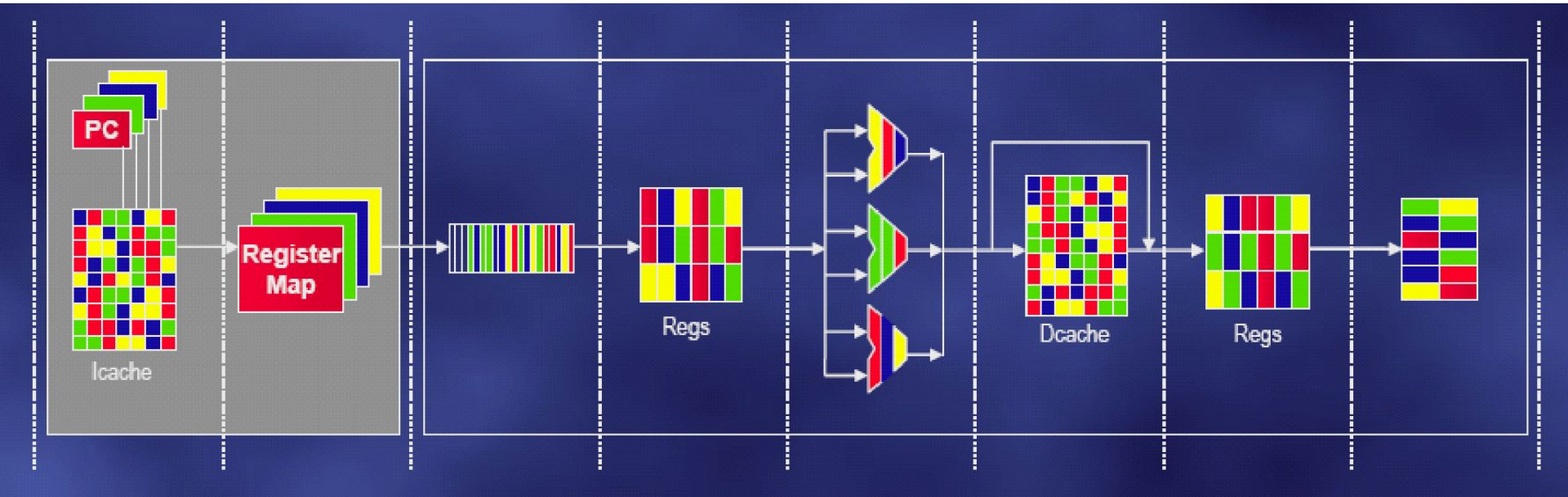


Clovertown
(4 ядра)

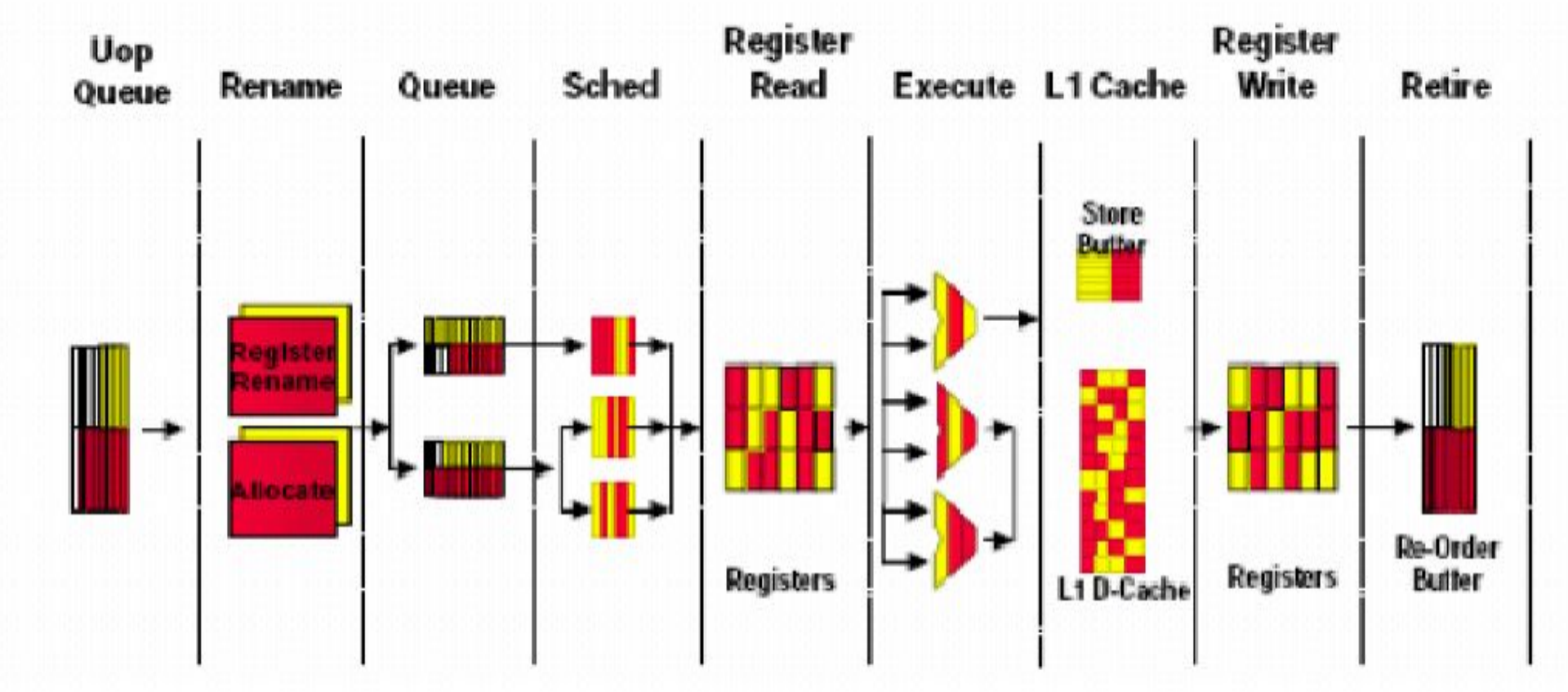


Alpha 21464
(4 потока)



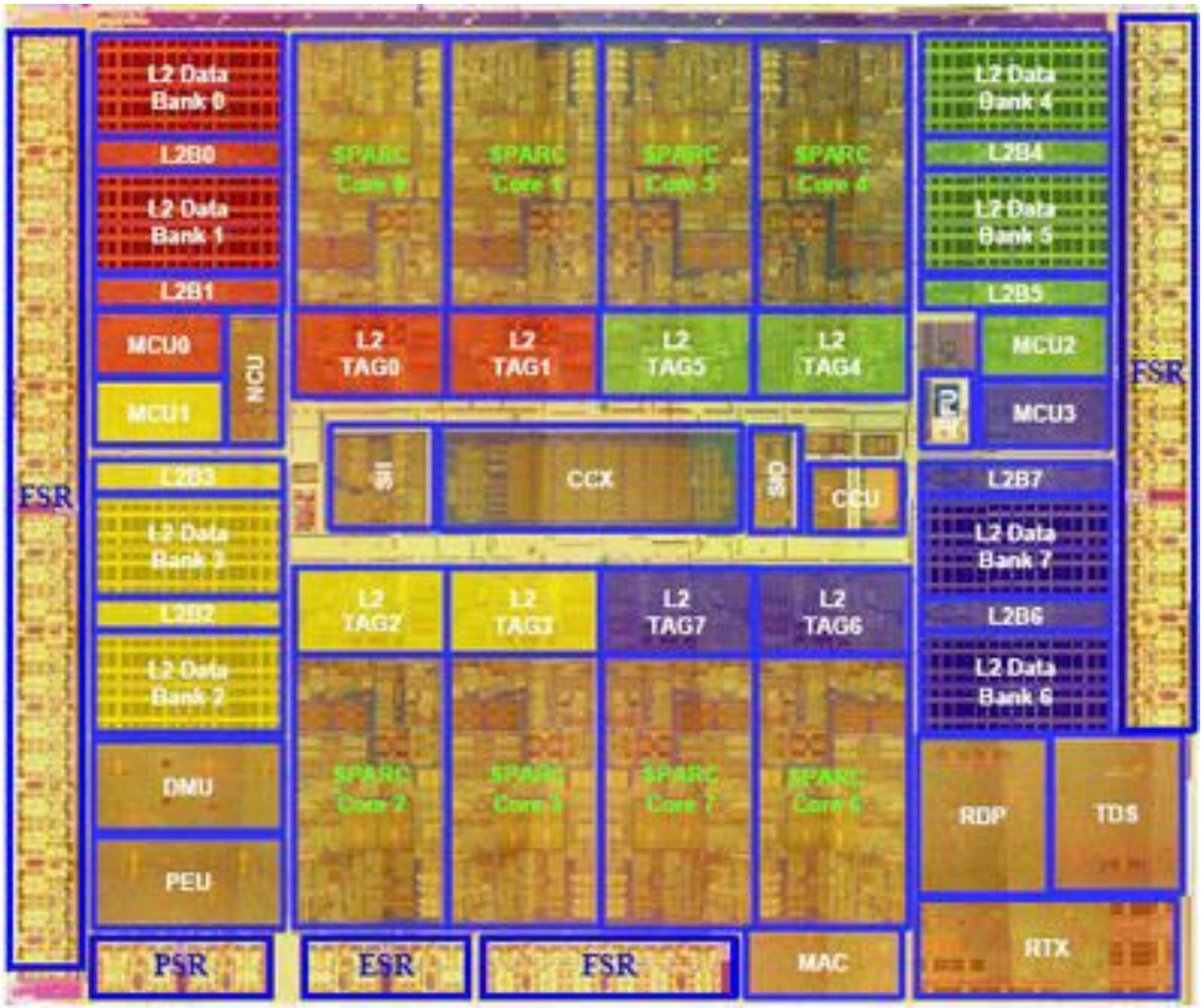
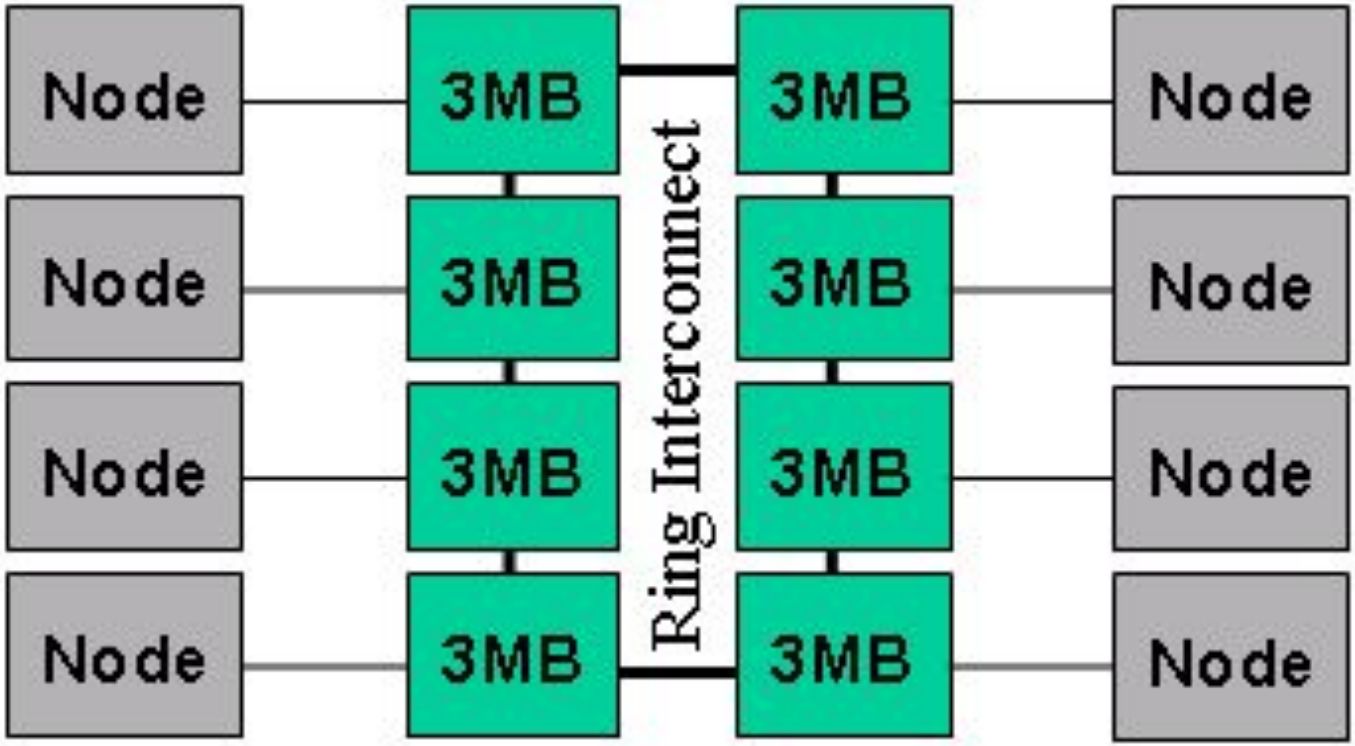


Pentium 4
(2 потока)



POWER 5
2 ядра 2 нити

UltraSPARC T2
(8 ядер x 8 потоков)



Keifer
(8 узлов x 4 ядер x 4 потока)