Нижегородский государственный университет
им. Н.И.Лобачевского

Факультет Вычислительной математики и кибернетики

Параллелизм как основа архитектуры ВС

Раздел 13

Модели многопоточных процессоров

Кудин А.В., к.т.н.

Содержание

- Ограничения однопоточных суперскалярных архитектур
- Причины длительных простоев конвейера
- Классификация потерь
- Использование параллелизма уровня потоков (TLP)
- Модификация суперскалярного CPU для поддержки SMT
- Сравнение суперскалярного и SMT конвейеров
- Модели многопоточного CPU
- Анализ производительности
- Влияние увеличения количества потоков
- Возможные стратегии планирования выборки инструкций

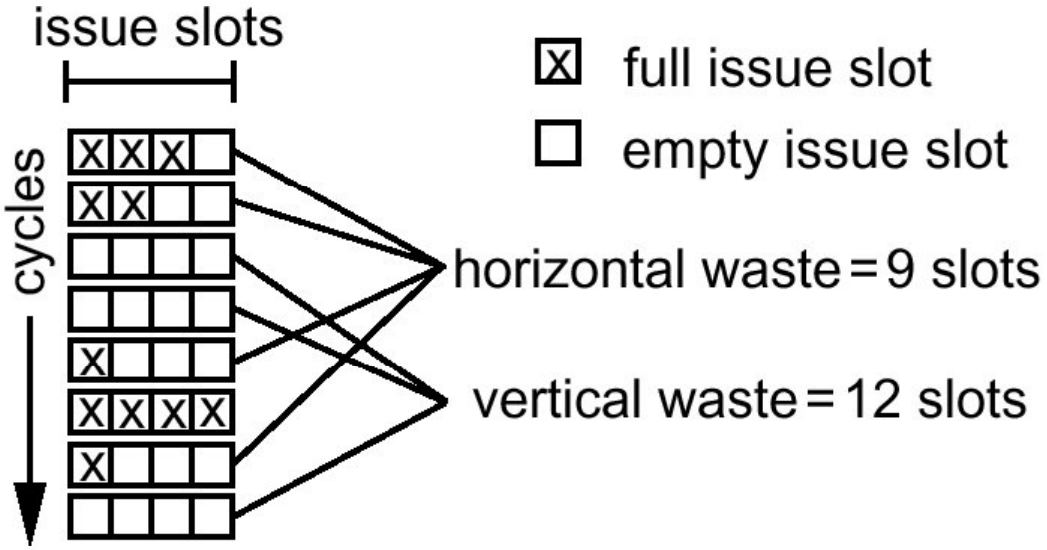
Ограничения суперскалярных архитектур

Классификация потерянных слотов выдачи инструкций:

- **Вертикальные потери**
- **Горизонтальные потери**

Причины возникновения пустых слотов – структурные конфликты, конфликты данных и управления.

Пример:
четырёхконвейерный CPU
идеальное IPC = 4
реальное IPC << идеальное IPC



Классические многопоточные процессоры

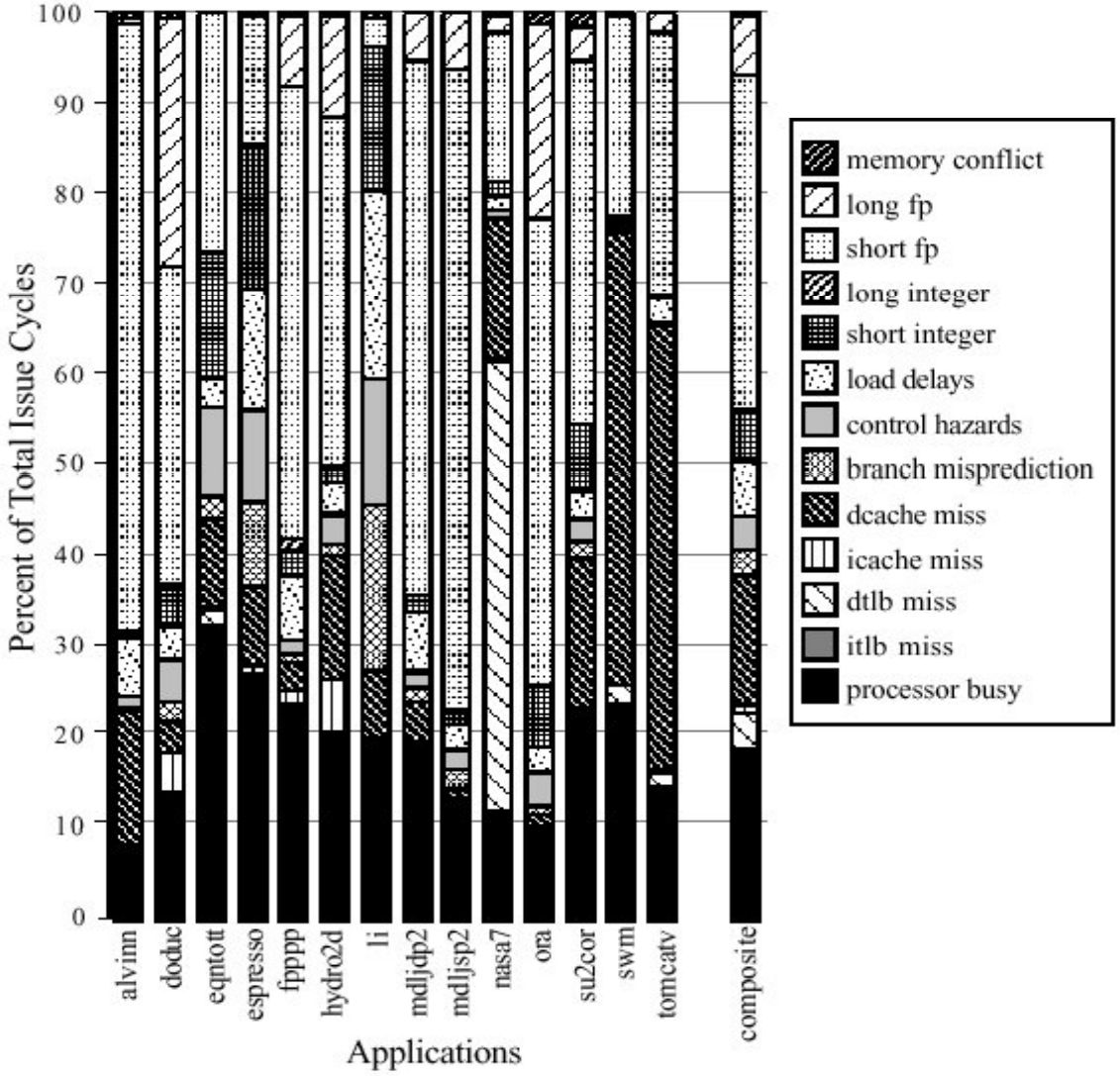
- Множество аппаратных контекстов, с точки зрения ОС это несколько логических процессоров
- Только один поток (один контекст) выдаёт инструкции каждый такт
- Не снижает горизонтальные потери
- Снижает вертикальные потери слотов выдачи
- Производительность ограничена параллелизмом уровня инструкций (ILP) в каждом отдельном потоке

Причины простоев конвейера

Пример:
восьмипоточный CPU
идеальное IPC = 8
реальное IPC ≈ 1.5

потеряно 81% слотов выдачи,
из них вертикальных 61%

нагрузка – SPEC92



Источник: TULLSEN, D.M., EGGERS, S.J., H.M. LEVY [1995], “Simultaneous multithreading: Maximizing on-chip parallelism”, *Proc. 22nd International Symposium on Computer Architecture* (June), pp.392-403

Конвейерная многопоточность

Simultaneous Multithreading (SMT) – эволюционная микропроцессорная архитектура, впервые представленная в 1995 году в университете Вашингтона Дином Тулсеном (Dean Tullsen)

- Предназначена для повышения эффективности использования аппаратных ресурсов в многопоточных суперскалярных микропроцессорах
- Использует параллелизм уровня потоков (Thread-level parallelism, TLP) на одном вычислительном ядре, позволяющий производить одновременную выдачу, исполнение и завершение инструкций из различных потоков в течении одного и того же такта (один физический SMT процессор действует как несколько логических процессоров, каждый из которых исполняет отдельный поток инструкций)
- Предоставляет эффективный механизм скрытия длительных простоев конвейера, снижая как вертикальные, так и горизонтальные потери

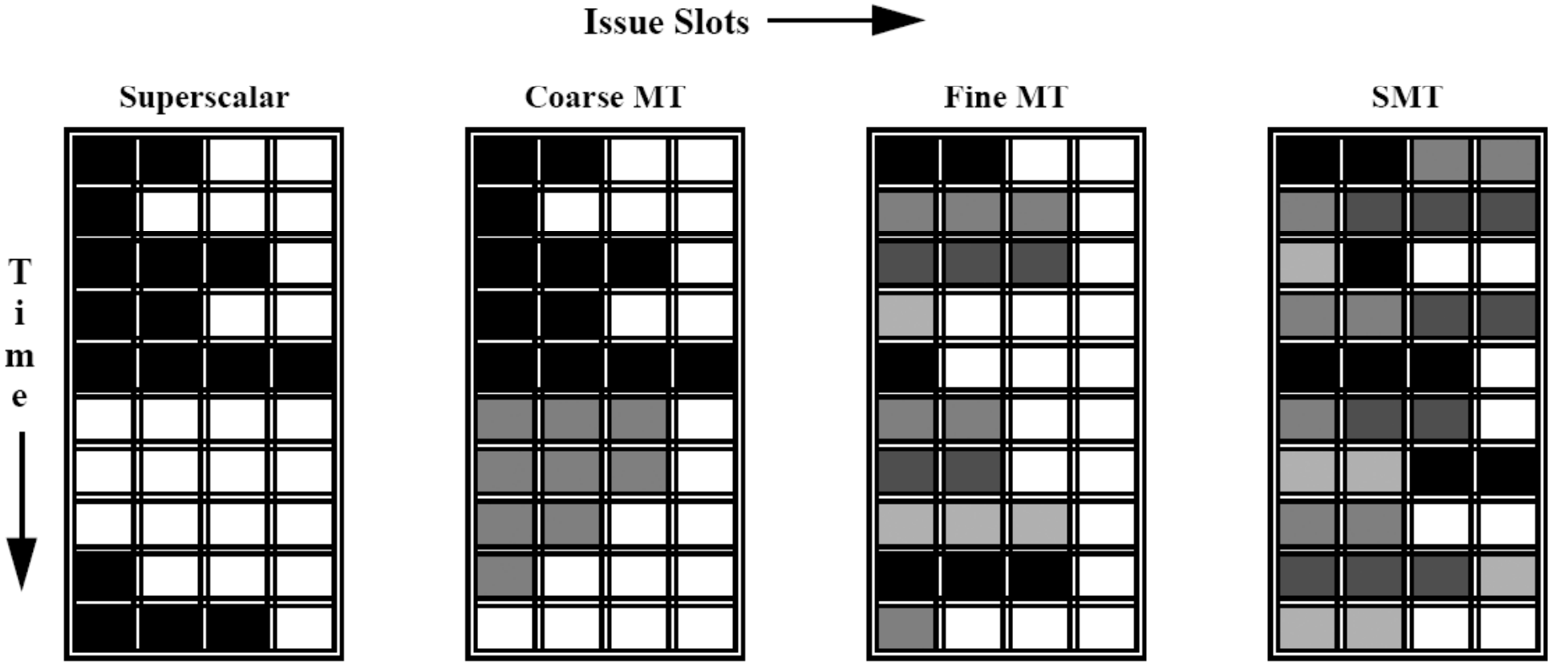
Экономическая целесообразность

Экономическая целесообразность технологии SMT:
рост производительности значительно больше, чем рост площади чипа и энергопотребления

Конвейерная многопоточность

- При множестве исполняющихся потоков в случаях промахов кеша, неверно предсказанных переходов и т.п. скрываются даже длительные штрафы (с большими задержками).
- Снижение и горизонтальных, и вертикальных потерь слотов выдачи ведёт к увеличению скорости выдачи инструкций (IPC).
- Функциональные устройства совместно используются всеми контекстами в каждом такте. Выдача инструкций для функциональных устройств множеством потоков повышает использование ресурсов процессора.
- Однако с целью поддержки множества исполняющихся потоков накладываются более жёсткие требования к размерам ресурсов CPU (кешей, буфера BTB, буфера TLB, регистров переименования и т.д.).

Сравнение эффективности



однопоточный CPU	грубая многопоточность	тонкая многопоточность	конвейерная многопоточность

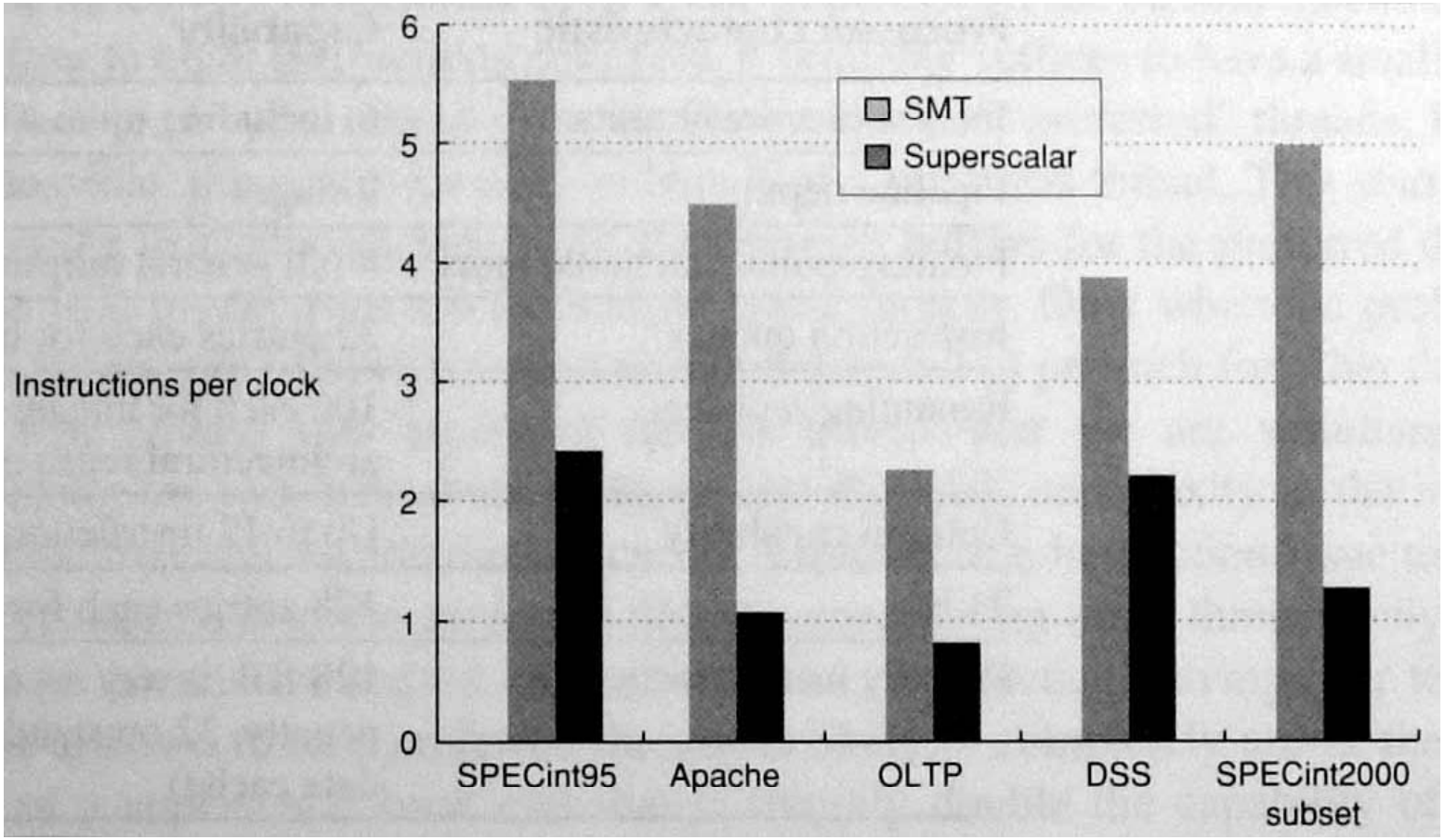
Необходимые изменения для поддержки SMT

- Множество контекстов исполнения и привязка каждой инструкции к своему контексту при прохождении всего конвейера
- Механизм выборки инструкций из множества потоков
- Отдельные для каждого потока механизмы завершения инструкций, менеджмента очереди инструкций и обработки исключений

Изменения для повышения производительности SMT

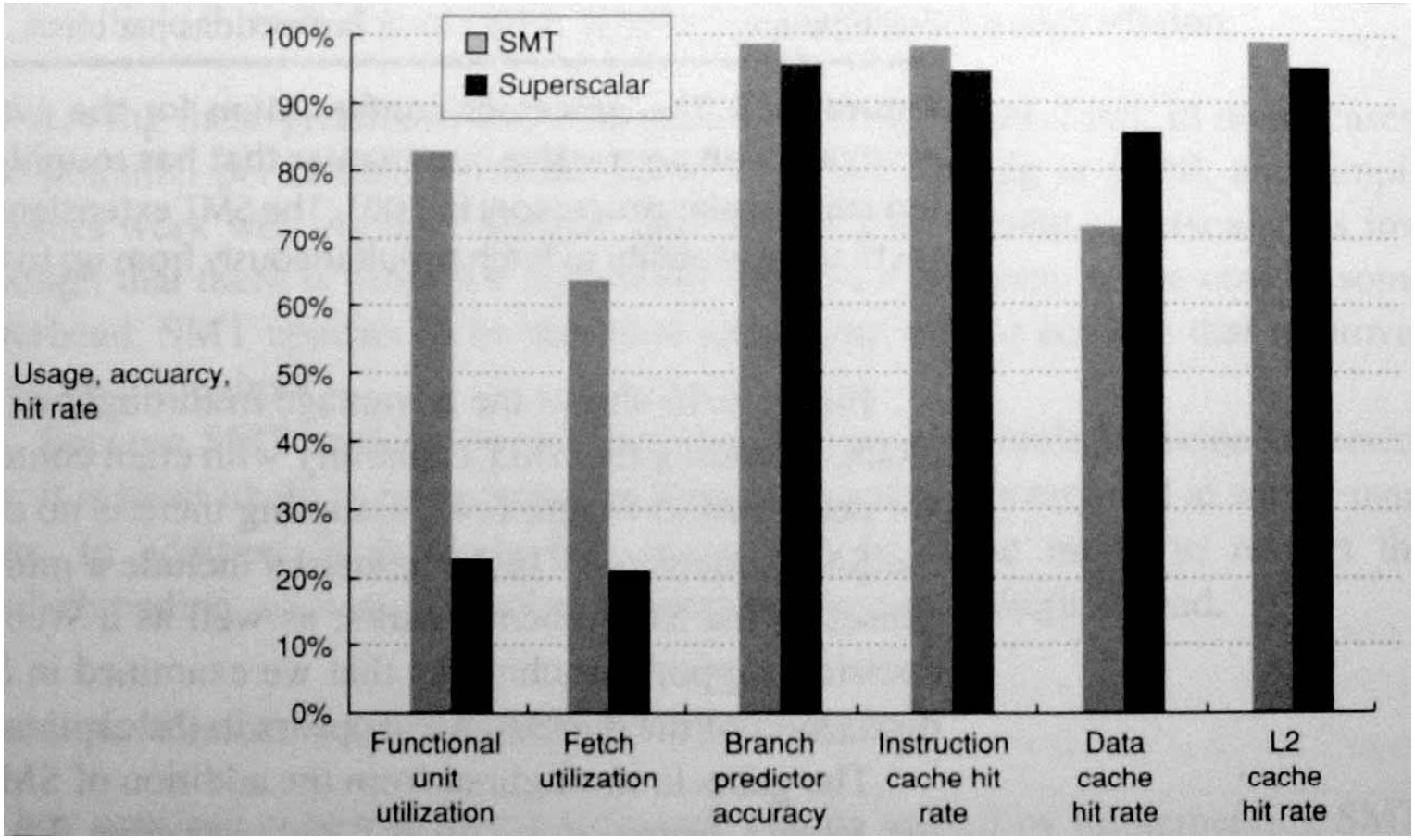
- Больший регистровый файл для поддержки переименования регистров
- Большая пропускная способность доступа к памяти
- Большие кеши для компенсации снижения производительности из-за совместного использования несколькими потоками (из-за снижения локальности)
- Больший BTB
- Больший TLB

Сравнение производительности



Источник: John L. Hennessy and David A. Patterson, Computer Architecture: A Quantitative Approach, 3rd Ed., Morgan Kaufmann, 2003

Сравнение эффективности ресурсов



Источник: John L. Hennessy and David A. Patterson, Computer Architecture: A Quantitative Approach, 3rd Ed., Morgan Kaufmann, 2003

Модели восьмипоточного процессора

Модели многопоточного CPU, который может выдавать до восьми инструкций за такт, отличаются способами использования слотов выдачи и функциональных устройств

- **Fine-Grain Multithreading**
- **SM: Limited Connection**
- **SM: Single Issue**
- **SM: Dual Issue**
- **SM: Four Issue**
- **SM: Full Simultaneous Issue**

Модели восьмипоточного процессора

Fine-Grain Multithreading

Классическая тонкая многопоточность. Только один поток выдаёт инструкции каждый такт, зато может использовать всю ширину выдачи процессора. Скрывает все источники вертикальных потерь, но не горизонтальных.

SM: Limited Connection

Ограниченные связи в конвейере. Каждый аппаратный контекст напрямую соединен только с некоторыми функциональными устройствами. Например, если аппаратура поддерживает восемь потоков и имеет четыре целочисленных устройства, каждое целочисленное устройство может получать инструкции в точности от двух потоков. Разбиение функциональных устройств по потокам в результате менее динамично, чем в других SMT моделях, но каждое функциональное устройство разделяемо (критический фактор при достижении высокой утилизации ресурсов).

Модели восьмипоточного процессора

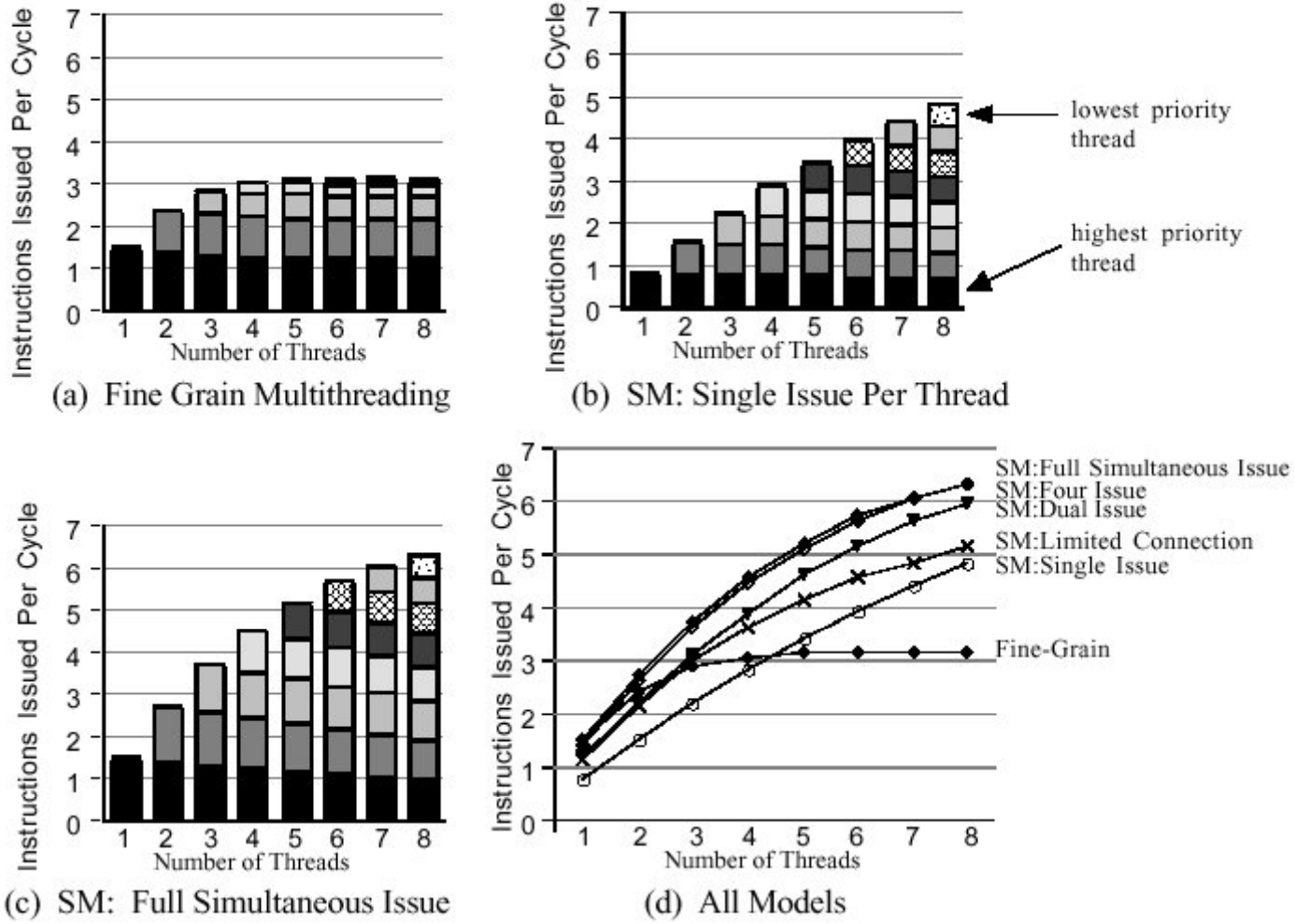
SM: Single / Dual / Four Issue

В этих трёх моделях ограничивается количество инструкций, которое каждый поток может выдать или иметь активными в окне планирования на каждом такте. Например, в SM: Dual Issue каждый поток может выдать максимум две инструкции за такт, и потребуется минимум четыре потока для заполнения восьми слотов выдачи в одном такте.

SM: Full Simultaneous Issue

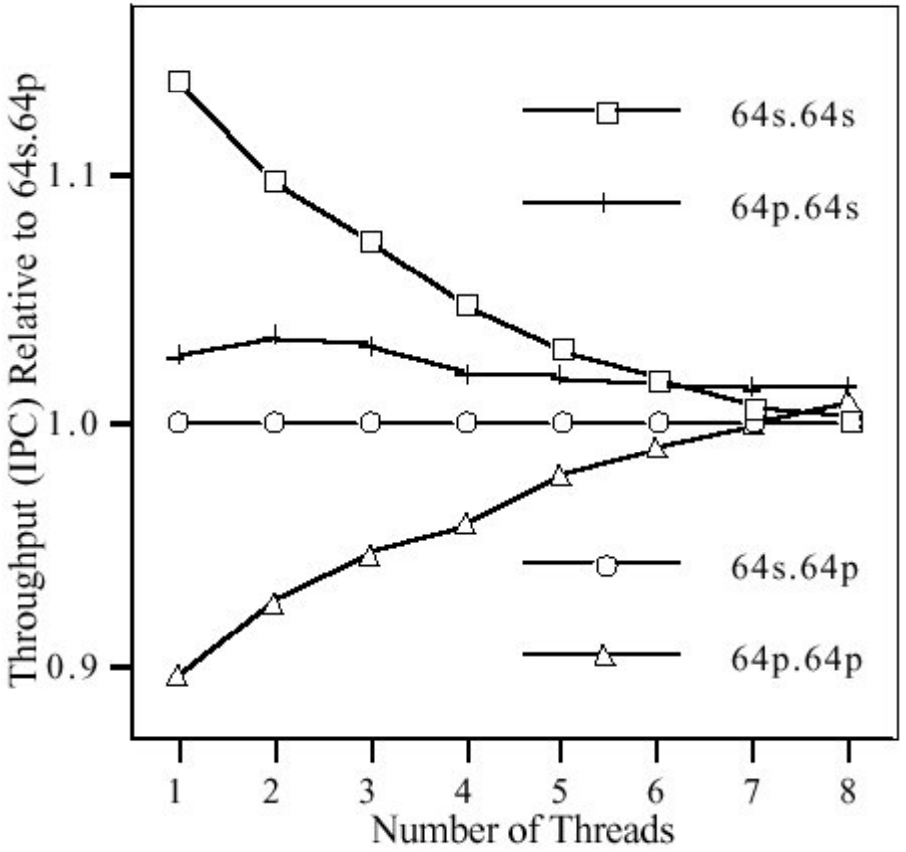
Полностью одновременная выдача – самая гибкая модель суперскалярного процессора с конвейерной многопоточностью: все восемь потоков конкурируют за каждый из восьми слотов выдачи каждый такт. Наивысшая сложность аппаратной реализации.

Сравнение производительности моделей



Источник: TULLSEN, D.M., EGGERS, S.J., H.M. LEVY [1995], “Simultaneous multithreading: Maximizing on-chip parallelism”, *Proc. 22nd International Symposium on Computer Architecture* (June), pp.392-403

Влияние кеша I уровня на производительность



Сравнение пропускной способности процессора (IPC) в зависимости от конфигурации кеша I уровня относительно конфигурации 64s.64p

объём кеша кода в Kb

совместный (shared)

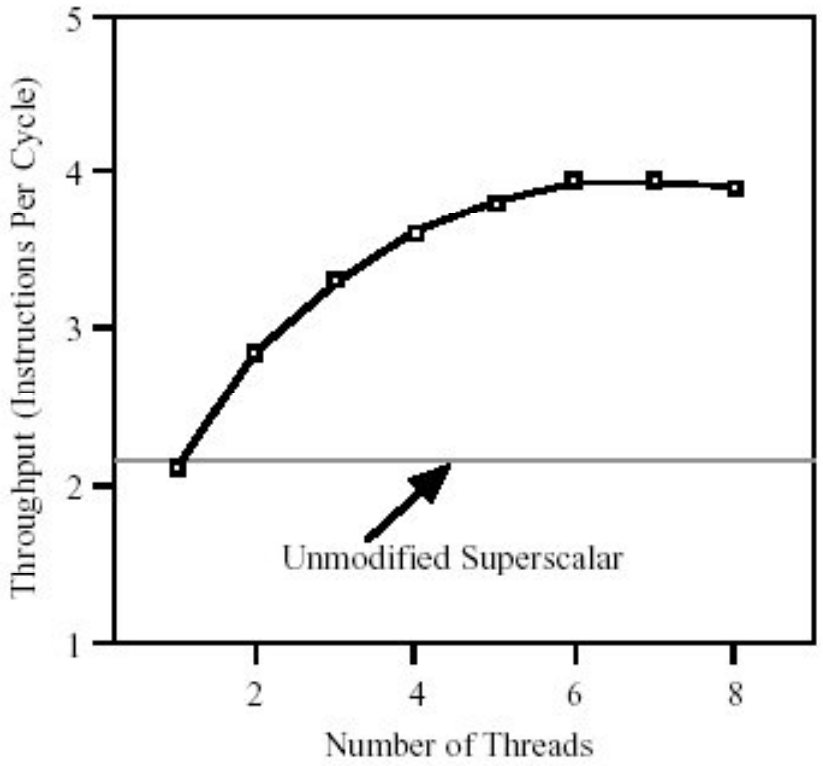
объём кеша данных в Kb

разделённый (private)

Источник: TULLSEN, D.M., EGGERS, S.J., H.M. LEVY [1995], “Simultaneous multithreading: Maximizing on-chip parallelism”, *Proc. 22nd International Symposium on Computer Architecture* (June), pp.392-403

Влияние количества потоков на ресурсы

Metric	Number of Threads		
	1	4	8
out-of-registers (% of cycles)	3%	7%	3%
I cache miss rate	2.5%	7.8%	14.1%
-misses per thousand instructions	6	17	29
D cache miss rate	3.1%	6.5%	11.3%
-misses per thousand instructions	12	25	43
L2 cache miss rate	17.6%	15.0%	12.5%
-misses per thousand instructions	3	5	9
L3 cache miss rate	55.1%	33.6%	45.4%
-misses per thousand instructions	1	3	4
branch misprediction rate	5.0%	7.4%	9.1%
jump misprediction rate	2.2%	6.4%	12.9%
integer IQ-full (% of cycles)	7%	10%	9%
fp IQ-full (% of cycles)	14%	9%	3%
avg (combined) queue population	25	25	27
wrong-path instructions fetched	24%	7%	7%
wrong-path instructions issued	9%	4%	3%



Источник: TULLSEN, D.M. et al. [1996]. “Exploiting choice: Instruction fetch and issue on an implementable simultaneous multithreading processor”, *Proceedings of the 23rd Annual International Symposium on Computer Architecture* (May), pages 191-202

Стратегии планирования выборки инструкций

Round Robin

Выборка по регулярному графику. Например, в RR 1.8 каждый такт из одного потока выбирается до восьми инструкций, а в RR 2.4 каждый такт из двух потоков выбирается до четырёх инструкций.

BR-Count

Наивысший приоритет у потоков с наименьшей вероятностью ложной спекуляции исполнения (с наименьшим числом невычисленных переходов).

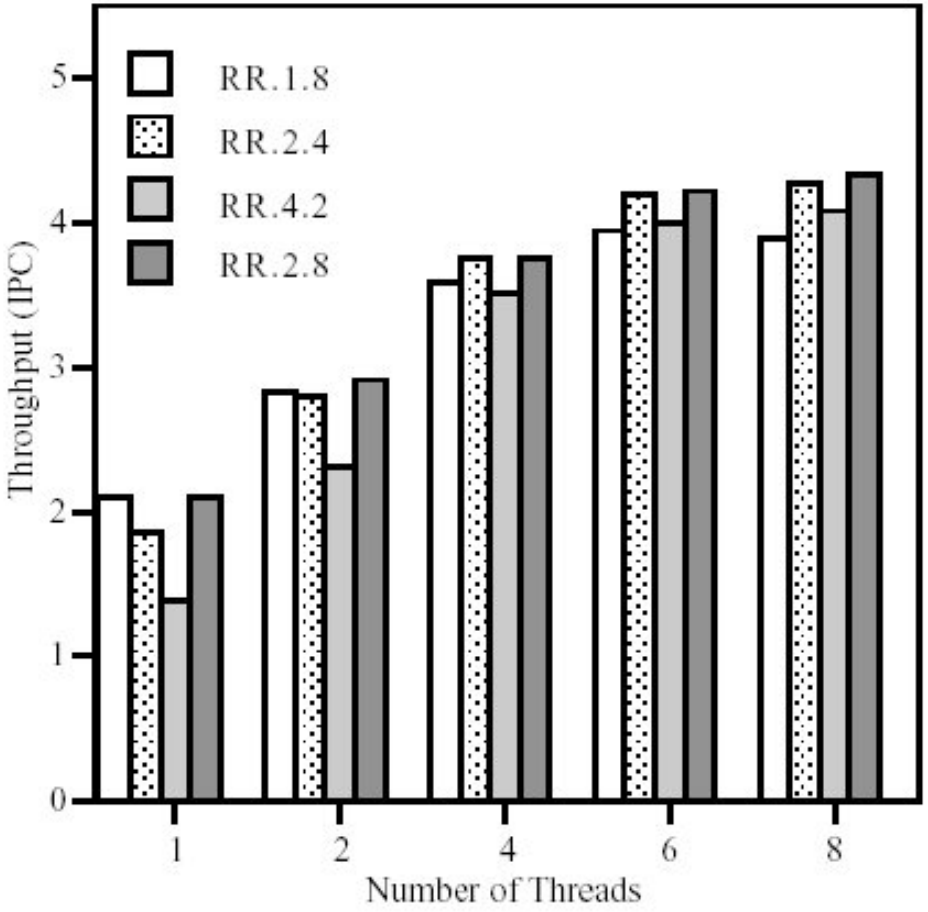
MISS-Count

Наивысший приоритет у потоков с наименьшей частотой промахов в кеше данных.

I-Count

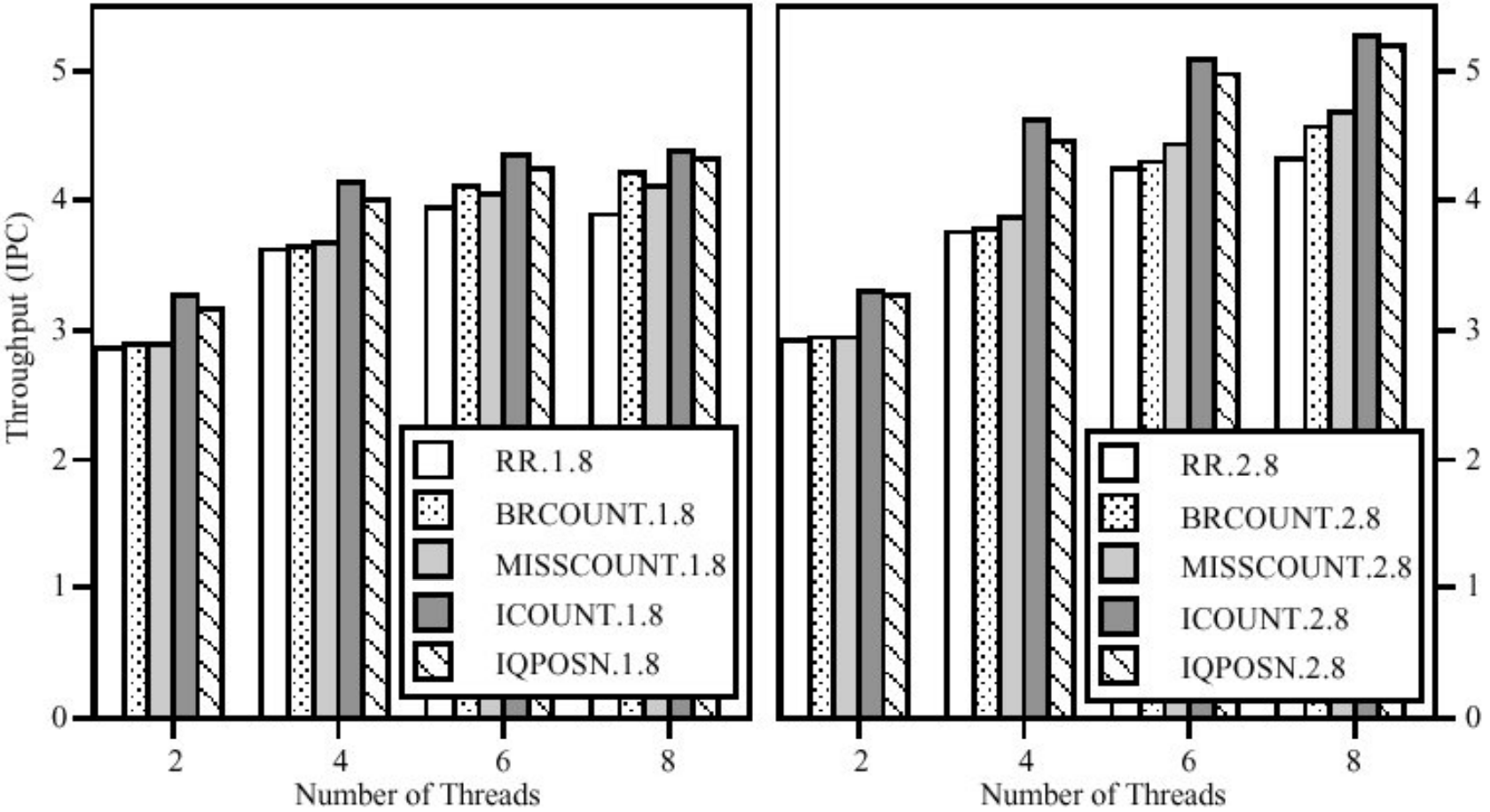
Наивысший приоритет у потоков с наименьшим числом инструкций в статической части конвейера (в очереди декодированных инструкций).

Производительность с Round Robin



Источник: TULLSEN, D.M. et al. [1996]. “Exploiting choice: Instruction fetch and issue on an implementable simultaneous multithreading processor”, *Proceedings of the 23rd Annual International Symposium on Computer Architecture* (May), pages 191-202

Сравнение эвристик выборки инструкций



Источник: TULLSEN, D.M. et al. [1996]. “Exploiting choice: Instruction fetch and issue on an implementable simultaneous multithreading processor”, *Proceedings of the 23rd Annual International Symposium on Computer Architecture* (May), pages 191-202

Сравнение эвристик выборки инструкций

Metric	1 Thread	8 Threads	
		RR	ICOUNT
integer IQ-full (% of cycles)	7%	18%	6%
fp IQ-full (% of cycles)	14%	8%	1%
avg queue population	25	38	30
out-of-registers (% of cycles)	3%	8%	5%

ICOUNT.2.8 увеличивает производительность по сравнению с RR.2.8 на 23% за счёт уменьшения помех в потоке инструкций, выбирая лучшую смесь инструкций

Заключение

- **Достоинства**
  - Преодоление ограничений производительности суперскалярной обработки, связанных с низким параллелизмом уровня инструкций (ILP), посредством применения параллелизма уровня потоков (TLP)
  - Повышение эффективности использования аппаратных ресурсов
  - Предоставление механизма скрытия длительных простоев конвейера
- **Недостатки**
  - Повышенные требования к производительности иерархии памяти
  - Повышенные требования к размерам ресурсов CPU

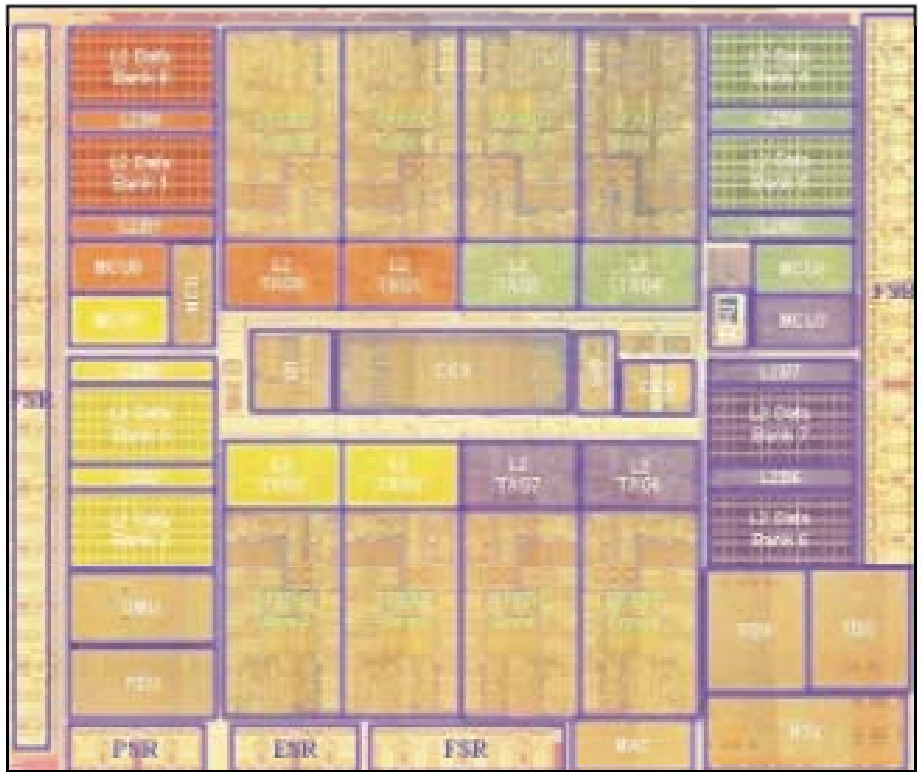
Пример: Niagara 2

8 ядер × 8 потоков

кеш I уровня кода 16K, данных 8K

кеш II уровня 4M для каждого ядра

Niagara 2



Пример: Power 6

32 ядра × 2 потока

кеш I уровня кода 64K, данных 64K
кеш II уровня 8M для каждого ядра
кеш III уровня 32M (16-входовый)

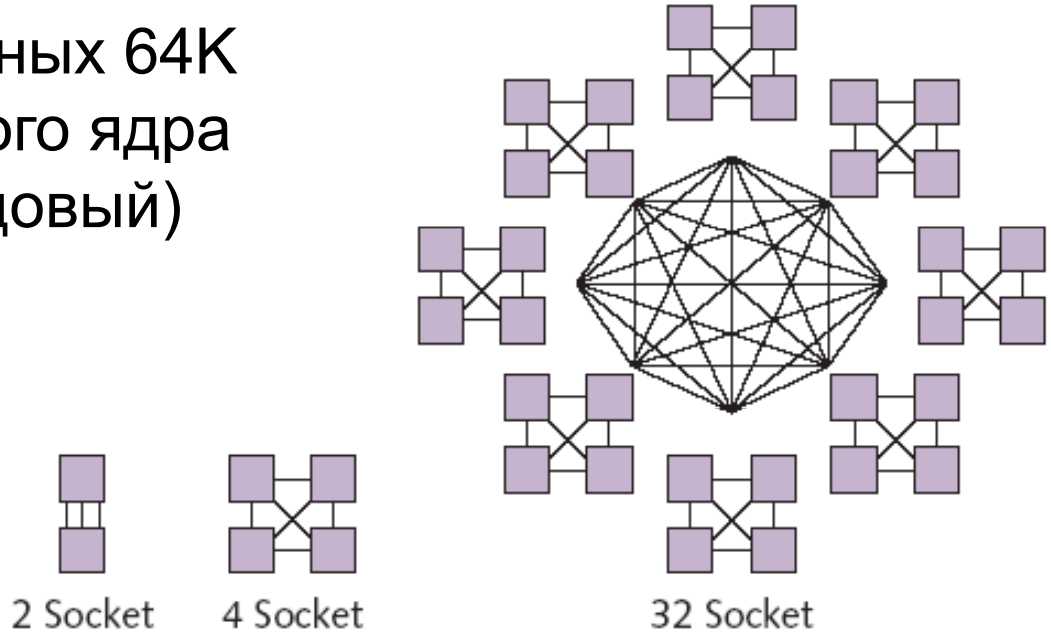


Figure 5. Basic SMP configurations for POWER6. POWER6 not only provides more cost/performance flexibility at the system-configuration level than previous generations, but the two-tier interconnection scheme used to link basic four-socket clusters is also new. IBM now appears to have adopted the more industry-standard practice of counting sockets (chips) rather than processor cores.

Вопросы для обсуждения

- На каких задачах эффективнее двухпоточный процессор, а на каких двухядерный?
- В каких случаях целесообразнее использовать однопоточный режим многопоточного суперскалярного процессора?
- Какие преимущества имеет классическая многопоточность над конвейерной многопоточностью?
- В чём отличие конвейерной многопоточности с политикой выборки Round Robin от классической тонкой многопоточности?
- Какие штрафы наиболее тяжело скрываются в конвейерной многопоточности?

Следующая тема

- **Симметричное мультипроцессирование**