



Гетерогенная многопроцессорная система на кристалле с производительностью 512 Gflops

Эйсымонт Алексей
ЗАО НТЦ «Модуль»

Алушта 2017



Введение (1)

Приложения для встроенных вычислителей:

- первичная обработка сигналов (преобразования Фурье, Адамара, фильтры...)
- многослойные нейронные сети

Нужна производительность десятки Tflops (FP32)

Обычный способ решения задачи: жестко специализированные процессоры под конкретную задачу или алгоритм (Domain Specific Accelerators)

- например: DSP, нейросетевые разных типов, криптопроцессоры, видео кодеры/декодеры...
- высокая эффективность меняется на узкую специализацию и слабую программируемость

Под разные задачи нужны разные DSA – это дорого и долго!

Цель: разработать проблемно-ориентированный процессор, который:

1. имеет высокую производительность (~500 Gflops FP32)
2. может применяться для широкого класса прикладных задач и алгоритмов (универсален)
3. энергоэффективен (10-20 Gflops/Вт)

Введение (2)

Мировой опыт построения проблемно-ориентированных процессоров:

- необходимо учитывать организацию вычислений в выбранной предметной области
- особое внимание - к проблеме высоких задержек выполнения обращений к памяти
- GPU+CPU на одной СБИС позволяют получить адаптируемость к разным типам вычислений - **гетерогенность архитектуры**
 - **GPU** – для задач с высоким параллелизмом по данным, требующих очень высокой производительности
 - **CPU** – для управления и для задач с низким параллелизмом по данным, но с хорошей пространственно-временной локализацией обращений к памяти (в таких задачах хорошо работает кэш-память)
- энергоэффективность

Пример: «суперкомпьютерная СнК»
NVIDIA Xavier (образцы Q4 2017)

- для беспилотных автомобилей
- 7 млрд транзисторов 16нм FinFet
- 8 ядер ARMv8 64p
- 512 CUDA-ядер GPU Volta
- 30 TOPS
- **20 Вт**



Выбор архитектуры NMC4 и способа работы с памятью в СнК NM6408

Приложение	Локализация обращений к памяти	
	пространственная	временная
первичная обработка сигналов	плохая	хорошая
нейросети	хорошая	плохая

Вывод: применение в архитектуре кэш-памятей данных не оправдано для выбранной области задач

Альтернатива: высокий темп выполнения обращений к памяти, необходима толерантность архитектуры к задержкам выполнения обращений к медленной памяти

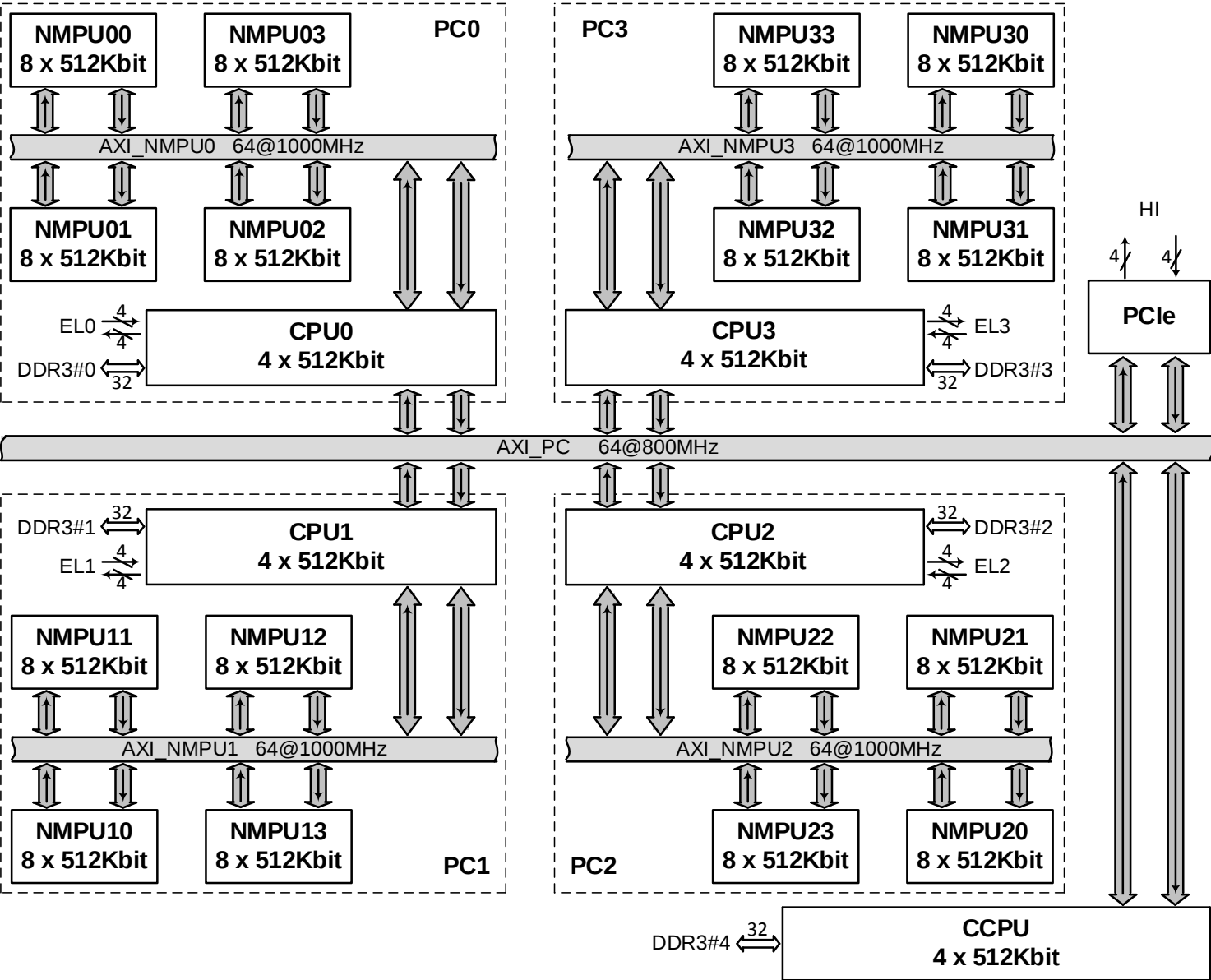
векторная или мультитредовая архитектура?

Итог:

- ▶ для высокопроизводительных ядер NMC4 выбрана **векторная архитектура**
- ▶ в качестве управляющих RISC-ядер используются **ARM Cortex A5**



Упрощенная структурная схема NM6408



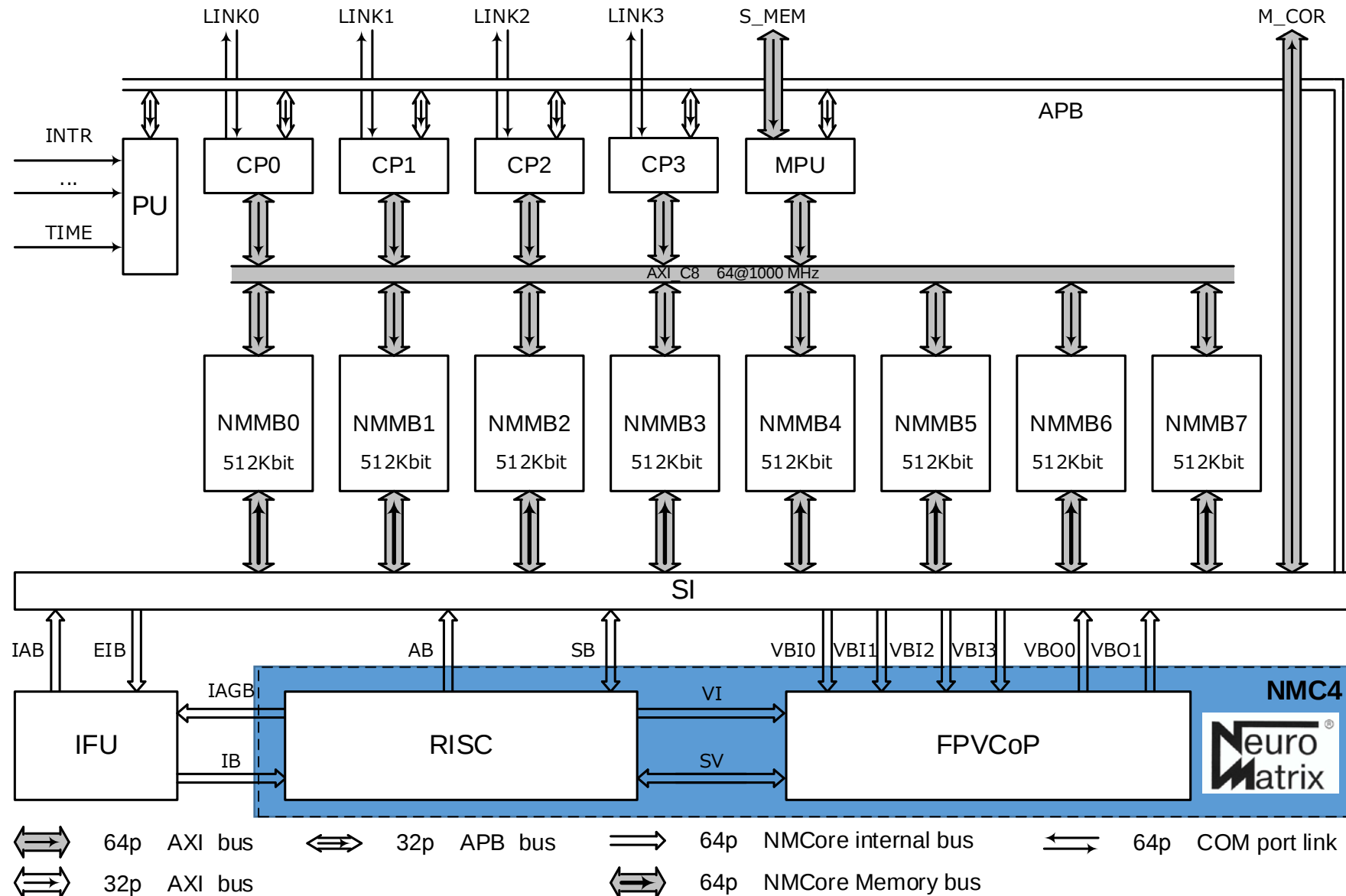
Процессорные узлы (ПУ) (тип/ядро/память/частота)			
NMPU _{ij}	NMC4	512КБ + L1I (8КБ)	1000
CPU _i	C-A5	256КБ + L1\$(2*32КБ)	800
CCPU	C-A5	256КБ + L1\$(2*32КБ) + L2(512КБ)	800

Процессный кластер PC _i		
4 * NMPU	@1ГГц	128Gflops
1 * CPU	@800МГц	
EL	4line@5ГГц	2ГБ/с дуплекс
DDR3-1600	32b@800МГц	6.4ГБ/с

Производительность 16 ПУ NMPU
512 Gflops

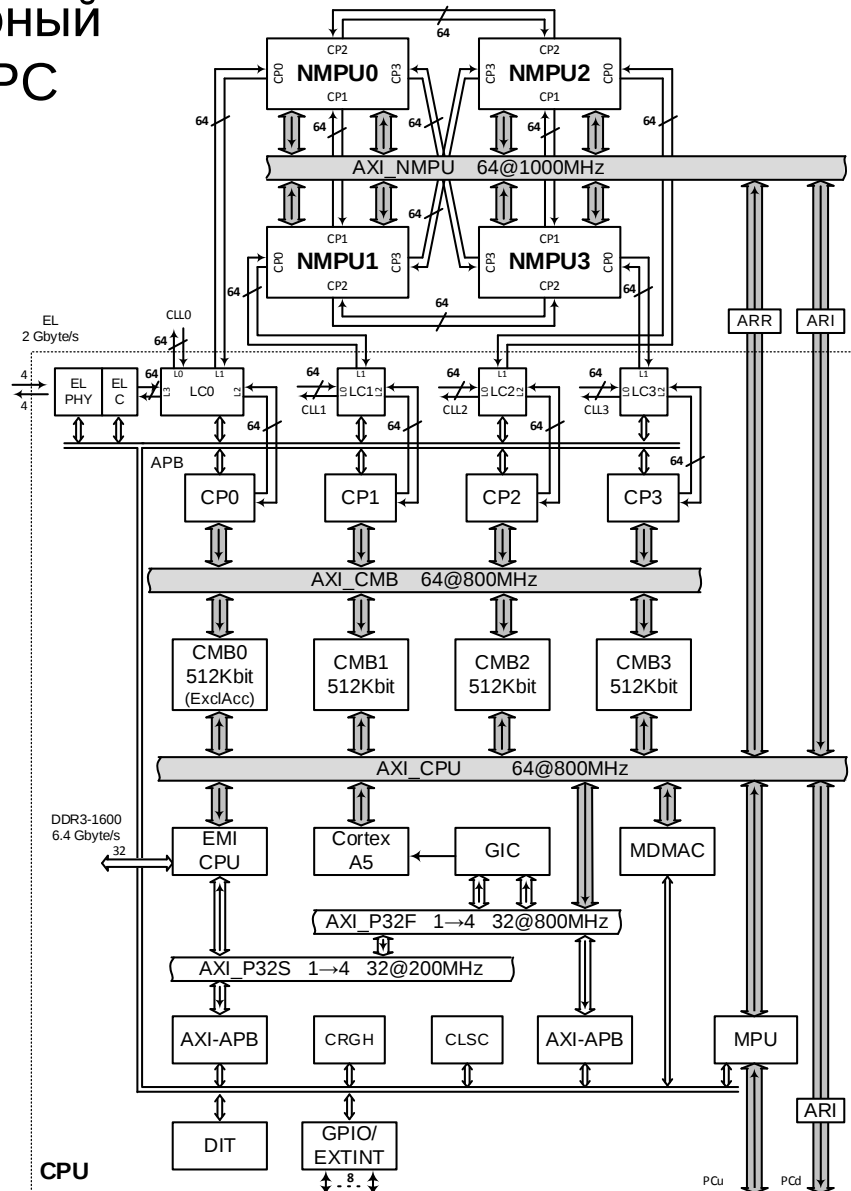


Структурная схема процессорного узла NMPU. Ядро NMC4.

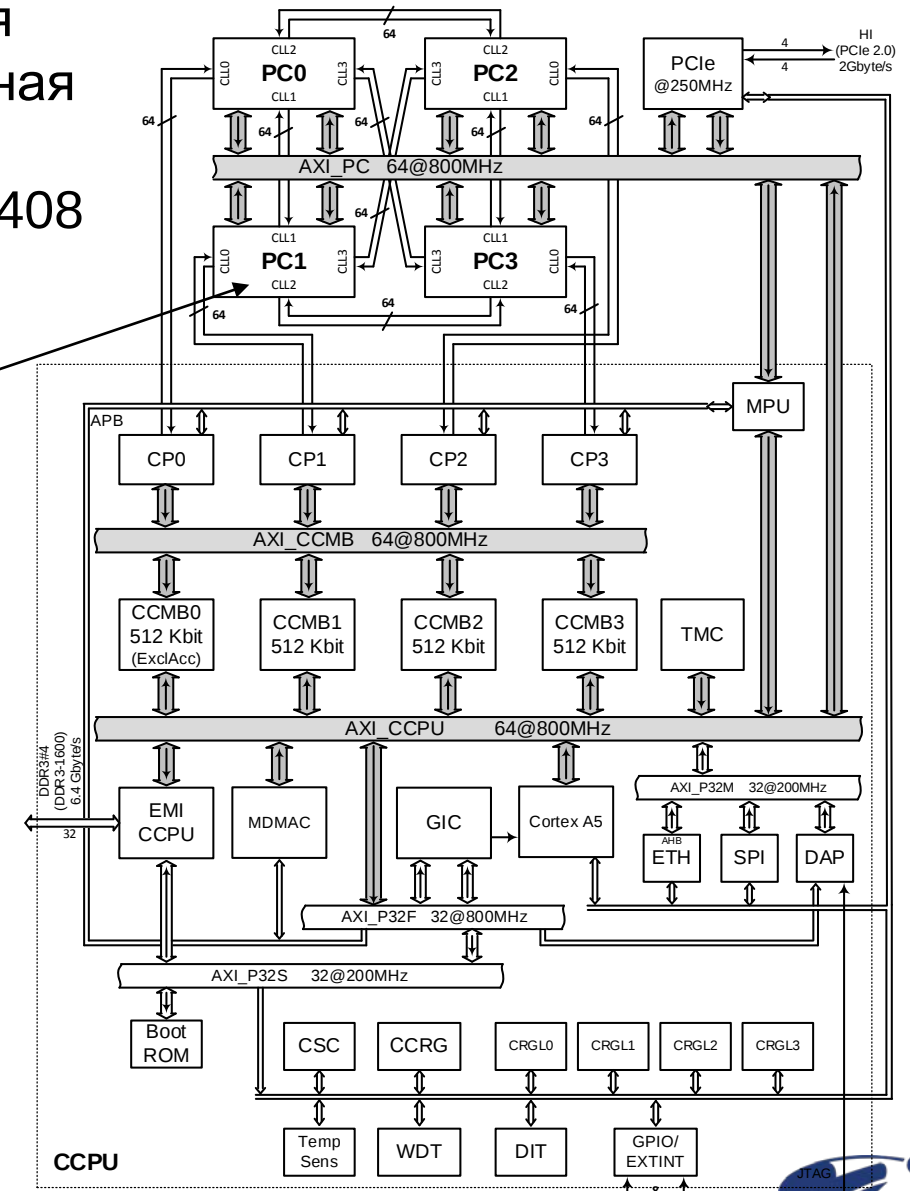


Структурная схема РС и общая структурная схема СБИС

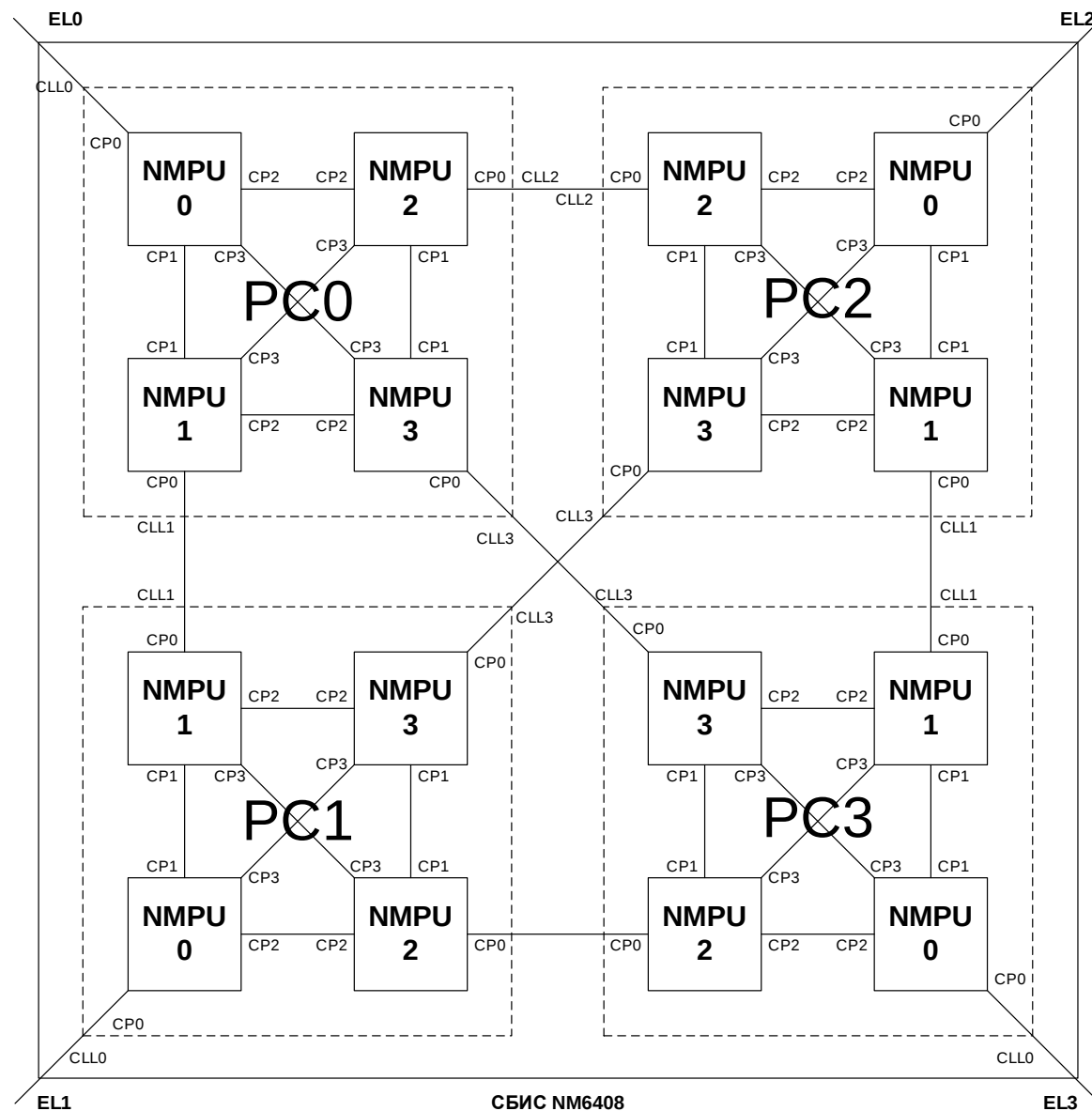
Процессорный кластер РС



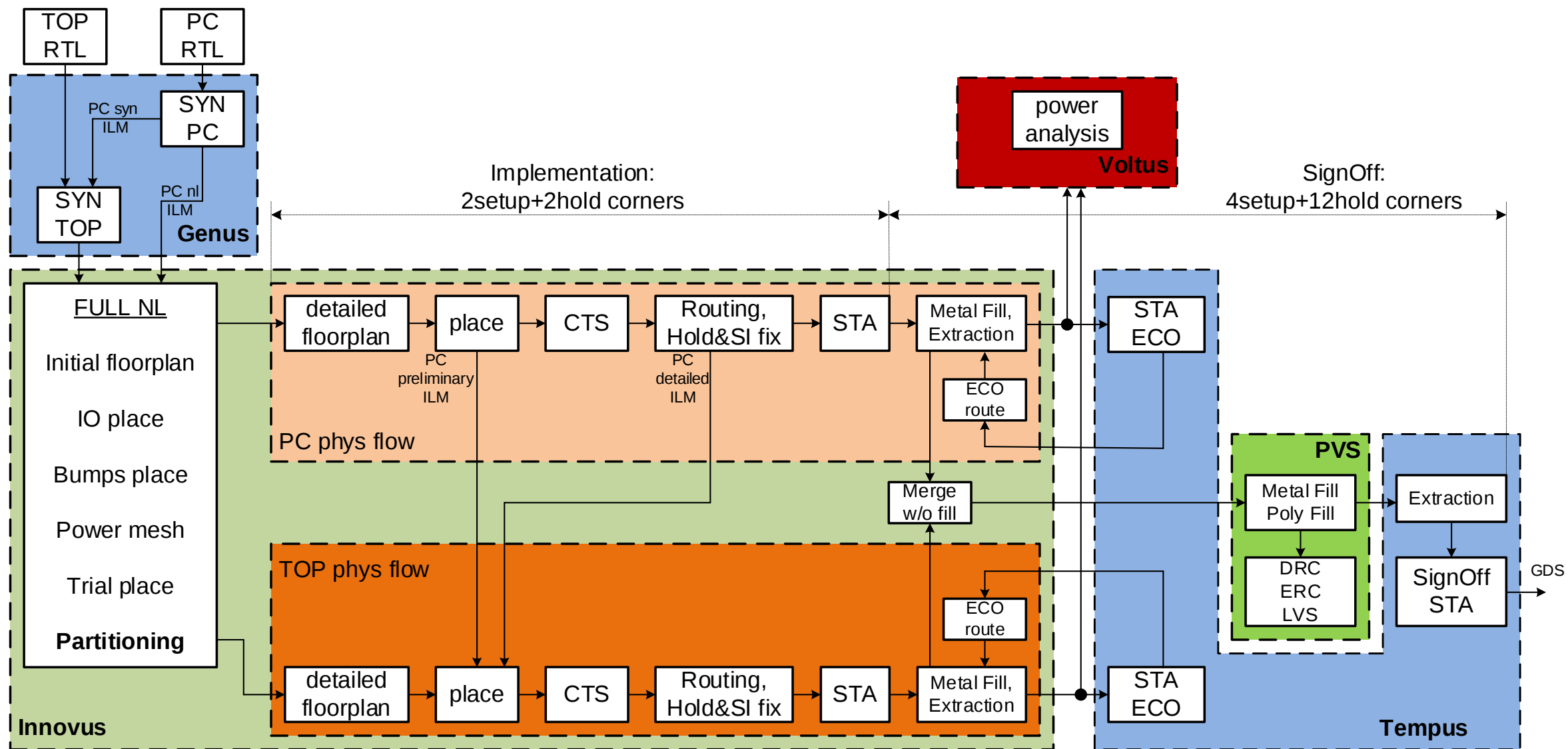
Общая структурная схема СЧК NM6408



Объединение NMPU через систему межпроцессорных каналов

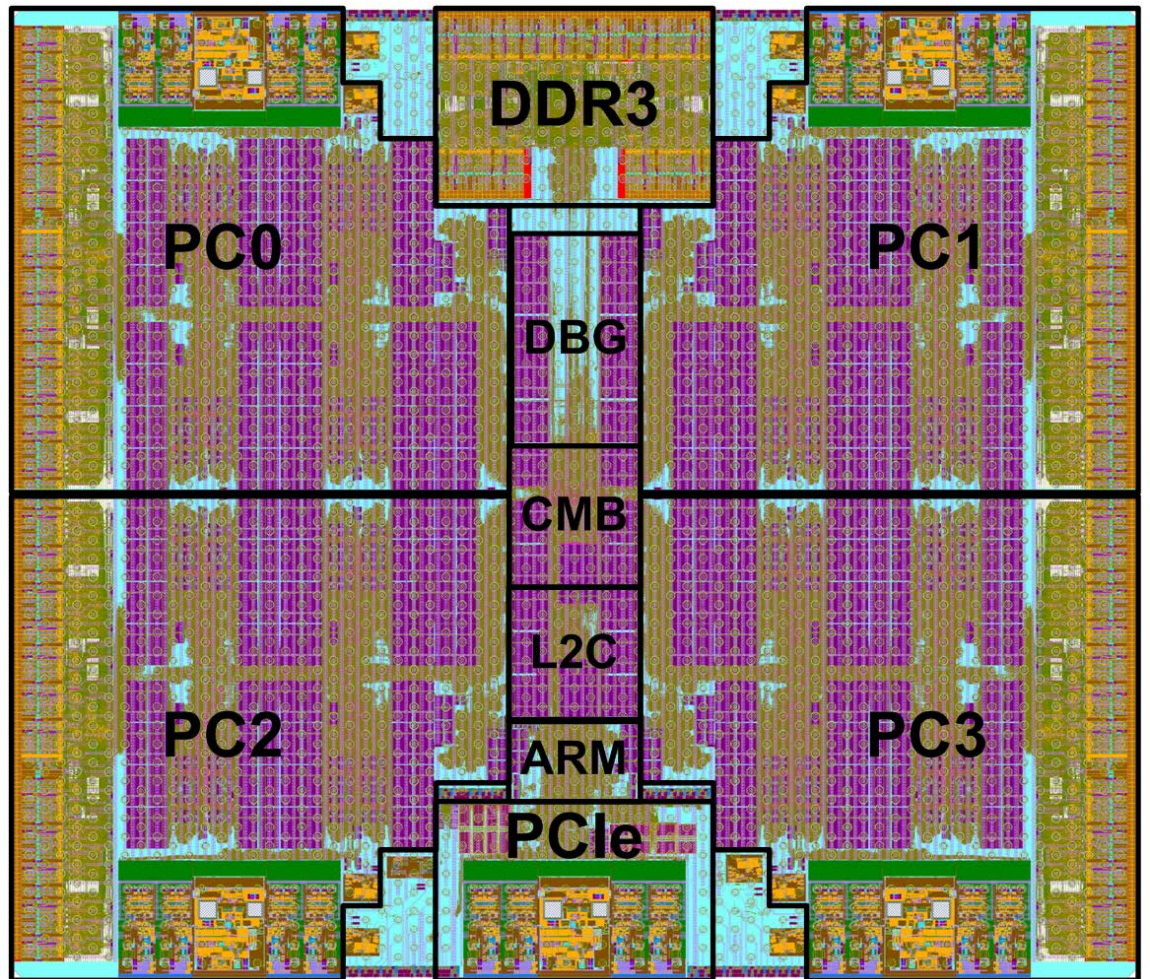
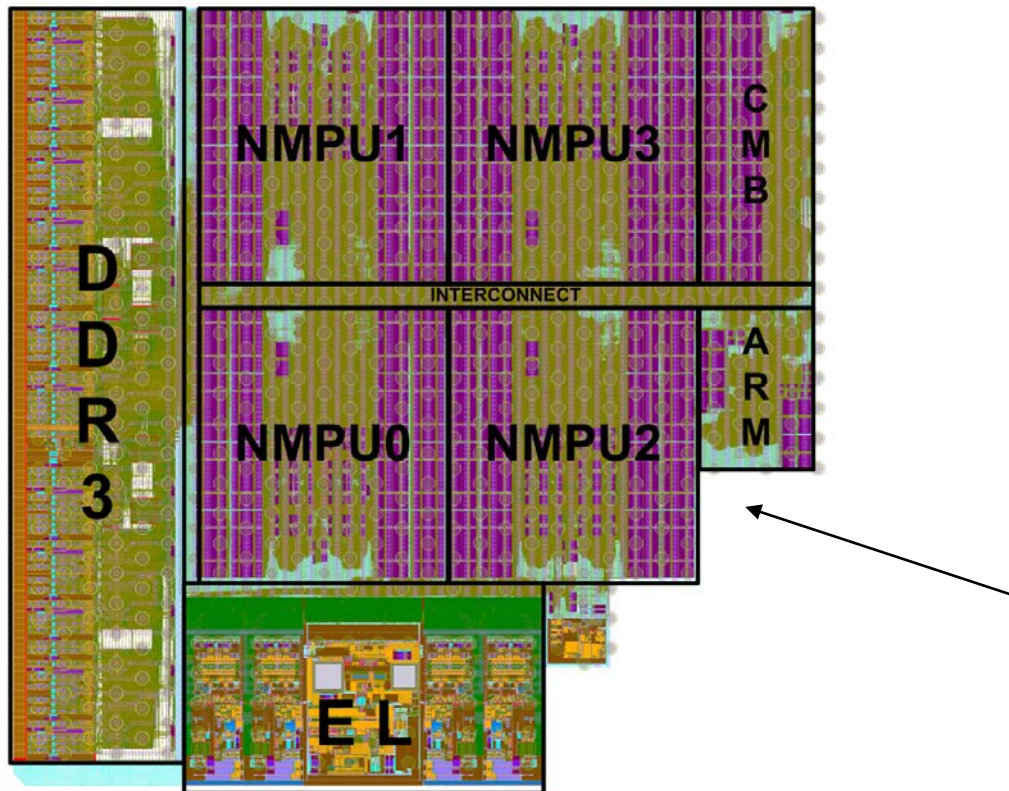


Маршрут физического проектирования (Cadence)



Полученная топология

Размер кристалла: 9.81 x 8,45 мм



Основные характеристики СнК NM6408

Состав (основные блоки):

- ▶ 16 векторных ядер NeuroMatrix NMC4
- ▶ 4 ядра ARM Cortex A5 (L1\$:32КБ+32КБ)
- ▶ 1 ядро ARM Cortex A5 (L1\$:32КБ+32КБ, L2\$:512КБ)
- ▶ внутренняя память 74МБит (9.25МБ)
- ▶ 5 интерфейсов DDR3-1600 32 разряда
- ▶ 4 внешних дуплексных коммуникационных порта 2ГБ/с в каждом направлении
- ▶ хост-интерфейс PCIe2.0 x4
- ▶ Ethernet 10/100
- ▶ SPI
- ▶ GPIO
- ▶ JTAG отладка ARM CoreSight

Основные характеристики СБИС:

- ▶ суммарная производительность
 - **FP32 – 512 Glops**
 - FP64 – 128 Gflops
- ▶ частота ядер NMC – 1GHz
- ▶ частота ядер ARM – 800MHz
- ▶ КМОП 28нм
- ▶ площадь кристалла ~83мм²
- ▶ 1.05 млрд транзисторов
- ▶ корпус BGA1444, шаг 1мм, FlipChip
- ▶ максимальная мощность – 35 Вт
- ▶ удельная производительность – 14.6 Gflops/Вт
- ▶ удельная мощность потребления – 0.42 Вт/мм²

Новизна полученных результатов

- ▶ высокая производительность СнК за счет векторных сопроцессоров:
 - СФ устройства, выполняющих до 8 FP32 операций за такт
 - потоковая обработка, за счет передачи данных через векторные регистры внутри ядра
- ▶ толерантность к задержкам обращений к памяти (высокий темп выполнения обращений)
 - векторная архитектура сопроцессора
 - сильно расслоенная, локализованная вблизи процессорных ядер внутренняя память
- ▶ энергоэффективность за счет расслоения и локализации памяти
- ▶ адаптируемость за счет гетерогенности процессорных ядер
- ▶ поддержка механизмов межпроцессорного взаимодействия
 - общее адресное пространство с блоками защиты памяти (MPU)
 - поддержка эксклюзивных операций в выделенных банках памяти (реализация семафоров)
 - асинхронная передача данных через каналы межпроцессорного обмена
 - многоярусная система межпроцессорных прерываний
- ▶ получен опыт проектирования большой СнК по технологии 28нм

Спасибо за внимание!

Вопросы?



- ▶ Internet: www.module.ru
- ▶ E-mail: rusales@module.ru
- ▶ тел.: +7 499 152-96-98
- ▶ факс: +7 499 152-46-61
- ▶ адрес: 4-я ул. 8-го Марта д. 3
а/я: 166, г. Москва,
125190, Россия

Механизмы межпроцессорного взаимодействия

- ▶ общее адресное пространство с блоками защиты памяти (MPU)
- ▶ поддержка эксклюзивных операций в выделенных банках памяти (реализация семафоров)
- ▶ асинхронная передача данных через каналы межпроцессорного обмена
- ▶ многоярусная система межпроцессорных прерываний

Программное обеспечение:

- ▶ поддержка стандартных аппаратно-программных инструментов ARM
- ▶ ассемблер
- ▶ RISC в NMC4 - компилятор C/C++
- ▶ векторный сопроцессор NMC4 – библиотеки функций на C (intrinsic, интринсики)
- ▶ параллельная модель программирования – собственная библиотека с элементами MPI и SHMEM
- ▶ средства отладки, трассировки и анализа выполнения параллельных программ

Принятые при разработке решения для достижения поставленных целей

Цели:

● **производительность**

● **программируемость**

● **гетерогенность**

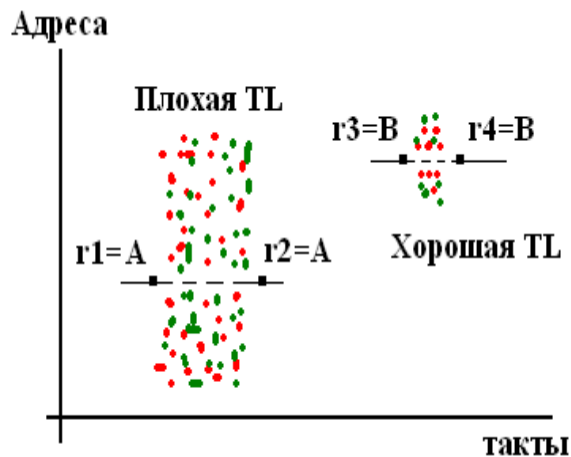
● **энергоэффективность**

- ▶ векторные устройства реализованы в виде динамически реконфигурируемых СФ устройств, способных выполнять до 8 операций за такт (●)
- ▶ для обеспечения высокого темпа обработки обращений от векторных функциональных устройств внутренняя память имеется при каждом функциональном устройстве и сильно расслоена, т.е. ярко выражена иерархия памяти (● ●)
- ▶ для сокращения количества обращений к памяти, данные в векторных устройствах передаются через большое количество векторных регистров (типа FIFO) минуя банки внутренней памяти (● ●)
- ▶ введено большое количество автономно работающих блоков ПДП (●)
- ▶ единое адресное пространство (● ●)
- ▶ используются разнотипные процессорные ядра NMC4 и Cortex-A5 (● ●)
- ▶ аппаратная поддержка средств межпроцессорного взаимодействия (● ●)

Пространственно-временная локализация обращений к памяти

Пространственная локализация (spatial locality, SL) - тенденция приложения выдавать обращения к памяти, в которых адреса находятся вблизи от адресов недавно выданных обращений.

Временная локализация (temporal locality, TL) – тенденция приложения выдавать обращения к памяти к тем же адресам, которые были в недавно выданных обращениях



Weinberg J. et al. Quantifying Locality In The Memory Access Patterns of HPC Applications. SC'05, November 12-18, 2005, 12 pp. (San Diego-LBNL)