

Аппаратная реализация искусственных нейронных сетей

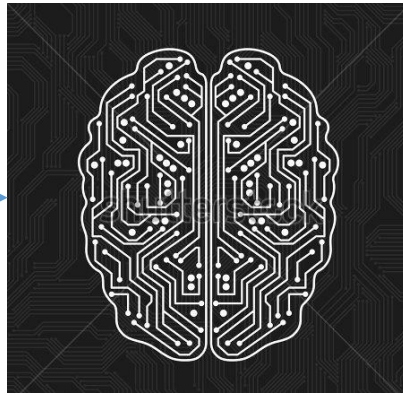
Негров Д.В., Захарченко С.В.



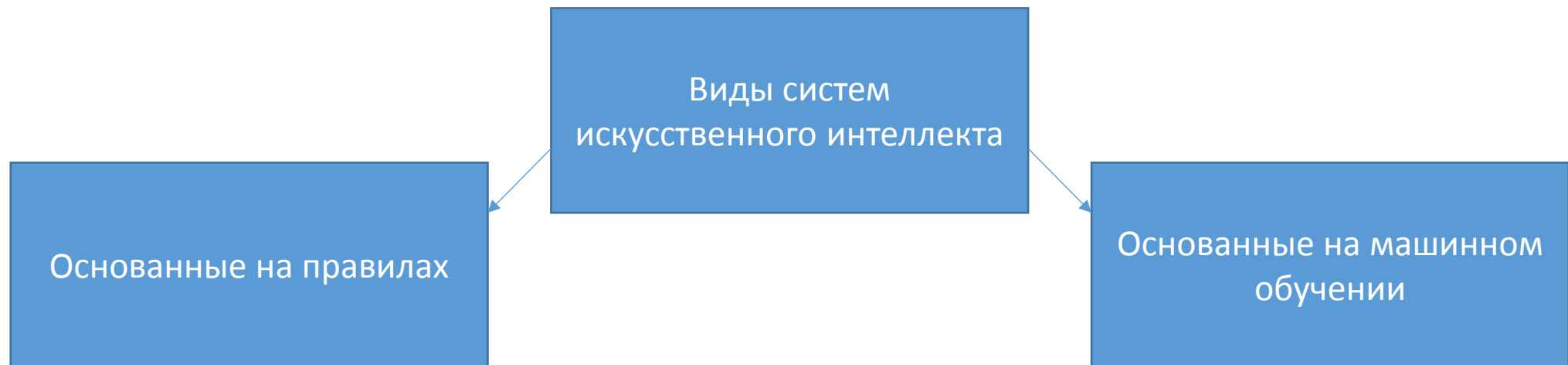
План доклада

- Машинное обучение и глубокие нейронные сети
- Особенности глубоких нейронных сетей
- Проблемы современных архитектур в задачах обучения
- Требования к эффективной архитектуре
- Обзор разрабатываемой архитектуры
- Сравнение с существующими решениями
- Реализация в виде Processor-in-memory

Подходы к созданию интеллектуальных систем

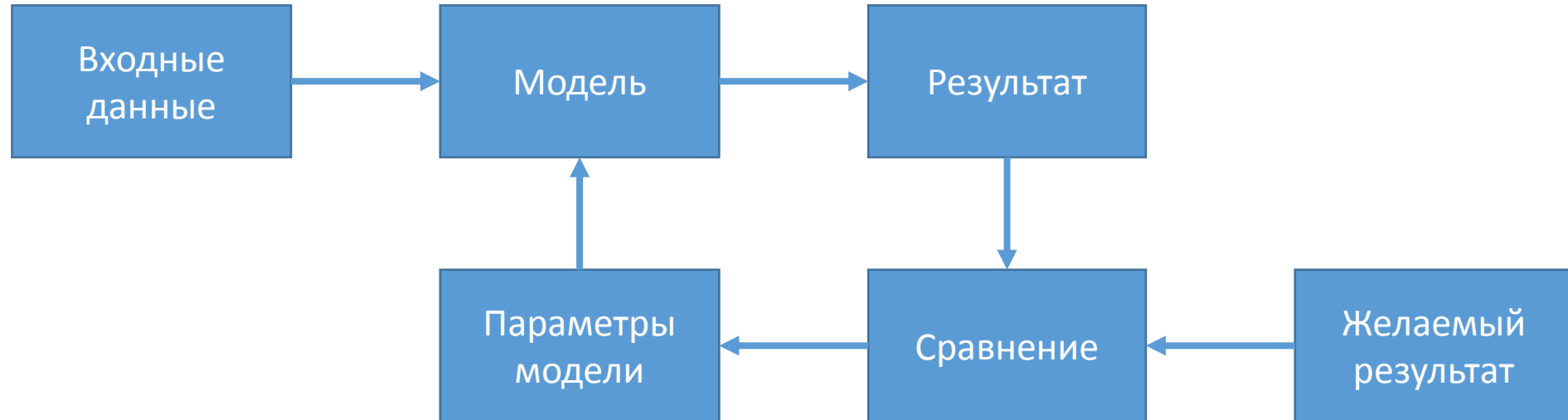


Cat

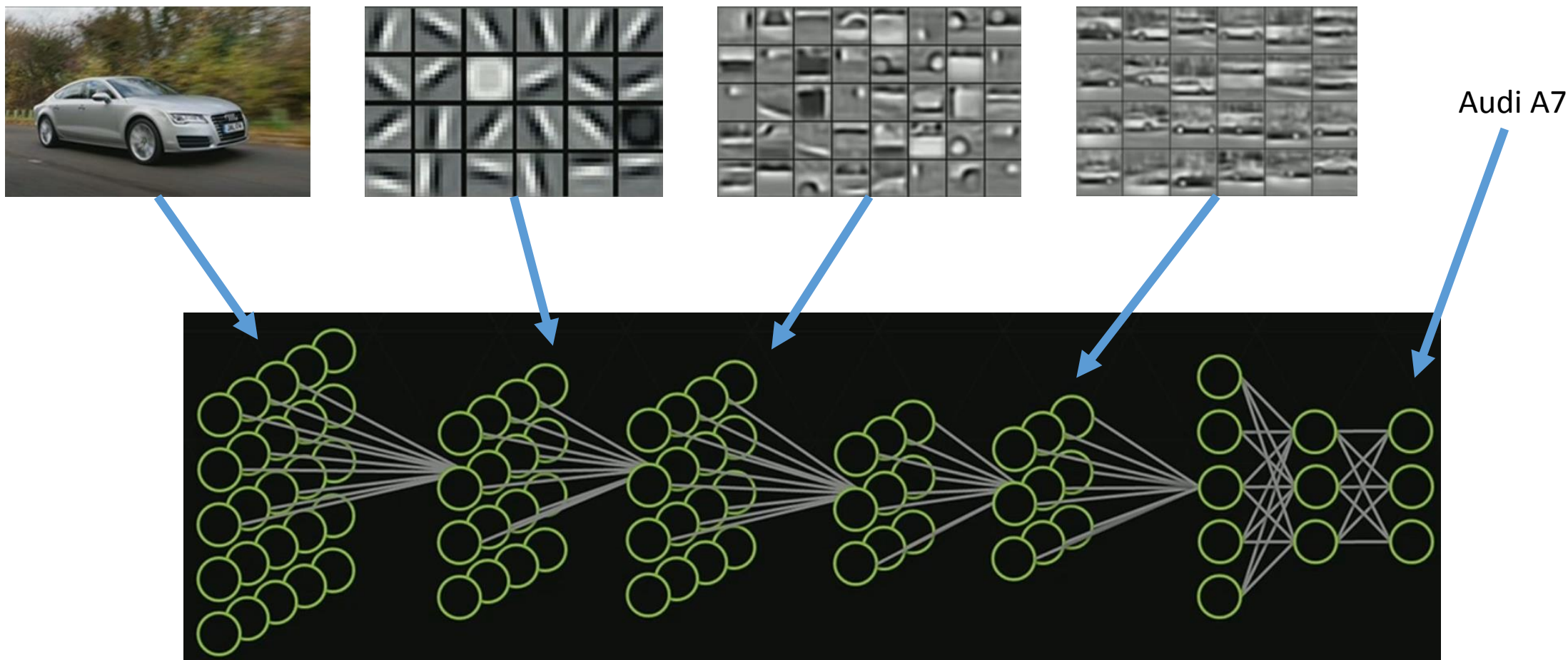


Подход глубокого машинного обучения

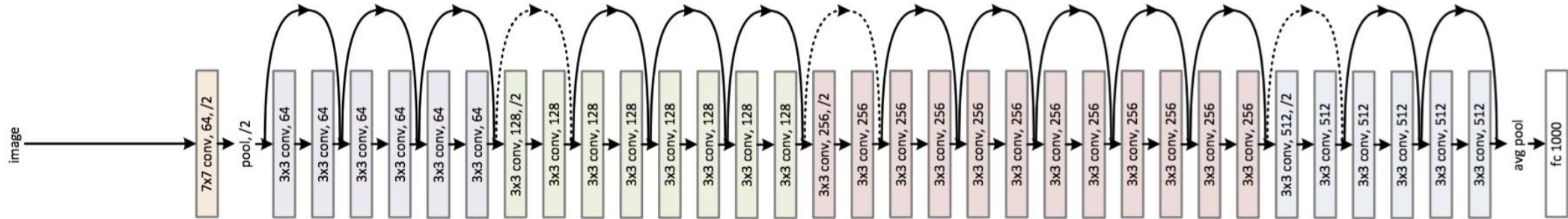
- Автоматический поиск закономерностей во входных данных, позволяющих вычислить правильный результат



Подход, основанный на искусственных нейронных сетях



Подход, основанный на ИНС



Большая глубина иерархии



Возможность анализа
абстрактных концепций

Большое количество связей



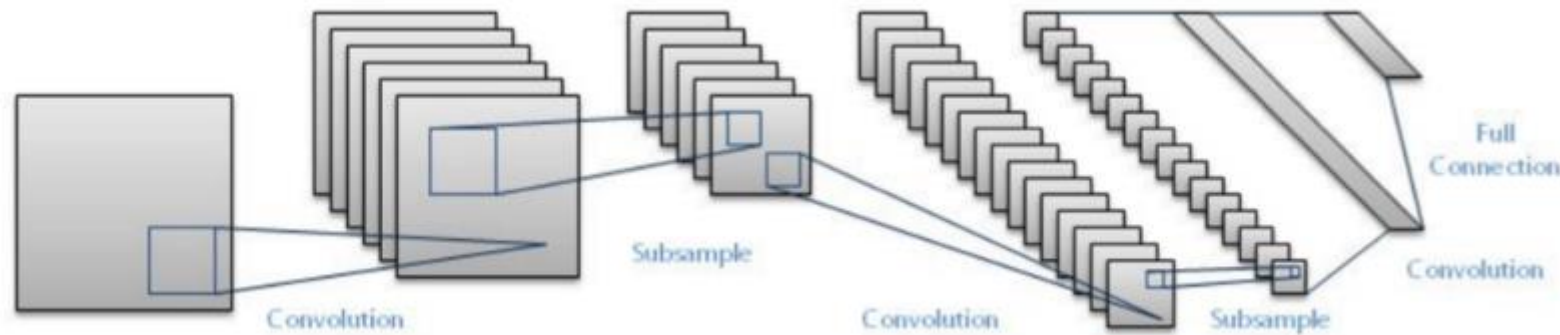
Возможность построения
сложных взаимосвязей

В определенном смысле, глубокие нейросети –
реализация графических байесовских моделей



Характерные особенности глубоких сетей

- Огромный размер модели – 10^6 (AlexNet) – 10^9 параметров.
- Большое количество необходимых вычислений.
- Специфические свойства разреженности и доступа к параметрам.
- Высокий параллелизм на уровне данных



Недостатки современных архитектур в задачах нейросетевого машинного обучения



CPU

- Предназначен для выполнения ПО общего назначения
- Сложный вычислительный конвейер
- Низкая производительность
- Высокая энергетическая стоимость операций

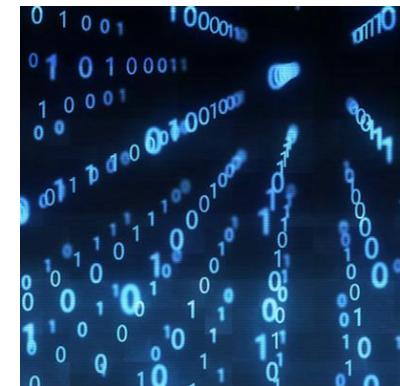
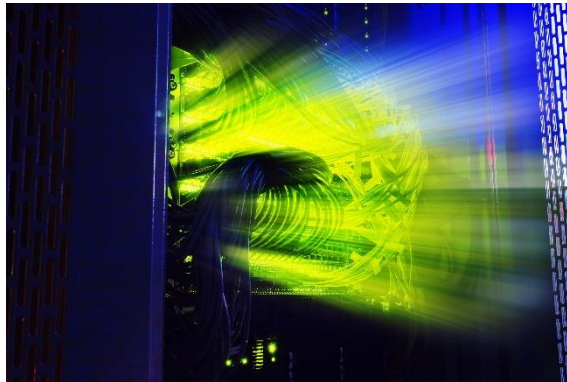
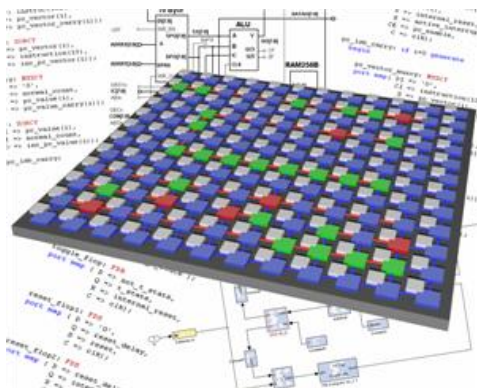


GPU

- Предназначен для выполнения большого количества однородных вычислений
- Падение производительности на ветвлениях
- Эффективно работает только с батчами
- Скрытие задержек памяти с помощью сверхдлинных конвейеров и большого числа потоков
- Средняя стоимость операции за такт

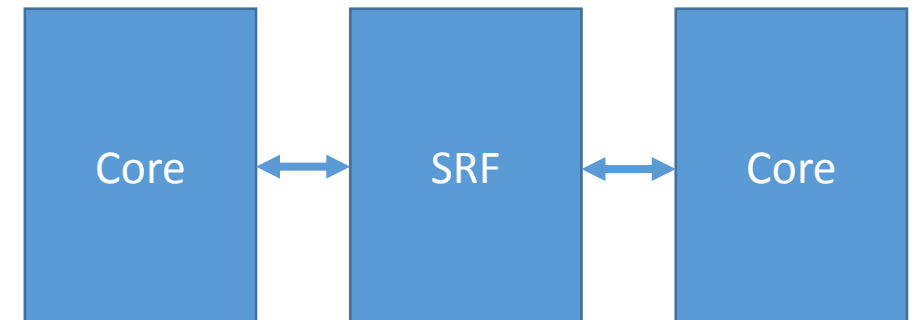
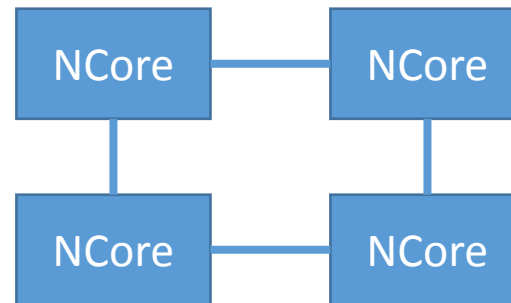
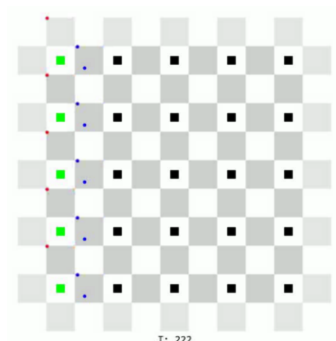
Необходимые свойства эффективной реализации

- Большое количество вычислительных модулей
- Малая длина конвейеров доступа к памяти
- Явное управление кешами или полное отсутствие внешней памяти
- Встроенная поддержка тензорных операций
- Эффективное представление данных



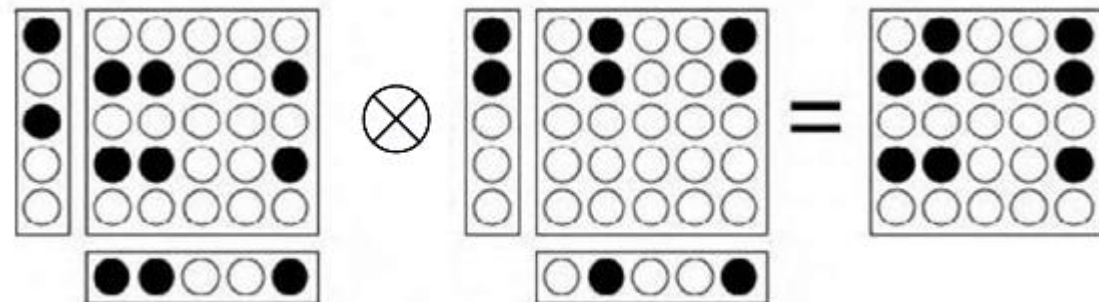
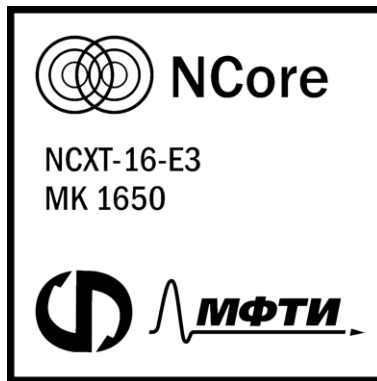
Свойства разрабатываемой архитектуры

- Большое количество компактных вычислительных ядер, погруженных в массив памяти
- Передача данных между ядрами с помощью доступа к общей памяти
- Возможность горизонтального масштабирования

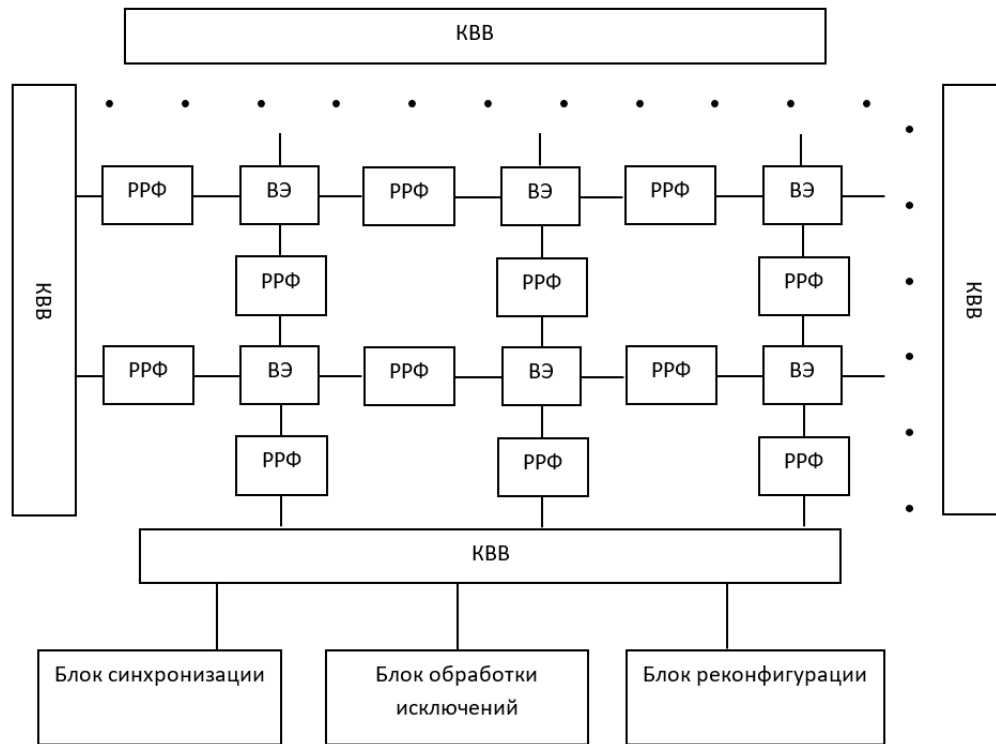


Свойства предлагаемой архитектуры

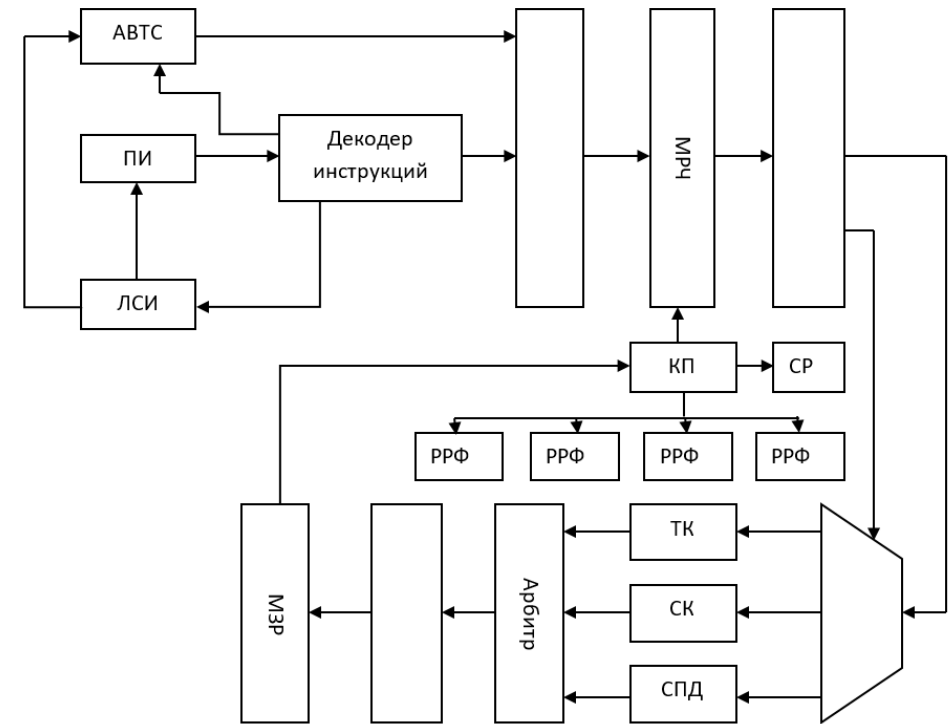
- Горизонтально масштабируемый manycore,
- Использование элементов, присущих DSP процессорам:
 - Аппаратный стек циклов
 - Инструкции векторной свертки
 - SIMD модель
- Генераторы адресов тензорных сечений
- Отсутствие блокировок в конвейерах – обработка графа вычислений при компиляции
- Возможность исполнения инструкций без их выборки



Общий вид архитектуры



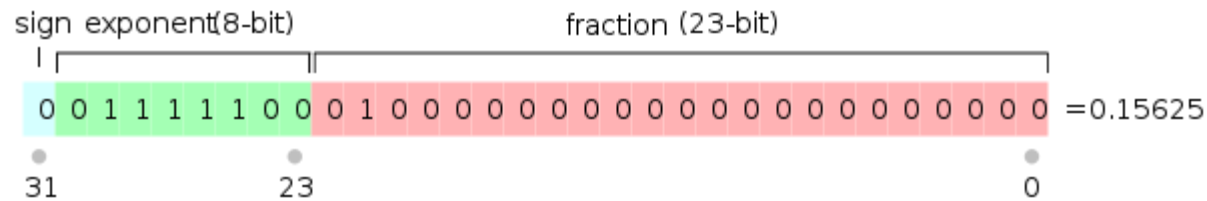
Вычислительный кластер (чип)



Исполнительное ядро

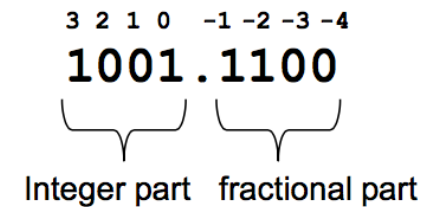
Гибкий формат данных

Floating point



- Большой размер
- Необходимость приведения экспонент при вычислениях
- Подходит для обучения

Fixed point



- Компактнее
- Малый динамический диапазон
- Не подходит для обучения при разумных размерах

Flex

Shared exponent

0110

Mantissa

0.1.0.1.1.1.0.0
1.1.0.0.0.1.0.1
0.0.1.1.1.0.0.0
0.1.1.1.1.0.1.0

- Большой динамический диапазон
- Ускорение тензорных операций за счет исключения приведения экспонент
- Эффективное использование памяти
- Подходит для обучения

Выполняемые алгоритмы

- Произвольно программируемая архитектура с PRAM параллелизмом и явным управлением памятью.
- Возможна реализация новых алгоритмов без изменений кремния.
- Эффективное выполнение произвольных алгоритмов тензорной алгебры:
 - Разреженные и полносвязные нейронные сети;
 - Рекуррентные нейронные сети;
 - Эффективное обучение на единичных векторах (без батчей).



Статус проекта

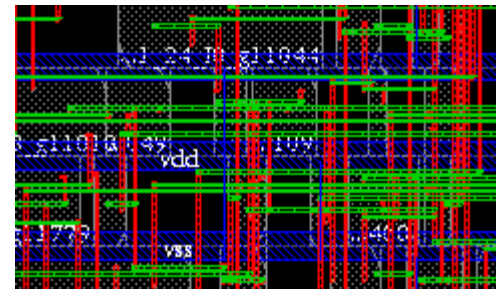
- Реализованы IP ядра системы, способные использоваться как в SoC, так и самостоятельном кристалле.
- Закончены системы компиляции промежуточных представлений, транслятор наиболее распространенных алгоритмов и интерфейс к DL фреймворкам (TensorFlow, Theano, Caffe2).
- Готова система симуляции для оптимизации алгоритмов.
- Создан стек ПО для работы с системой (TensorFlow target).
- Реализован демонстратор на FPGA (Xilinx 7 series).
- Tapeout менее, чем через месяц.



theano



Caffe2



Сравнение с другими системами

	Ncore 2x2	Ncore 16x16	TPU	TPU2	Volta	Nervana	BM1680
Техпроцесс	65 nm	65 nm	28 nm	28 nm	12 nm	TBA	28 nm
Площадь кристалла	~ 10 mm ²	4*20 mm ²	331 ²	?	815 mm ²	TBA	?
Производительность, TOP/S	0,1	6,4	23	45	100*	TBA	2
Тип данных	Flex16	Flex16	FxP8	FP23	FP32	Flex16	FP32 (?)
TDP, W	2	14	40	200	300	TBA	41
Энергоэффективность, TOP/J	0,05	0,45	0,58	0,22	0,33	TBA	0,05
Формат инструкций, Программная модель, Программируемость	RISC PRAM FLEX	RISC PRAM FLEX	CISC Systolic PREFEF	CISC Systolic PREDEF	RISC SIMT FLEX	? ? ?	RISC (?) ? FLEX
Применение	LEARN INFER FLEX	LEARN INFER FLEX	- INFER RIGID	LEARN INFER RIGID	LEARN INFER FLEX	LEARN INFER FLEX	?

* - Только для тензорных операций 4x4*4x4+4x4 (FP16)

In-memory processing

Доступ во внешнюю DDR память – медленная и дорогая операция



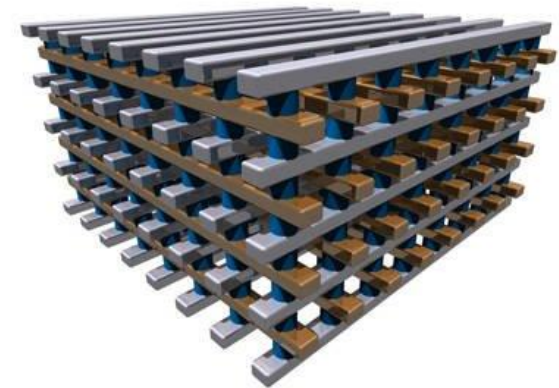
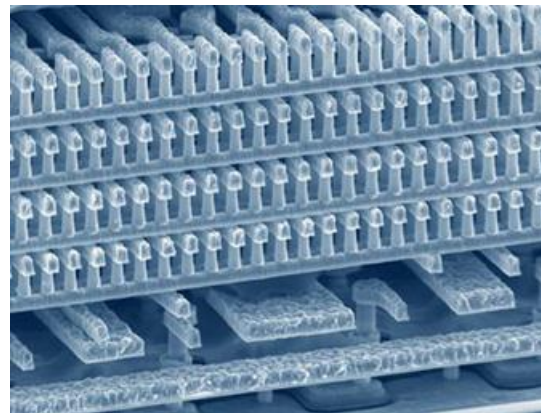
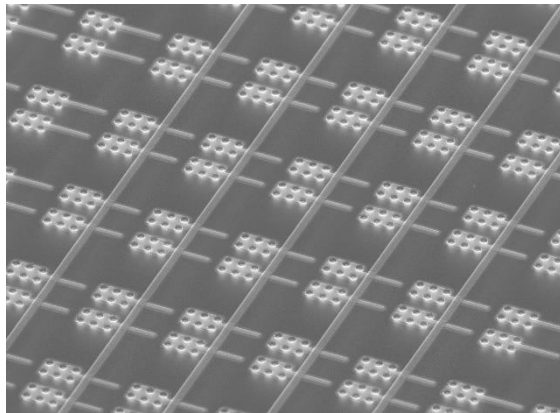
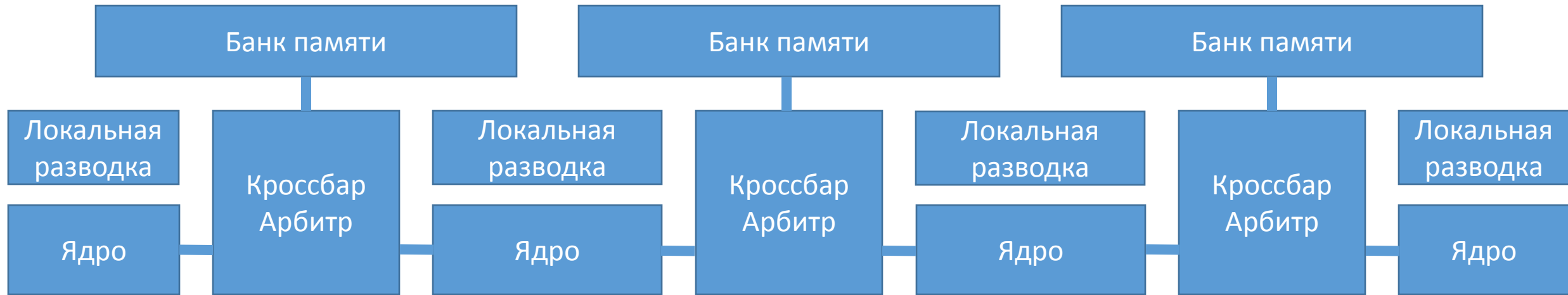
Возможность значительно ускорить вычисления и повысить энергоэффективность за счет размещения вычислительных ядер непосредственно в памяти

Новые виды памяти (ReRAM, FeRAM на оксидах переходных металлов):

- Имеют высокую плотность и возможность интеграции в логический процесс.
- Могут быть интегрированы в BEOL процесс.
- Обеспечивают крайне низкое потребление энергии.

Возможно достичь плотности памяти, сравнимой с DRAM и значительно большей пропускной способности по сравнению с HBM при значительно более низком энергопотреблении.

Возможности реализации РИМ на перспективных видах энергонезависимой памяти



Спасибо за внимание

