Реализация потактового эмулятора нейросетевого ускорителя на основе VERILATOR

Сергей Мушкаев

начальник сектора прикладного программного обеспечения, компания НТЦ «Модуль»

2022

Module RESEARCH CENTRE

МЕТОДЫ МОДЕЛИРОВАНИЯ

- **RTL – моделирование**
  - Потактовая моделирование
  - Временные диаграммы
  - Низкая скорость (100 тактов/сек)
- **ПЛИС – прототипирование**
  - Высокая скорость (70 МГц)
  - Ограничение по объему ПЛИС
- **VERILATOR – моделирование**
  - Потактовое моделирование
  - X10-x20 скорость моделирования
  - Временные диаграммы
  - C++, Трассы, дампы, временные метрики
  - Многопоточность
  - Масштабируемость
  - Открытый исходный код.

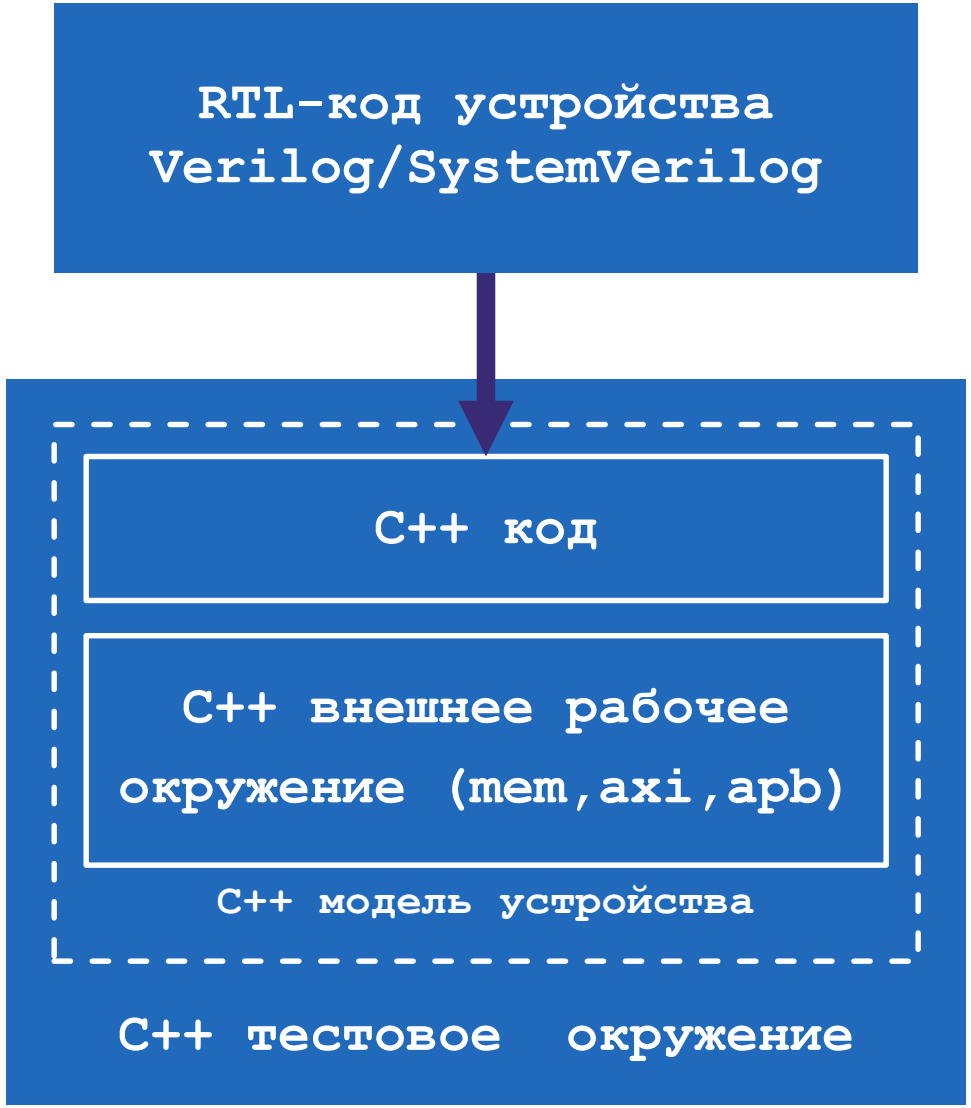
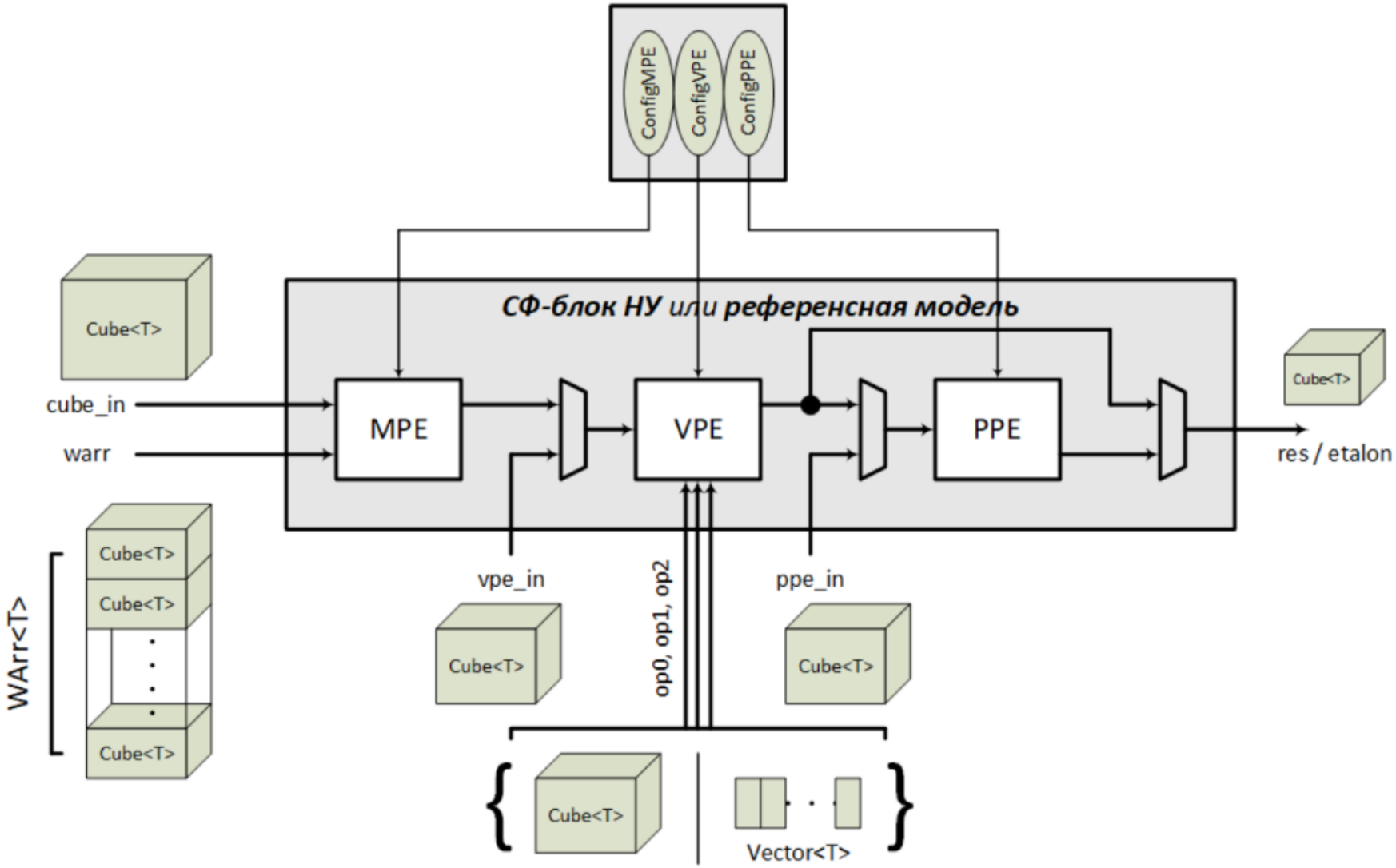
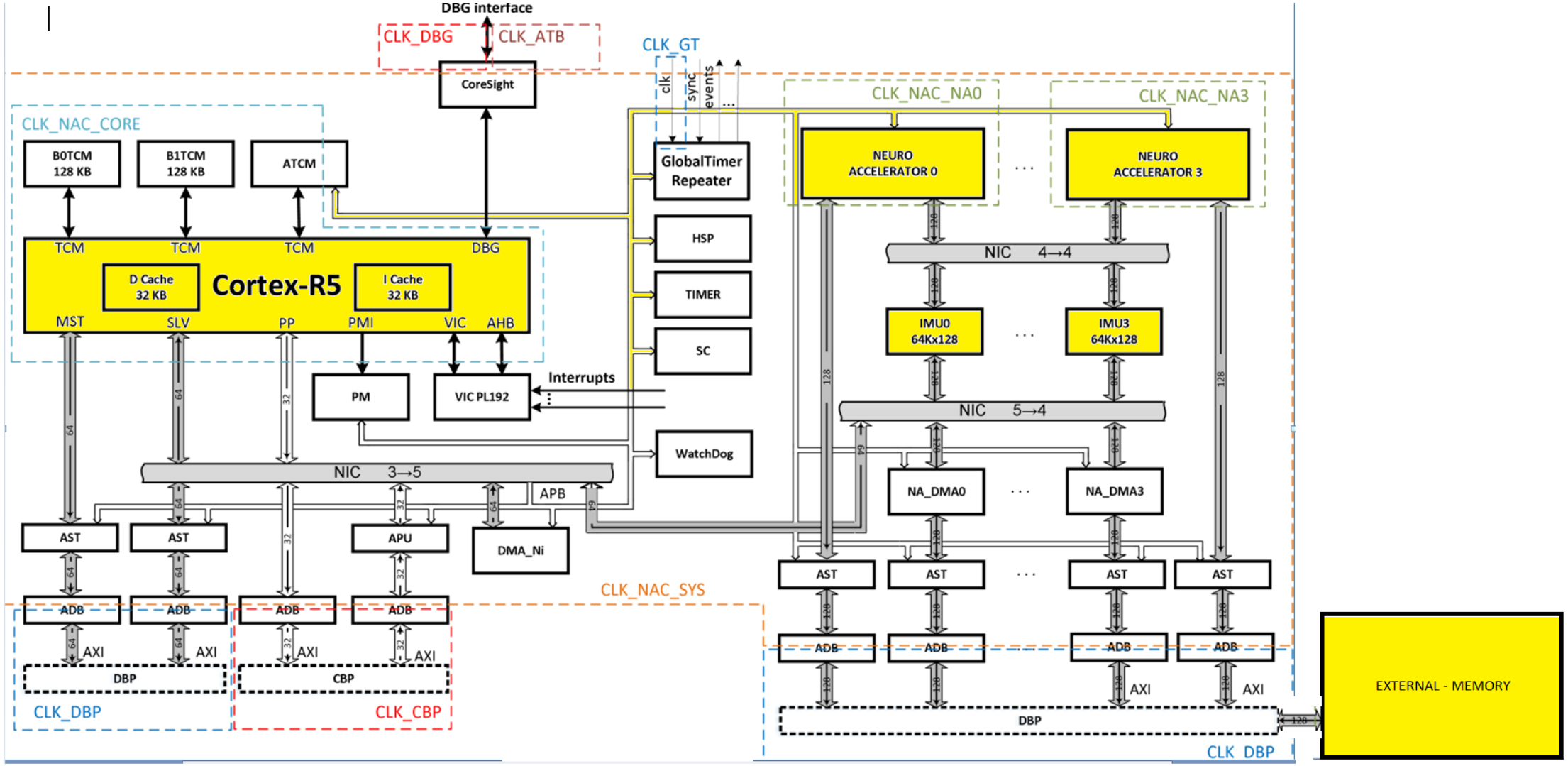


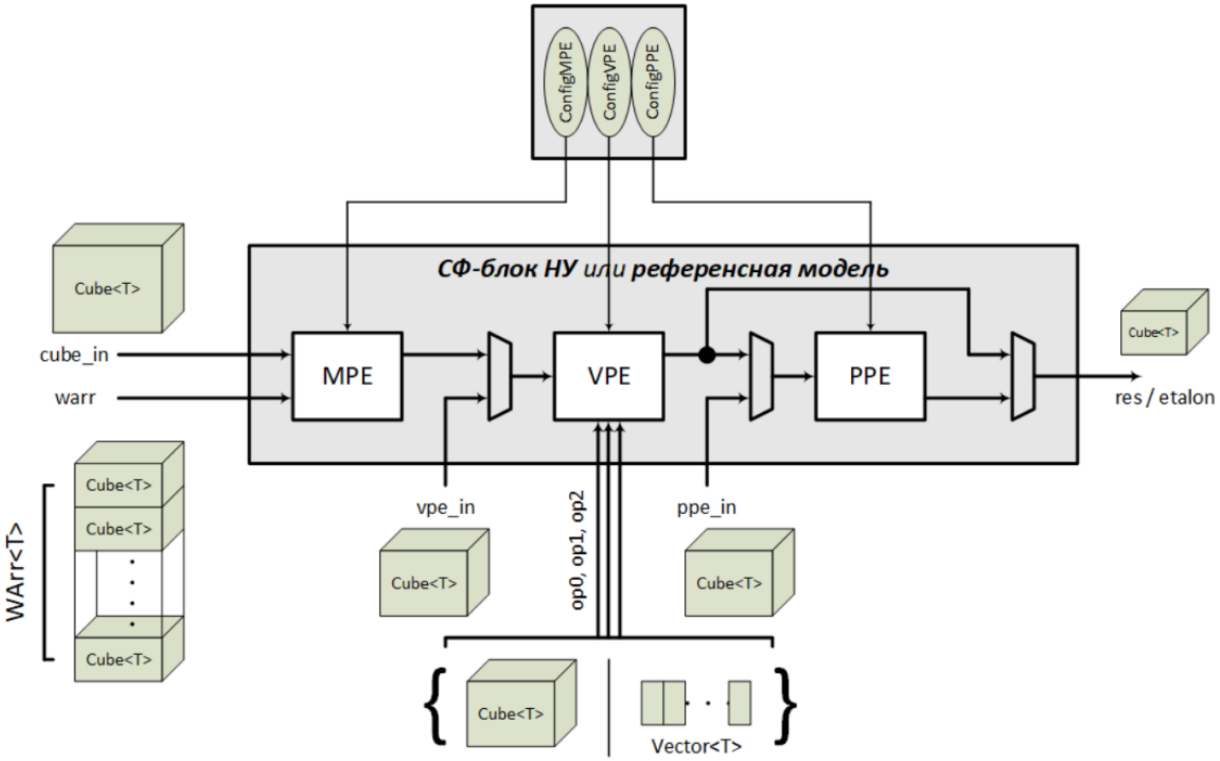
СХЕМА ПРЕОБРАЗОВАНИЯ ДАННЫХ СФ-БЛОКОМ НЕЙРОСЕТЕВОГО УСКОРИТЕЛЯ

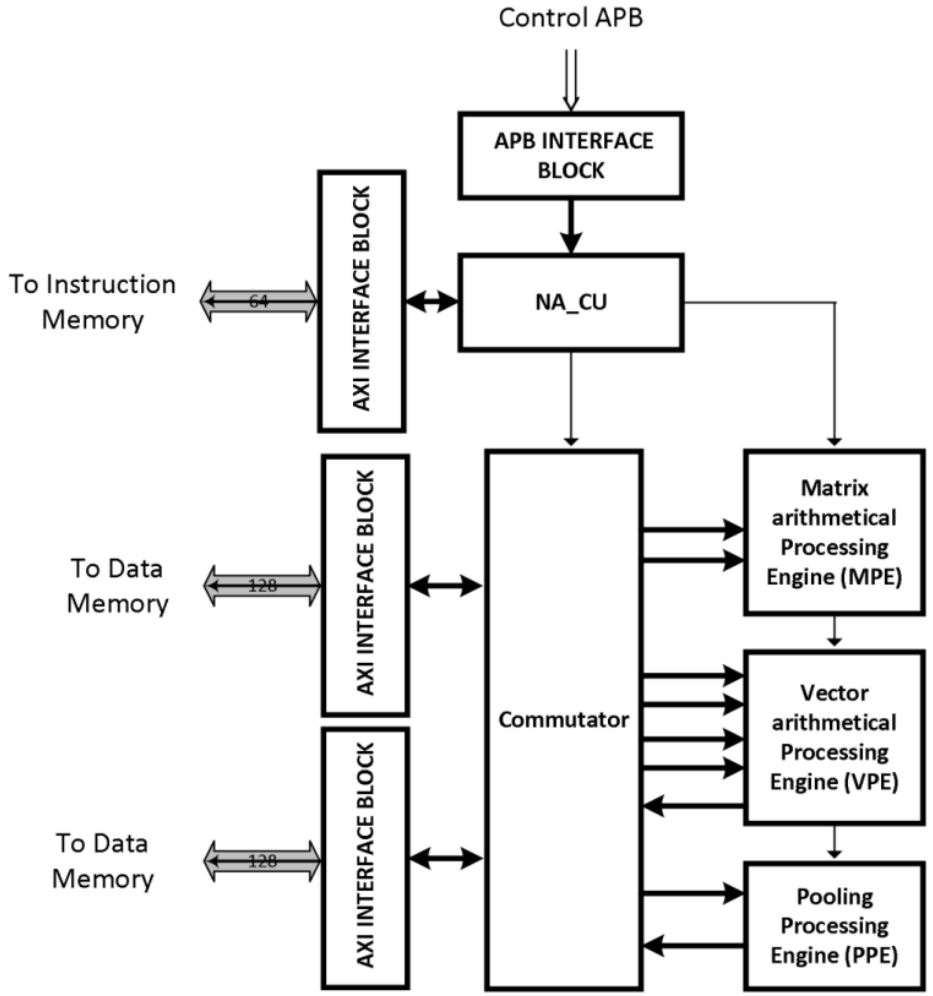


Кластер нейросетевых ускорителей (NAC)



СТРУКТУРНАЯ СХЕМА НЕЙРОСЕТЕВОГО УСКОРИТЕЛЯ

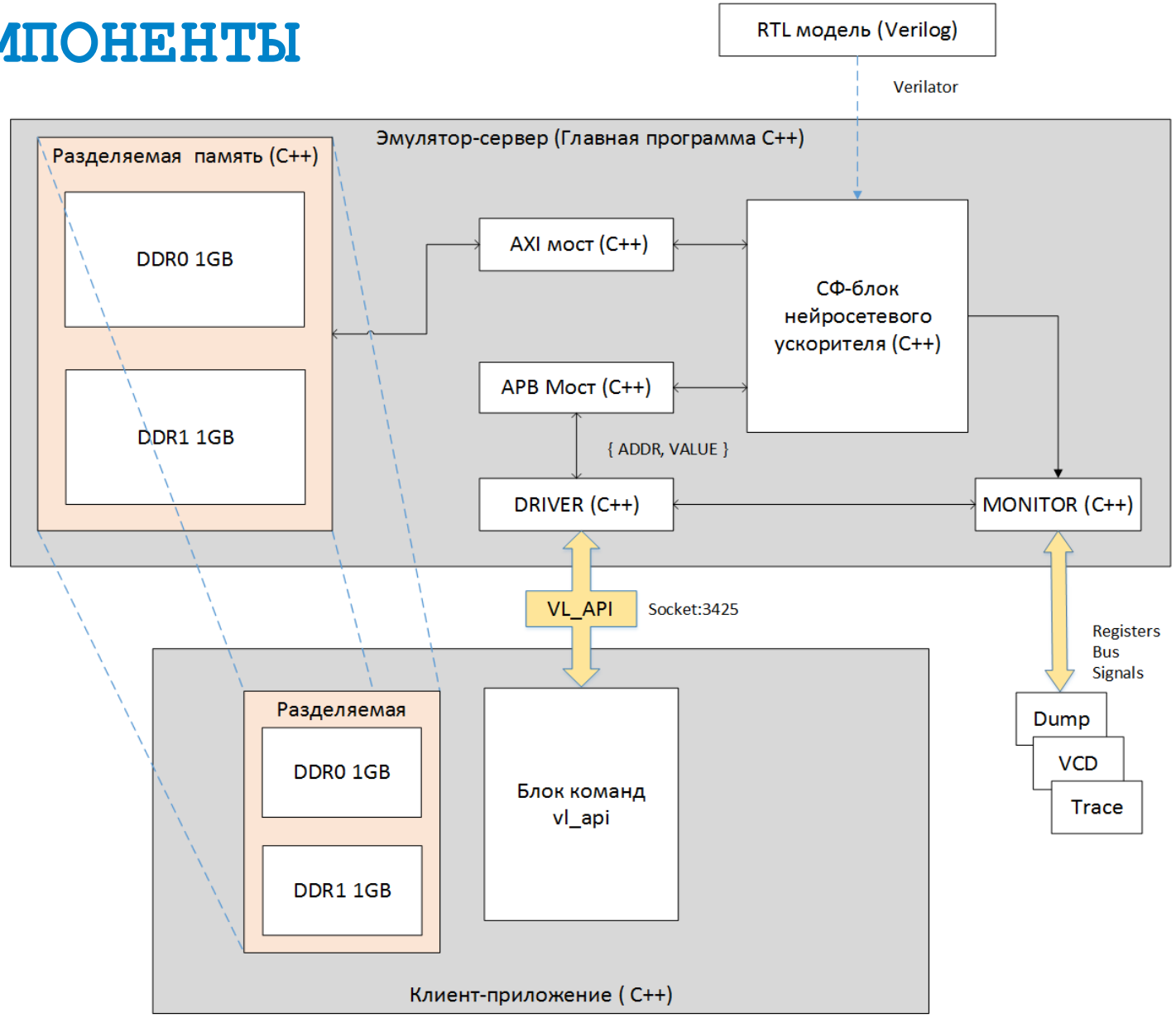




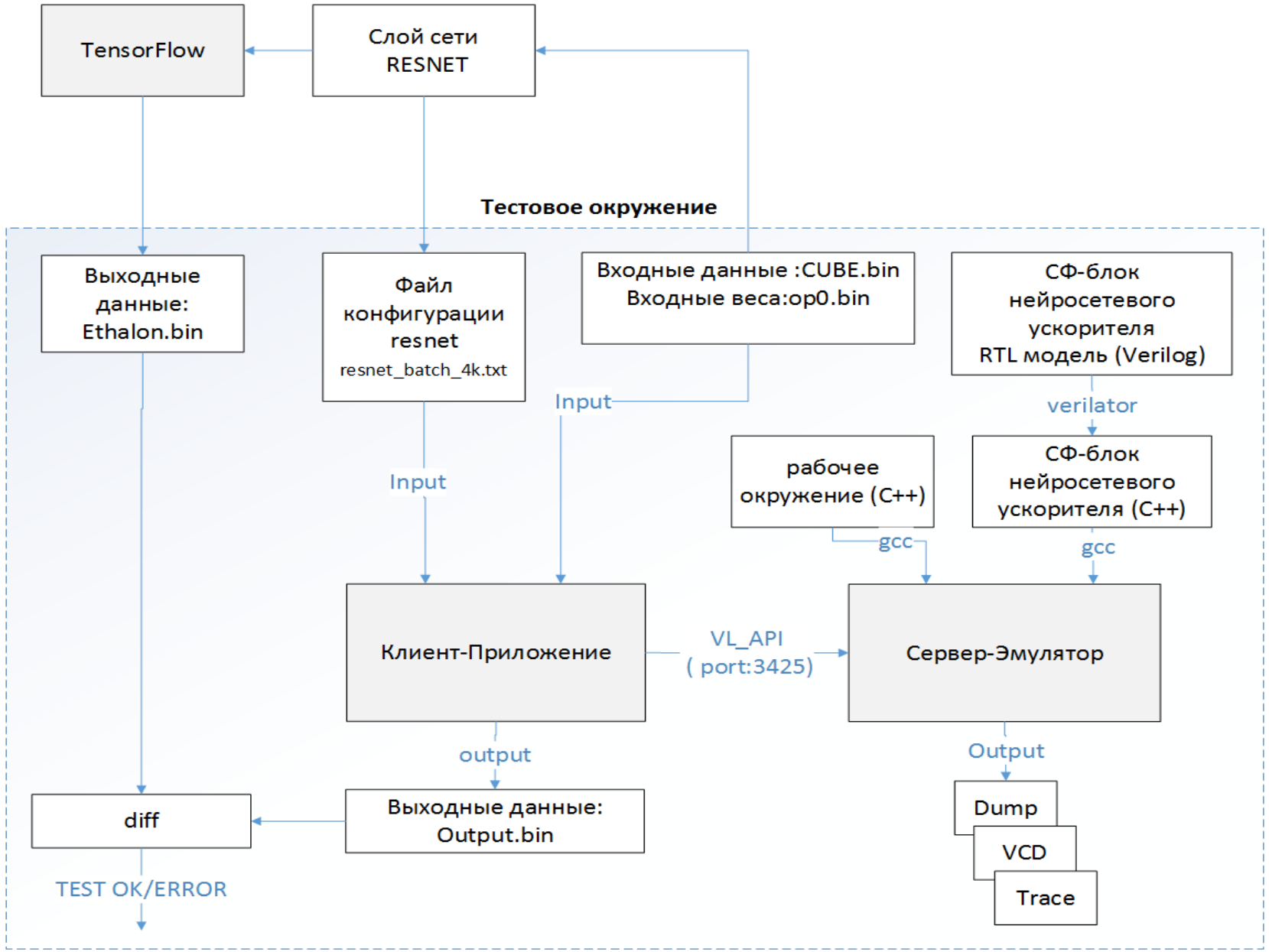
MPE - Матричный арифметический блок, используемый для вычисления свертки;
VPE – Векторный арифметический блок, используемый для поэлементных операций и функции активации;
PPE – Блок субдискретизации.

Module
НАУЧНО-ТЕХНИЧЕСКИЙ ЦЕНТР

СТРУКТУРА ЭМУЛЯТОРА И ЕГО ОСНОВНЫЕ КОМПОНЕНТЫ



ТЕСТОВОЕ ОКРУЖЕНИЕ



ВЫХОДНЫЕ МЕТРИКИ ЭМУЛЯТОРА

Трасса изменения регистров

```
(5343) FULL_BUFF[03c003ff]
(5349) FULL_BUFF[03c007ff]
(5358) FULL_BUFF[07c007ff]
(5471) FULL_BUFF[078007ff]
(5489) FULL_BUFF[07800fff]
(5498) FULL_BUFF[0f800fff]
(5599) FULL_BUFF[0f000fff]
(5629) FULL_BUFF[0f001fff]
(5638) FULL_BUFF[1f001fff]
(5727) FULL_BUFF[1e001ffe]
(5743) FULL_BUFF[1e001ffc]
(5759) FULL_BUFF[1e001ff8]
(5769) FULL_BUFF[1e003ff8]
(5775) FULL_BUFF[1e003ff0]
(5778) FULL_BUFF[3e003ff0]
(5791) FULL_BUFF[3e003fe0]
(5807) FULL_BUFF[3e003fc0]
(5823) FULL_BUFF[3e003f80]
(5839) FULL_BUFF[3e003f00]
(5855) FULL_BUFF[3c003f00]
(5909) FULL_BUFF[3c007f00]
(5918) FULL_BUFF[7c007f00]
(6049) FULL_BUFF[7c00ff00]
(6058) FULL_BUFF[fc00ff00]
(6081) FULL_BUFF[f800ff00]
(6189) FULL_BUFF[f800ff01]
(6198) FULL_BUFF[f801ff01]
(6209) FULL_BUFF[f001ff01]
(6330) FULL_BUFF[f001ff03]
(6337) FULL_BUFF[e001ff03]
(6338) FULL_BUFF[e003ff03]
(6465) FULL_BUFF[c003ff03]
(6470) FULL_BUFF[c003ff07]
```

Трасса чтения записи управляющих регистров

```
(504) reg[0x00001340] write 0x00008000
(515) reg[0x00001344] write 0x00000020
(526) reg[0x00001348] write 0x00000020
(537) reg[0x0000134c] write 0x00000020
(548) reg[0x00001350] write 0x00000002
(559) reg[0x00001354] write 0x0000007f
(570) reg[0x00001358] write 0x00000000
(581) reg[0x0000135c] write 0x00000007
(592) reg[0x00001360] write 0x0000007f
(603) reg[0x00001364] write 0x00000000
(614) reg[0x00001368] write 0x00000007
(625) reg[0x0000136c] write 0x00000000
(636) reg[0x00001370] write 0x00000000
(647) reg[0x00001374] write 0x00000000
(658) reg[0x00001378] write 0x00000000
(669) reg[0x0000137c] write 0x00000000
(680) reg[0x00001400] write 0x00000050
(691) reg[0x00001408] write 0x0001ffff
(702) reg[0x0000140c] write 0x0002000f
(713) reg[0x00001420] write 0x00000000
(724) reg[0x00001424] write 0x0000007f
(735) reg[0x00001428] write 0x0000007f
(746) reg[0x0000142c] write 0x0000007f
(757) reg[0x00001430] write 0x00007f00
(768) reg[0x00001434] write 0x003f8000
(779) reg[0x00001438] write 0x000000e0
(790) reg[0x0000143c] write 0x00000100
(801) reg[0x00001440] write 0x00008000
(812) reg[0x00001444] write 0x00000020
(823) reg[0x00001448] write 0x00000020
(834) reg[0x0000144c] write 0x00000020
(845) reg[0x00001450] write 0x00000002
```

VCD Временные диаграммы

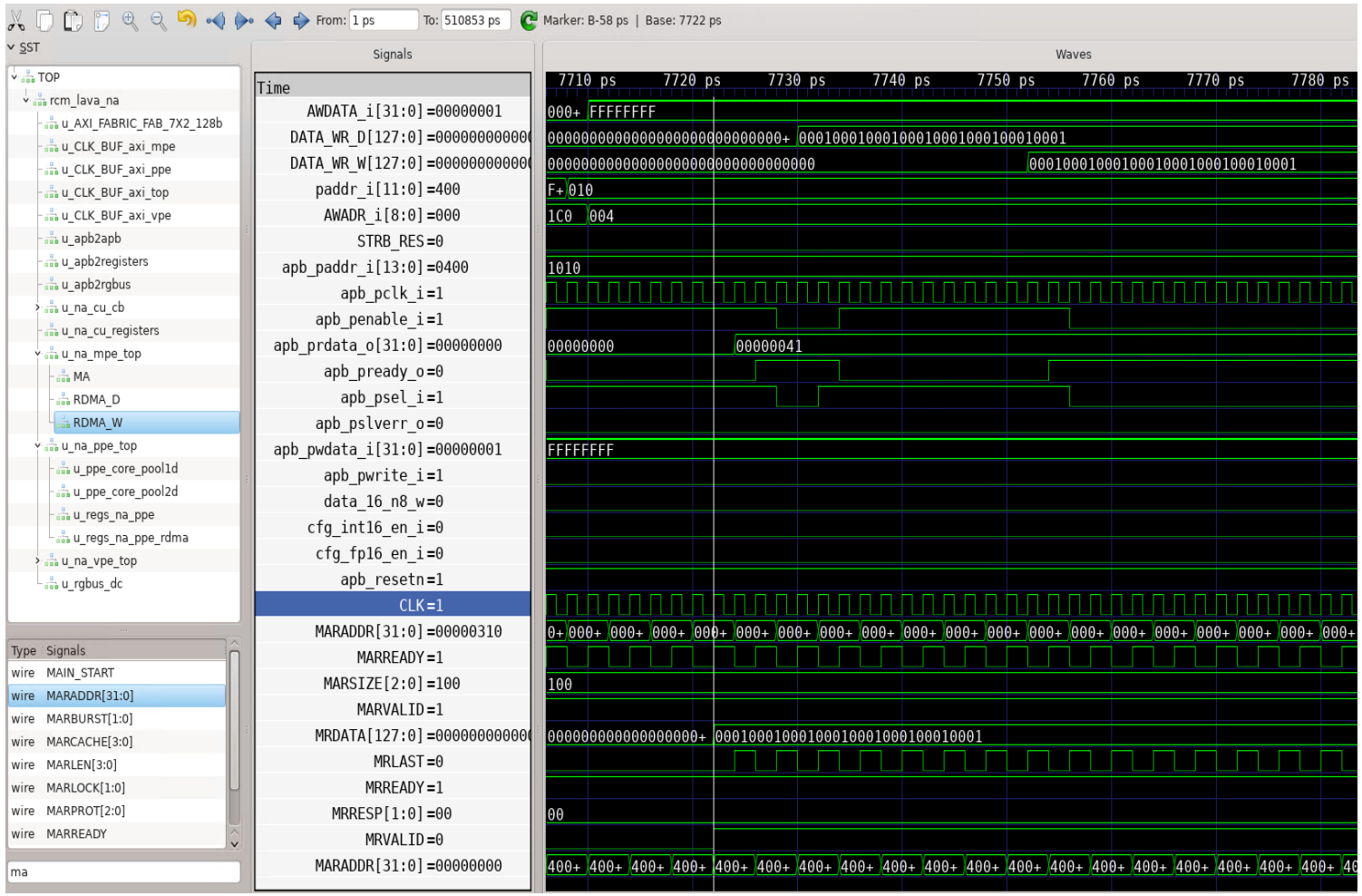


СХЕМА СОВМЕСТНОГО МОДЕЛИРОВАНИЯ ARM ЯДРА И СФ-БЛОКА НЕЙРОСЕТЕВОГО УСКОРИТЕЛЯ

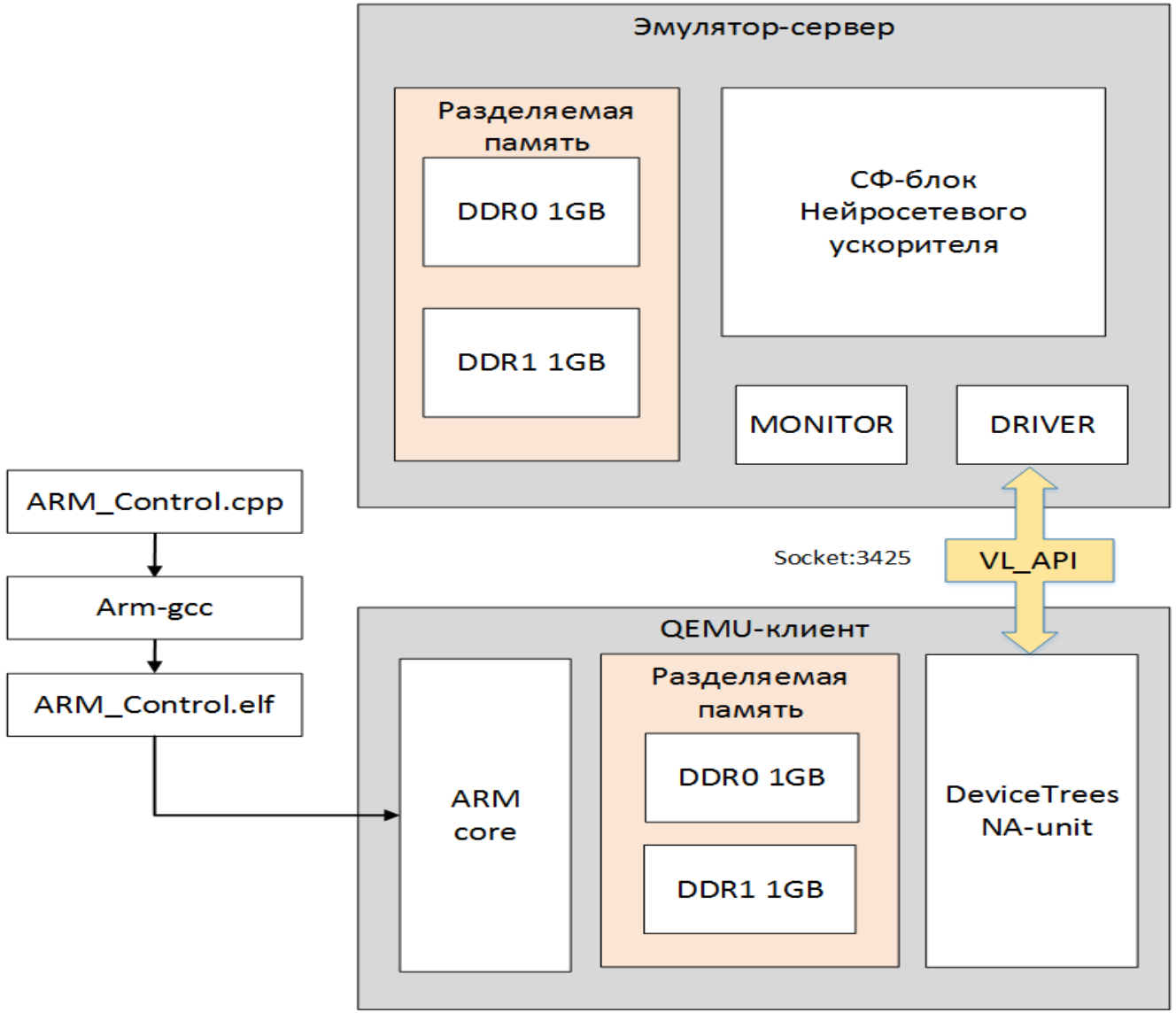
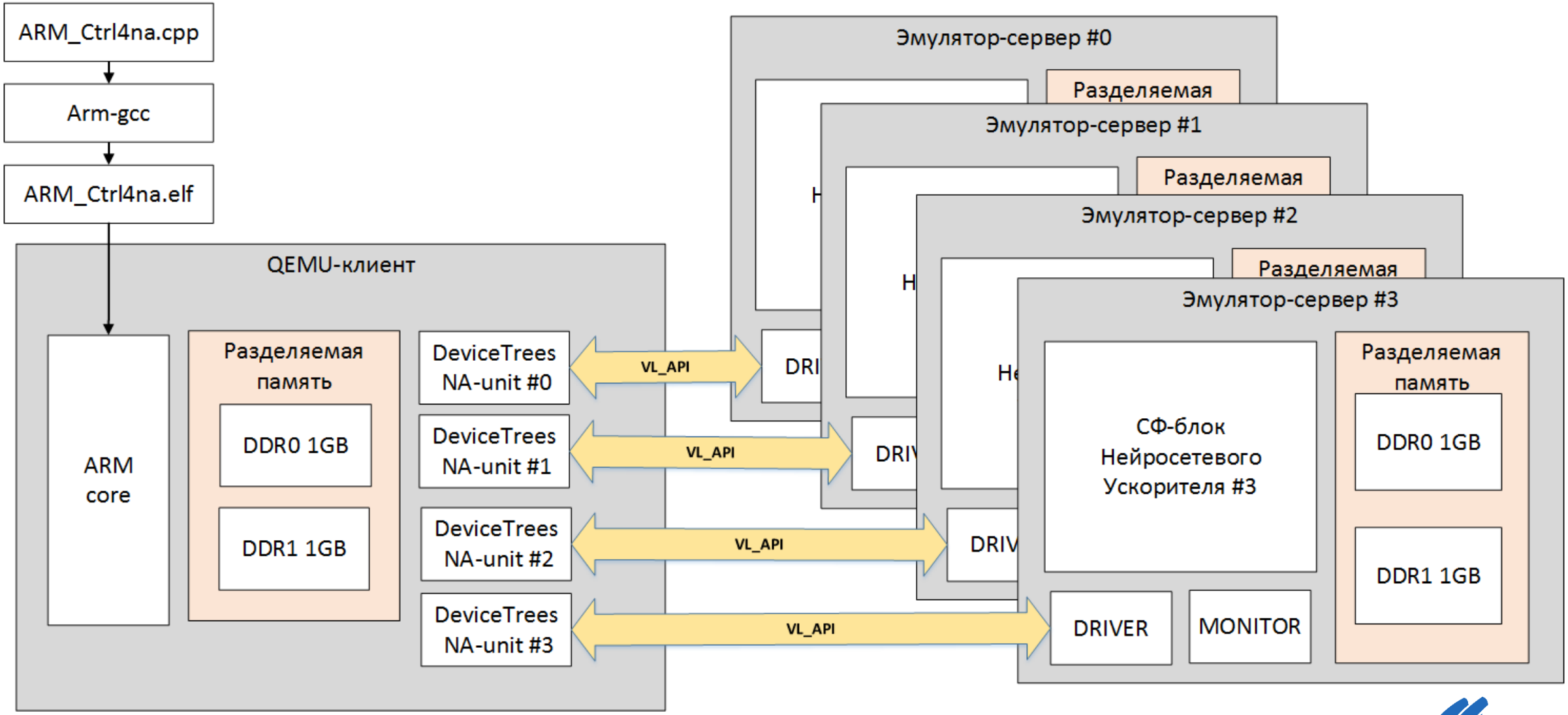


СХЕМА СОВМЕСТНОГО МОДЕЛИРОВАНИЯ КЛАСТЕРА ИЗ 4-Х СФ-БЛОКОВ И ARM-ЯДРА



СРАВНЕНИЕ ПРОИЗВОДИТЕЛЬНОСТИ ЭМУЛЯТОРА И RTL-МОДЕЛИРОВАНИЯ

Моделирование	Время моделирования	Реальное время	Такты	Скорость моделирования такты/сек
RTL	1:18:28 (78 мин)	459263 ns	431 000	91
VERILATOR	359.1сек (6 мин)			1200

Ускорение – 13 раз и более

Module
НАУЧНО-ТЕХНИЧЕСКИЙ ЦЕНТР

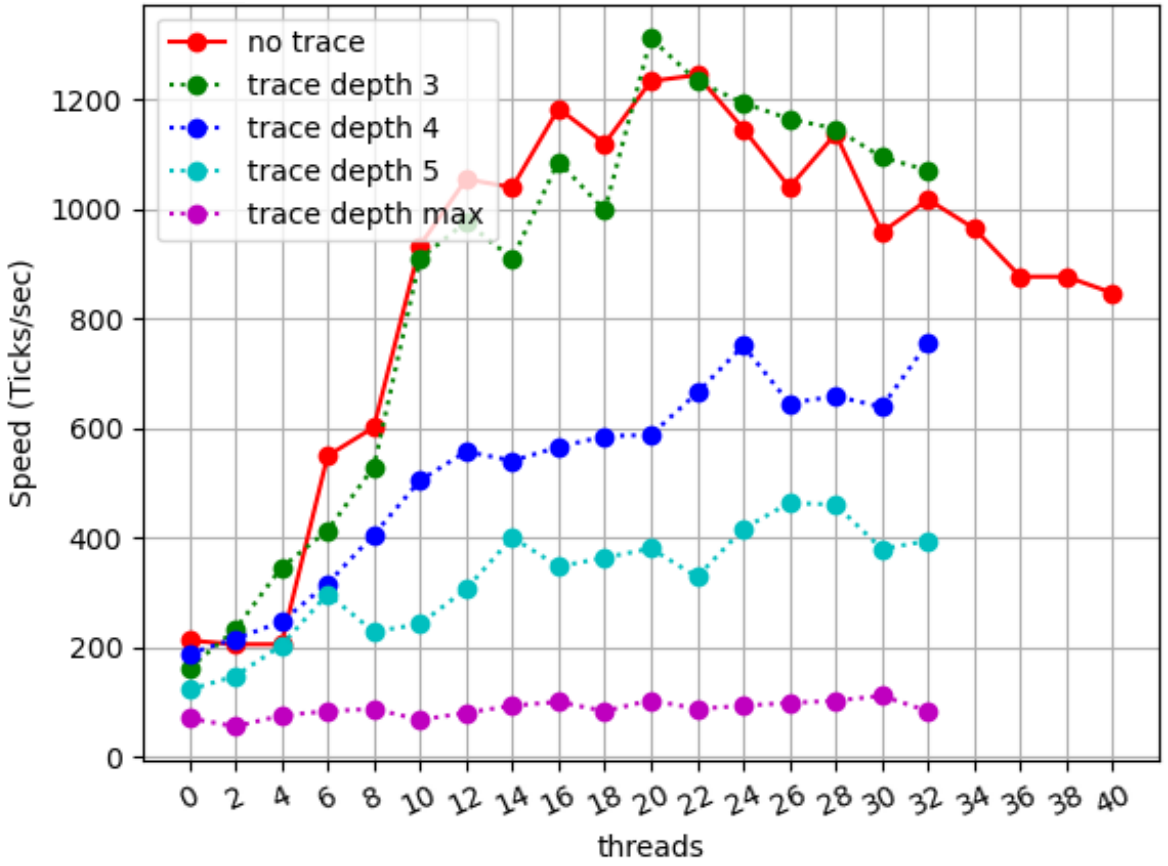
ОЦЕНКА ПРОИЗВОДИТЕЛЬНОСТИ ЭМУЛЯТОРА В НЕЙРОСЕТВЫХ ЗАДАЧАХ

Сеть	Кол-во операций	Время моделирования	Такты	Скорость Такты/сек
YOLOv2	14732084224			
YOLOv3	32932037632			
Resnet50	4089184256	8067.45s (2:14:27)	14 807 000	1847
Alexnet	716767232	7226.08s (2:00:26)	14 095 000	1981
SqueeseNet	818924576	2901.27s (0:48:21)	5 545 000	1921
Inception_v3	5713216096			
NasNet	314415872			
MobileNet27	300774272			
GAN	428343296			
SOD	26484498432			
Inception_v4	6120358752			
MobileNet3	216589760			

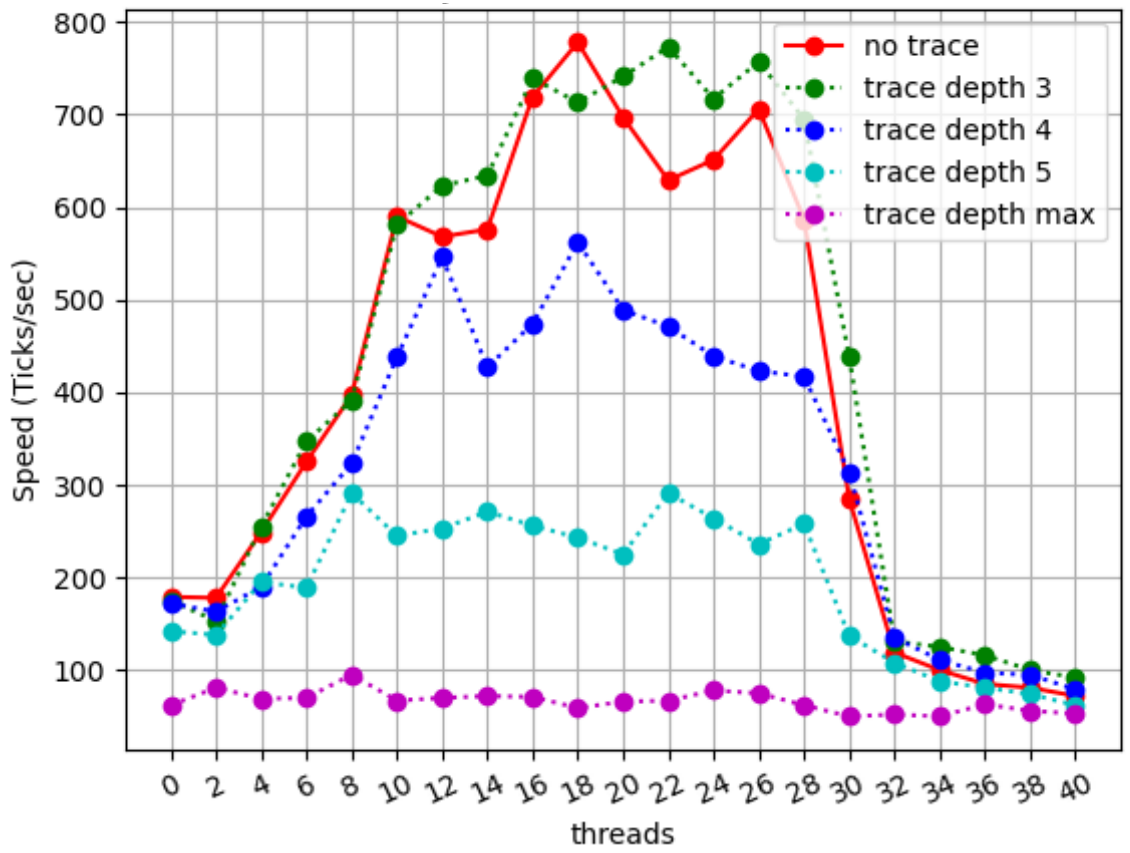
Module
НАУЧНО-ТЕХНИЧЕСКИЙ ЦЕНТР

АНАЛИЗ МНОГОПОТОЧНОСТИ

2-процессорный сервер на 104 ядра (hyper-threading)



2-процессорный сервер на 32 ядра (hyper-threading)



РЕЗУЛЬАТЫ ПО РАЗРАБОТКЕ ЭМУЛЯТОРА

- Разработан изолированный точный потактовый эмулятор СФ-блока нейросетевого ускорителя .
- Разработан внешний C++ интерфейс для подключения к Эмулятору сторонних приложений. C++/ Python
- Разработан эмулятор QEMU ядра ARM, обеспечивающего совместное моделирование как с одним потактовым эмулятором СФ-блока нейросетевого ускорителя так и с кластером из 4 сф-блоков.
- Обеспечена работа Эмулятора в ОС Linux (Ubunta, Red Hat, Debian), Windows 7,10.
- Достигнуты значения скорости эмуляции тейросетевых задач в 1500-2000 тактов в секунду (0.5- 2 часа)

Module
НАУЧНО-ТЕХНИЧЕСКИЙ ЦЕНТР

ВЫВОДЫ, ПРЕИУЩЕСТВА VERILATOR

- Потактовое моделирование
- x10-x20 скорость моделирования относительно стандартного RTL-моделирования
- Временные диаграммы VCD настраиваемой глубиной трассировки
- C++ код (простота разработки пользовательских приложений)
- Масштабируемость (построение сложных моделей кластеров, СнК)
- Создание произвольных трасс, дампов памяти, и других временных метрик
- Многопоточность (оптимизация производительности)
- Компьютер (от 16-32 MB RAM. Linux /Windows)
- Открытый исходный код

Спасибо за внимание!

Контакты

www.module.ru

Москва, 4-я улица 8 Марта, д.3
Россия, 125190, г. Москва, а/я 166

тел.: +7 495 531-3080
факс: +7 499 152-4661

rusales@module.ru

Module
RESEARCH CENTRE