xperience / ai

512 КБ памяти хватит всем! ReID с камерой на MCU

Александр Сморкалов

Xperience.AI

Приложения:

- Компьютерное зрение
- Ретейл
- Автоматизация

Платформы:

- Сервер / ПК
- Мобильные
- ПЛИС
- Встраиваемые системы

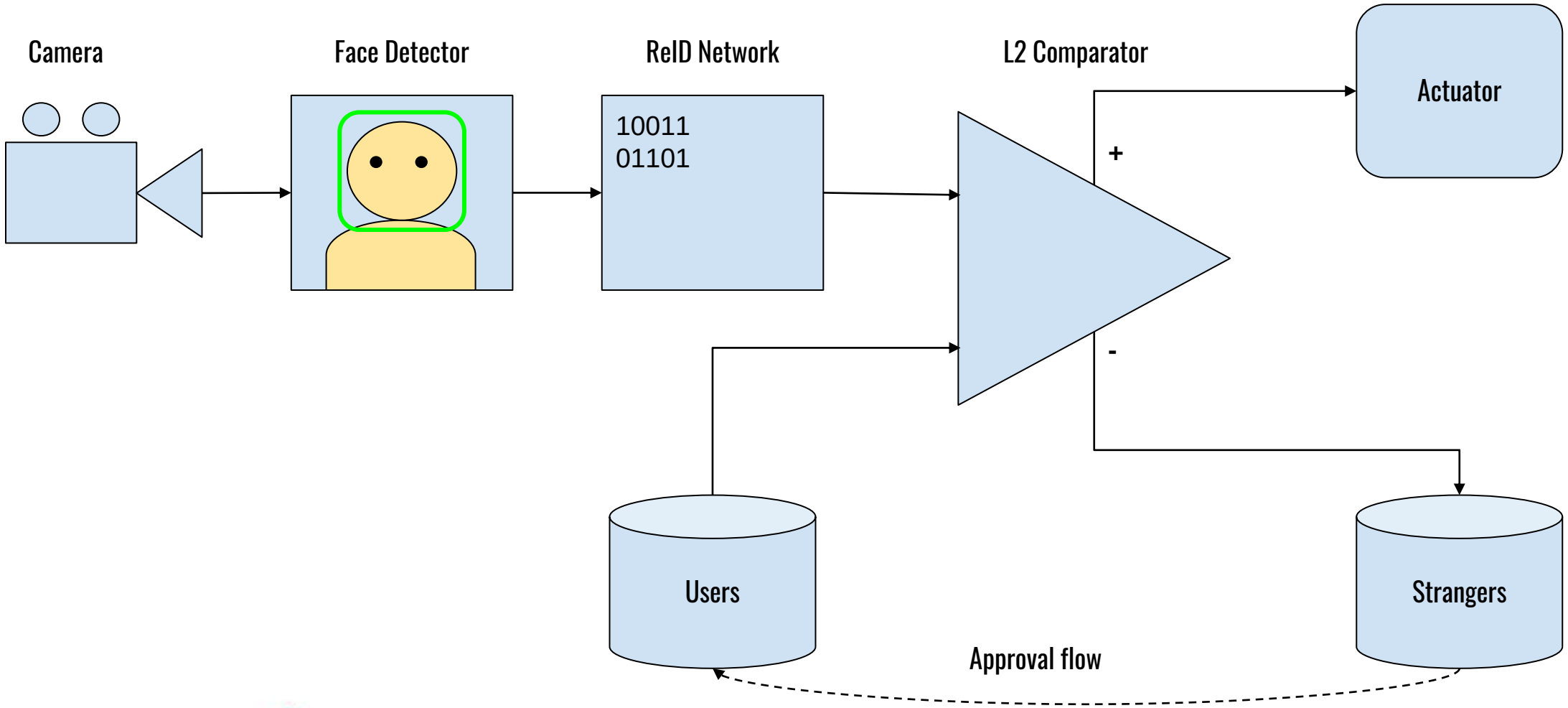
xperience/ai

Задача: Умный дверной звонок

- Захват с камеры и отображение потока на экране (QVGA)
- Детектирование лица
- Идентификация человека по известным дескрипторам (10-20 человек)
- Управление списком пользователей

xperience/ai

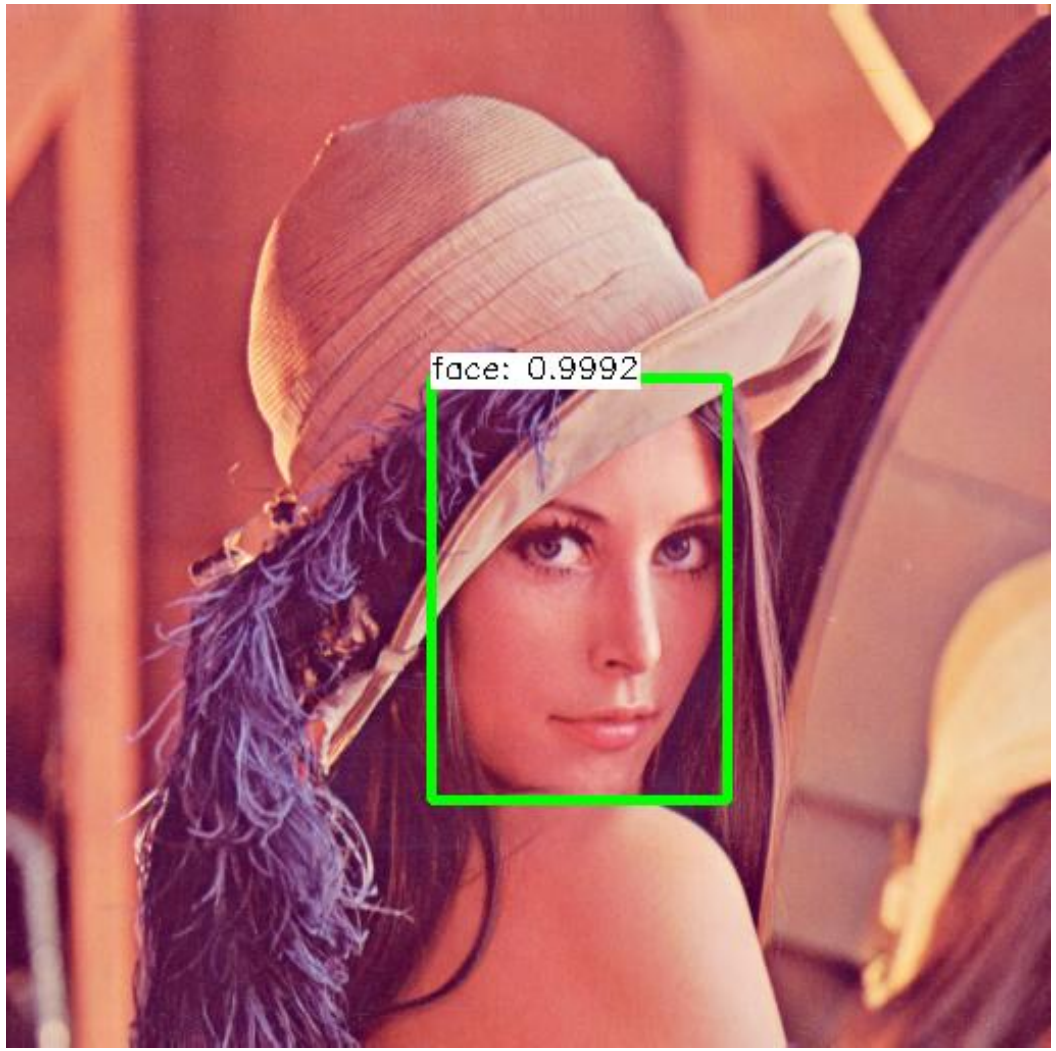
Декомпозиция



xperience/ai

Декомпозиция: детектор

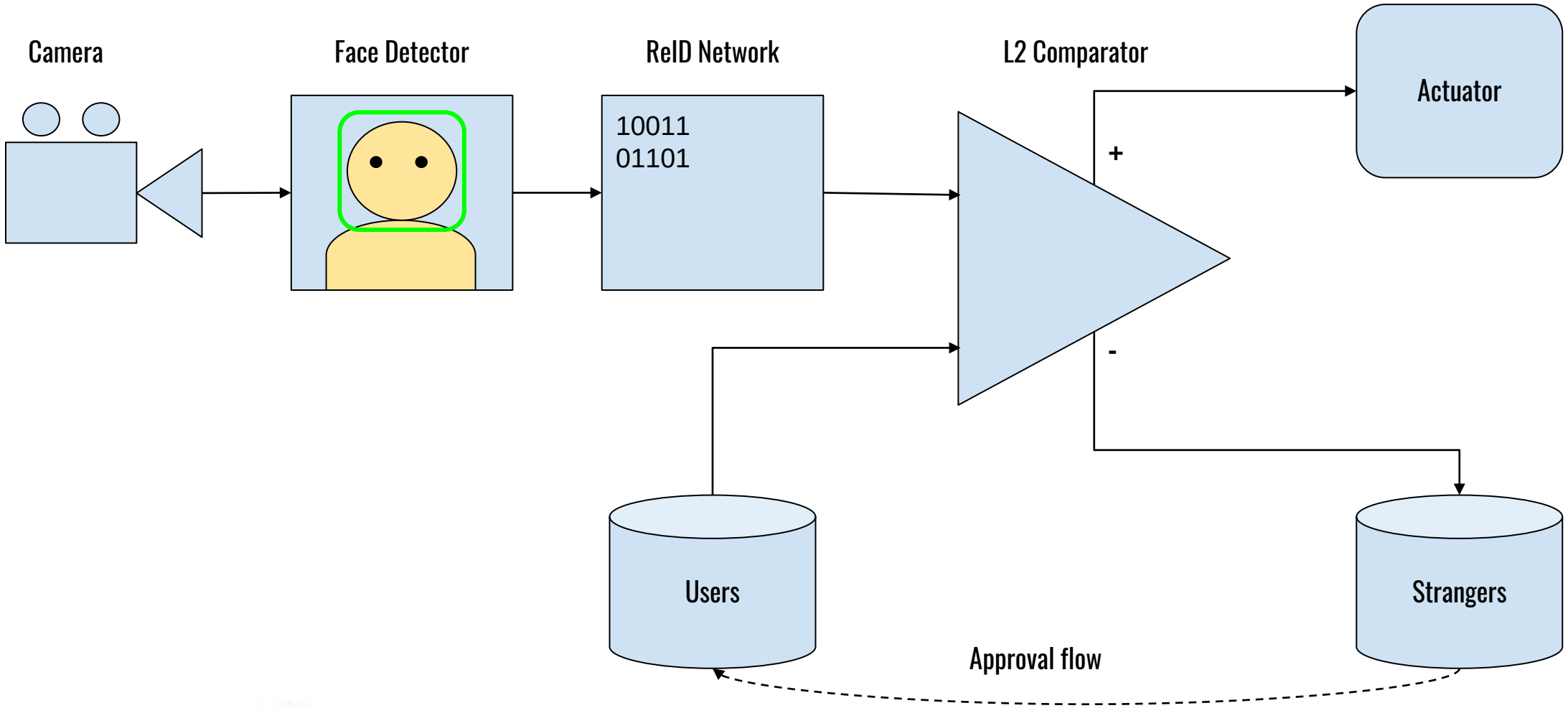
- Координаты прямоугольников с лицами
- Отклик для каждого лица, аналог вероятности
- Опционально: ключевые точки, например глаза



Декомпозиция: дескриптор & ReID

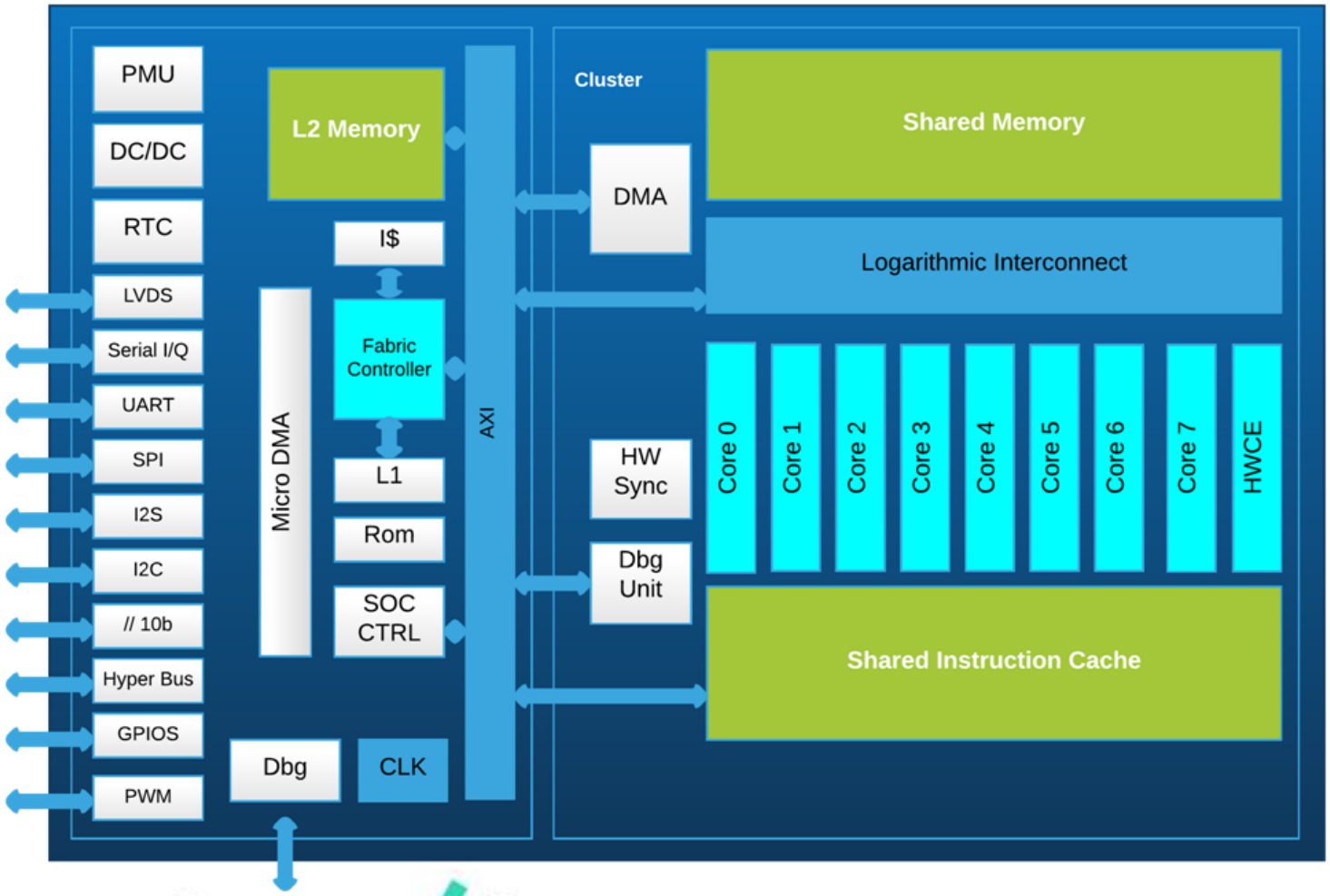
- Дескриптор — вектор признаков описывающих характерные черты изображения
- Дескрипторы должны образовывать линейное пространство
- Дескрипторы должны быть близки по введённой метрике для изображений одного человека
- Дескрипторы должны быть максимально далеки для изображений разных людей
- Идентификация — поиск дескриптора с наименьшим расстоянием до искомого в базе с известными людьми

Декомпозиция



xperience/ai

GAP8: Введение



- GAP8 - RISC-V MCU
- L2 512 KiB
- Cluster L1 64 KiB
- FC L1 8KiB
- HWCE
- 1D/2D cluster-DMA
- FC micro-DMA
- HyperBUS для FLASH и RAM (L3)

xperience/ai

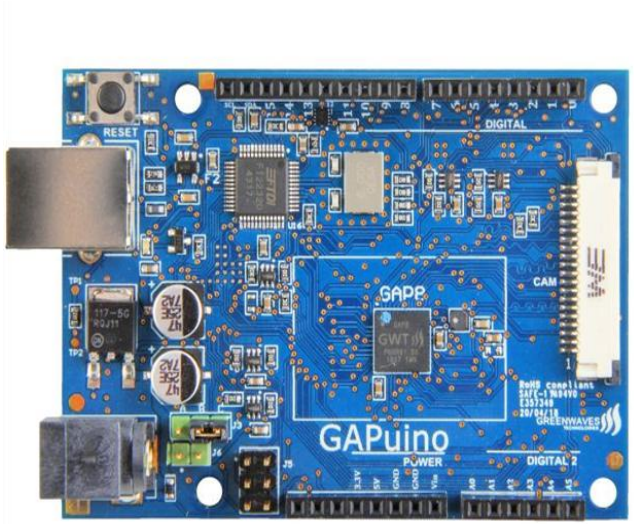
https://greenwaves-technologies.com/manuals/BUILD/HOME/html/index.html

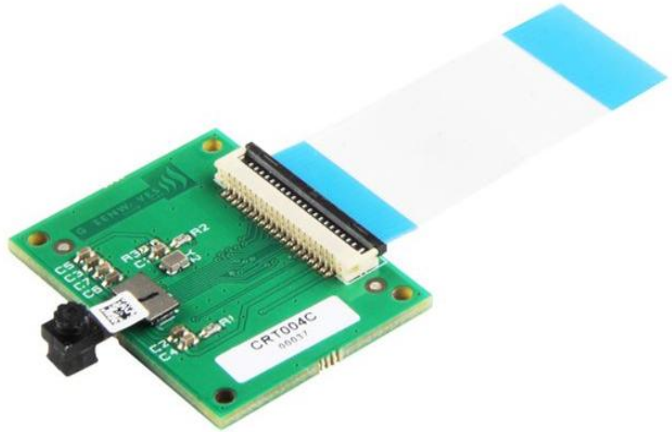
GAP8: SDK & Autotiler

- Поддержка PulpOS, FreeRTOS, MBed OS
- PMSIS — универсальный API для GPIO и периферии
- API для управления вычислительным кластером
- Autotiler — C-фреймвок для генерации кода обработки данных на кластере по тайлам
- API для управления HWCE
- Примеры и демо для вычислительных задач

Что есть

- Плата GAPuino + QVGA-камера + ЖК-дисплей
- API для управления камерой и дисплеем
- Готовая реализация детектора лиц на каскадах Хаара
- CNN примитивы в Autotiler
- Пример для MNIST, CIFAR, некоторые примитивы для CV





xperience/ai

План

- Выбор сети, тренировка, квантизация
- Запуск сети на GAP “в вакууме”
- Оптимизация выполнения сети
- Интеграция сети в приложение
- Борьба за память
- Документация, тестирование, публикация

Выбор сети

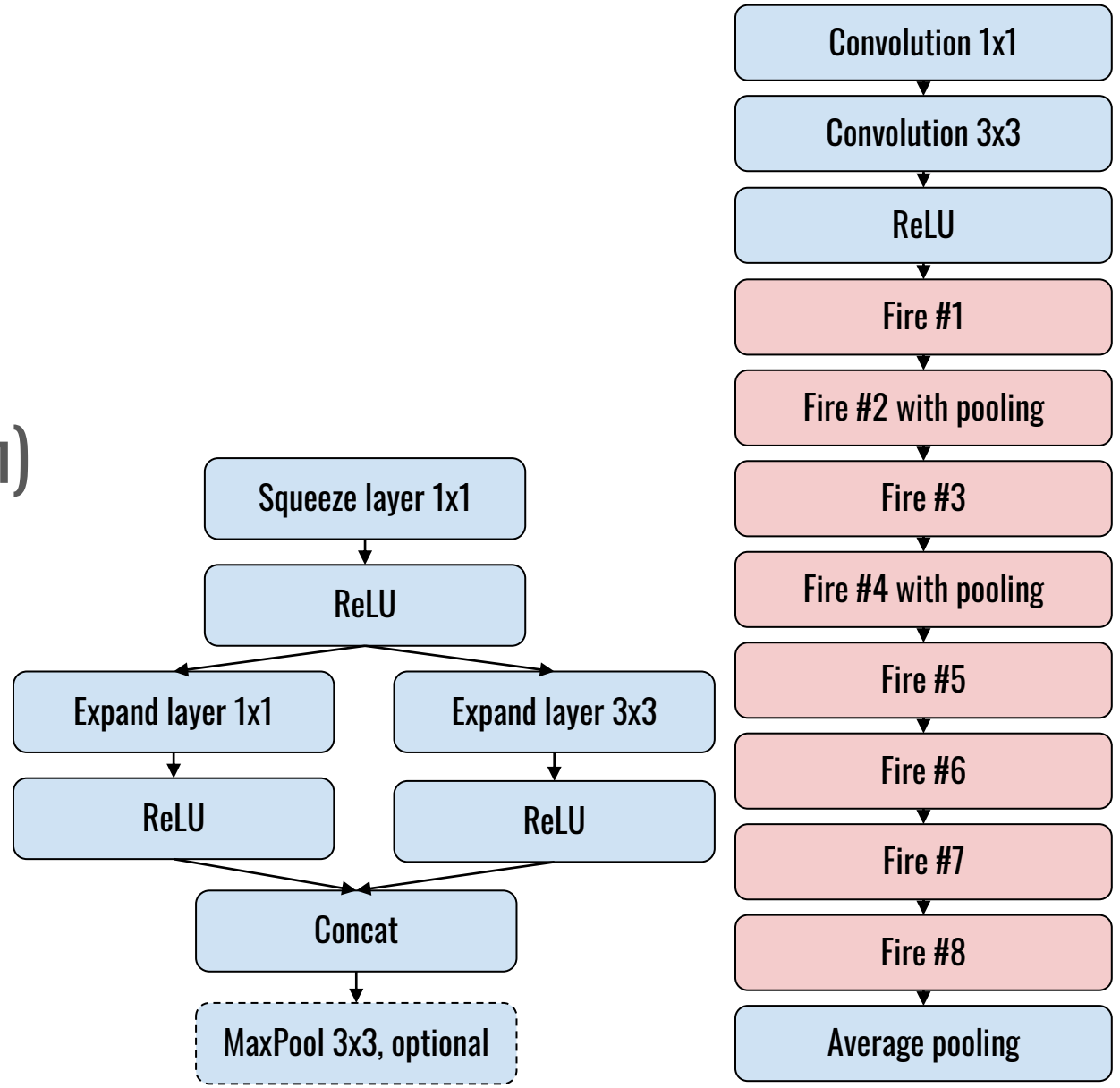
- Минимальное количество примитивов
- Оптимизация памяти в run-time
- Нет реализации depthwise-свёрток
- Без residual/skip connection

Архитектура	**Миллионов параметров**
SqueezeNet 1.1	0.722
SqueezeNet 1.0	0.735
ShuffleNet	0.905
ShuffleNet v2	1.254
MobileNet v1	2.224
MobileNet v2	3.207

xperience/ai

Адаптация Сети

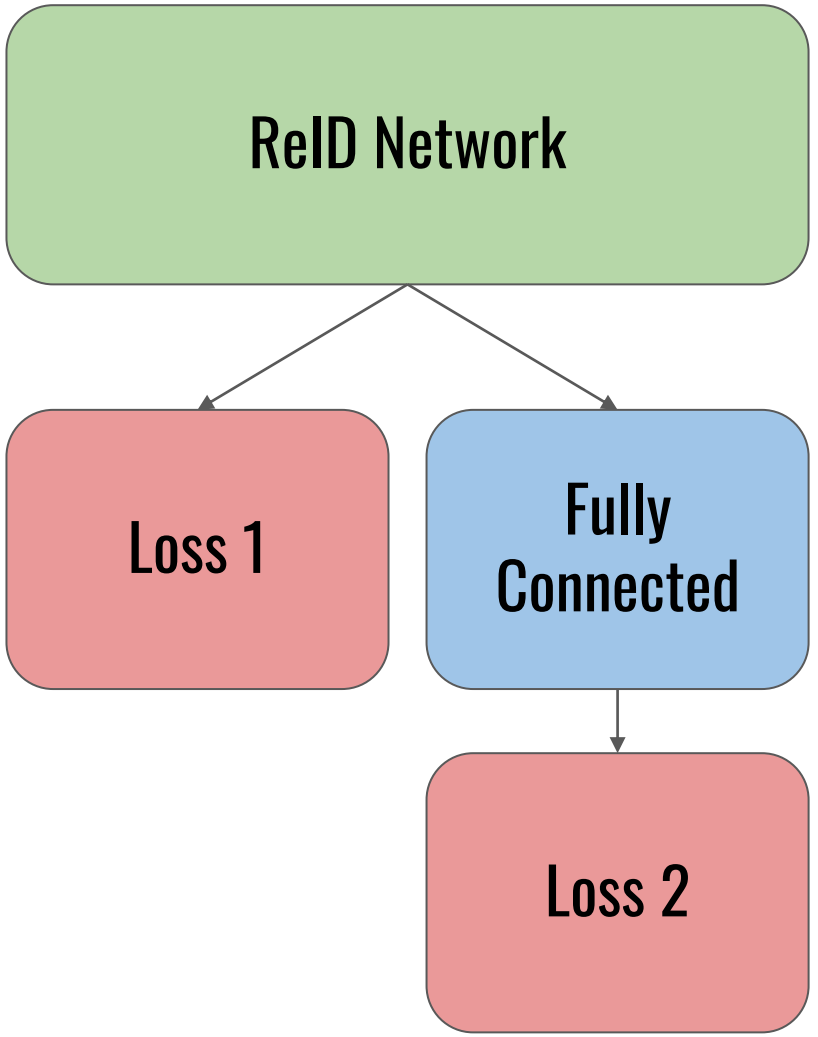
- Вход 128x128
- Черно-белое изображение (1 канал)
- Размер дескриптора 512
- Оптимизация памяти в run-time
- Только целочисленные (Fixed Point) операции



xperience/ai

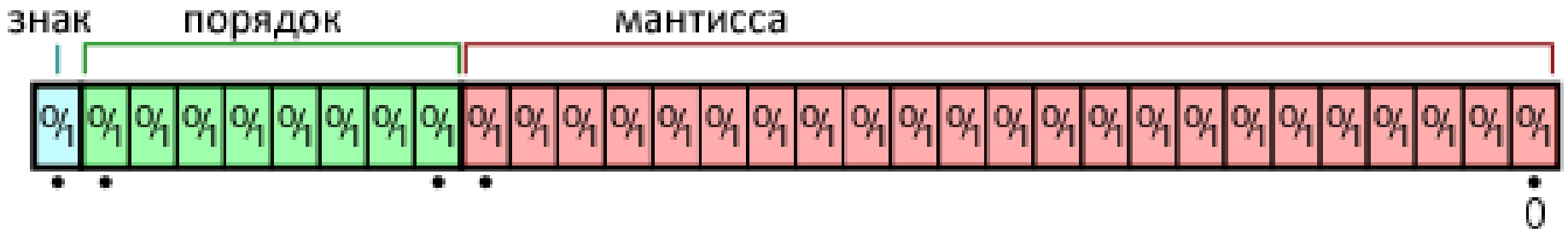
Тренировка Сети

- Задача классификации
- Dataset: VGGFace2
- Loss: Cross Entropy + Lifted Structured Loss
- Нет выравнивания лица
- Нормализация входа склеена с первой свёрткой



xperience/ai

Квантизация: Float Point vs Fixed Point

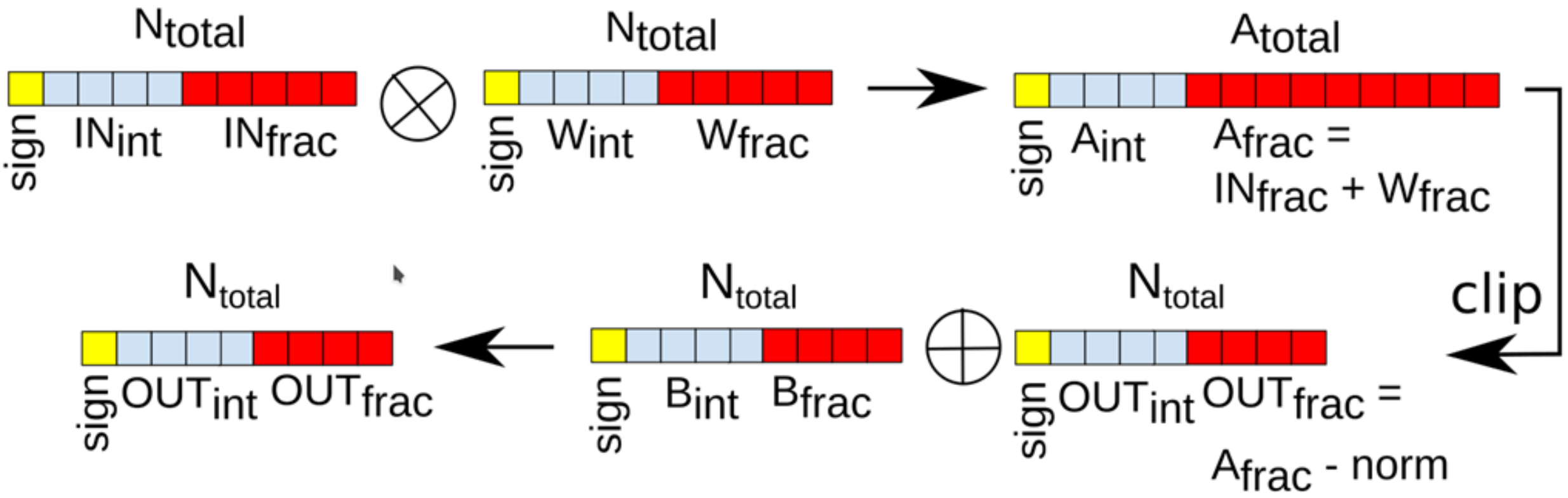


$(-1)^s \times M \times B^E$, где s — знак, B-основание, E — порядок, а M — мантисса

При E == const, порядок можно не хранить, мантисса и знак м.б. представлены как целое число со знаком (int)

- Сложение и вычитание — как для целого со знаком
- Умножение — как для целого со знаком, разрядность дробной части удваивается

Fixed Point & Квантизация



Квантизация: Float Point -> Fixed Point

1. Прогоняем через сеть набор тестовых данных
2. Строим распределение весов и активаций
3. Определяем максимальное по модулю значение веса и активации для каждого слоя
4. Количество бит для целой части: $\log_2(\max(|X|))$, остальное — дробная часть
5. Одинаковая квантизация перед Concat

Квантизация: PyTorch

- Квантизация в 16 бит
- Количество бит для целой части: $\log_2(\max(|X|))$
- Clip после каждого умножения
- GAP квантизация в PyTorch: https://github.com/xperience-ai/gap_quantization/

xperience/ai

Проблемы с сетью на GAP

- Веса (1.37 MiB) не умещаются в L2
- Вход, выход, веса для слоя должны быть в L2
- Вход, выход, веса для следующего слоя — тоже в L2
- Выход некоторых слоёв до Pooling не укладываются в L2. Нужно объединение слоёв
- Массивы должны быть физически непрерывны

xperience/ai

Укладка памяти: свёртки

Layer input

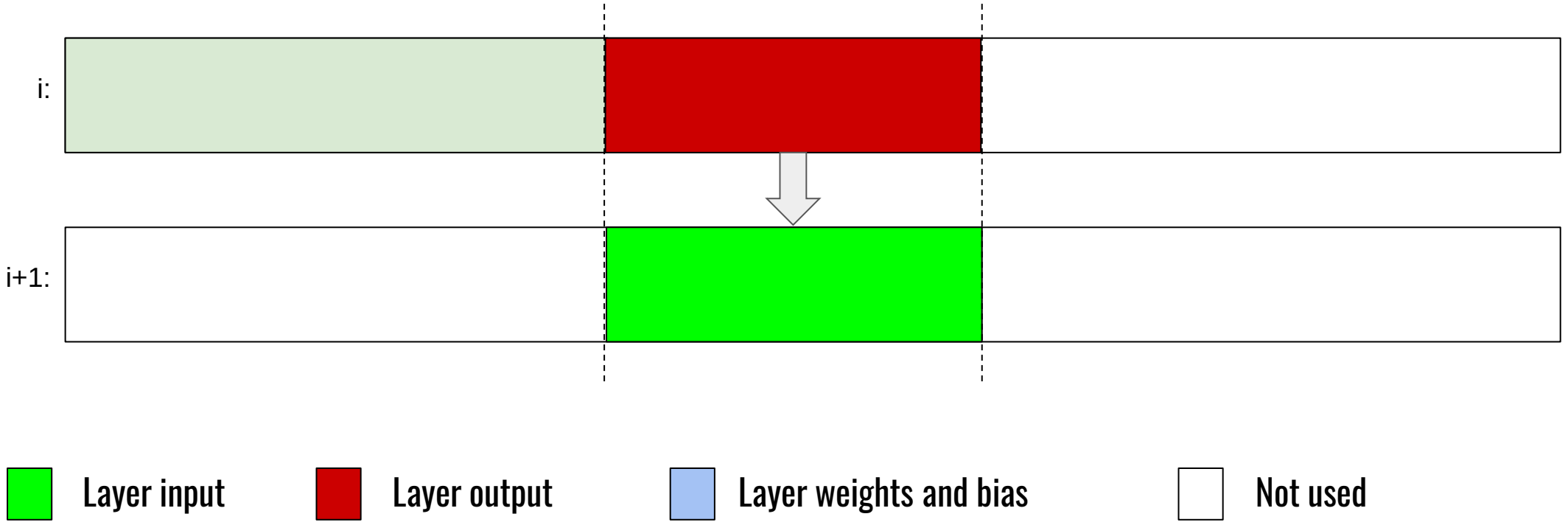
Layer output

Layer weights and bias

Not used

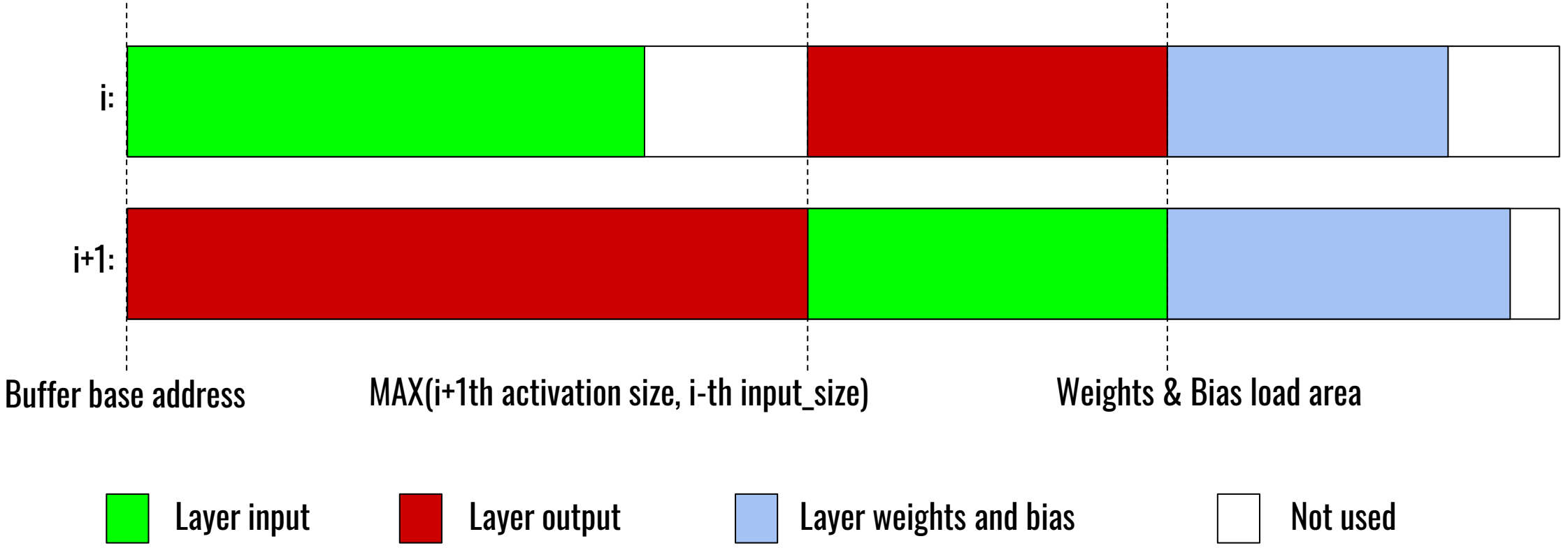
xperience/ai

Укладка памяти: свёртки



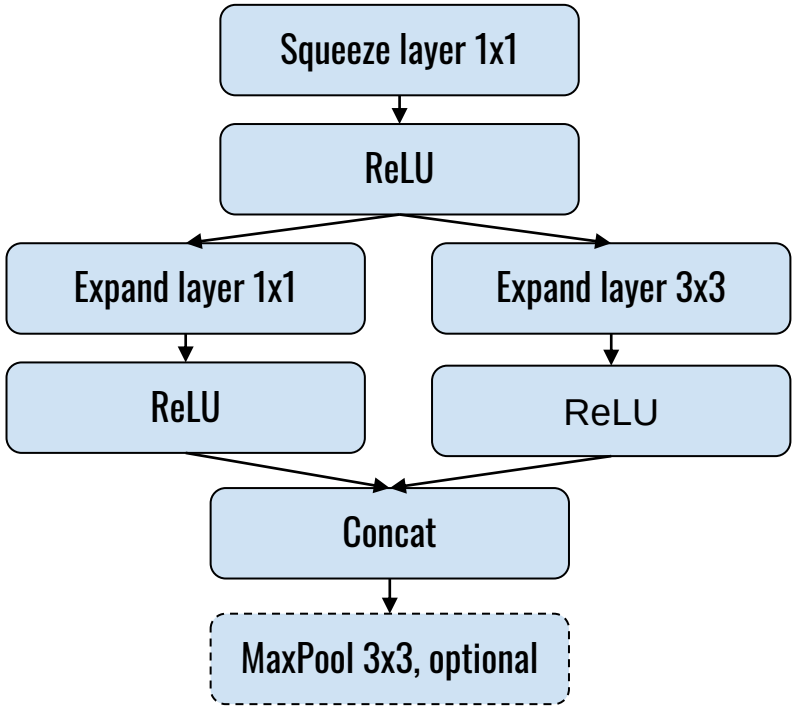
xperience/ai

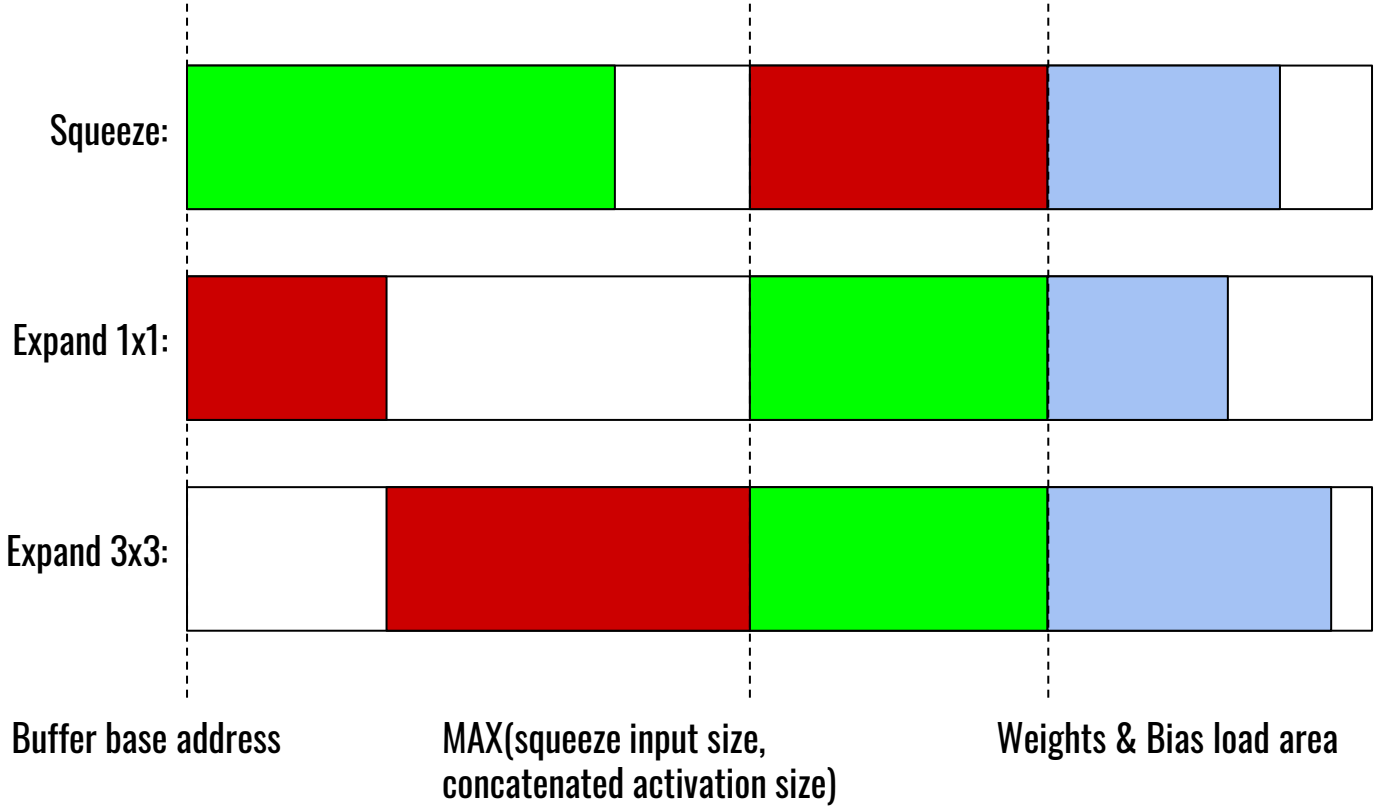
Укладка памяти: свёртки



xperience/ai

Укладка памяти: Fire Module





xperience/ai

Demo: region `L2' overflowed by XXX bytes

- Сеть 344 KiB
- Каскадный детектор 54KiB + промежуточные буферы
- Кадр для отрисовки 75KiB
- Входной сети регион (ROI) 32KiB
- Дескрипторы известных пользователей N*1KiB
- *.text*

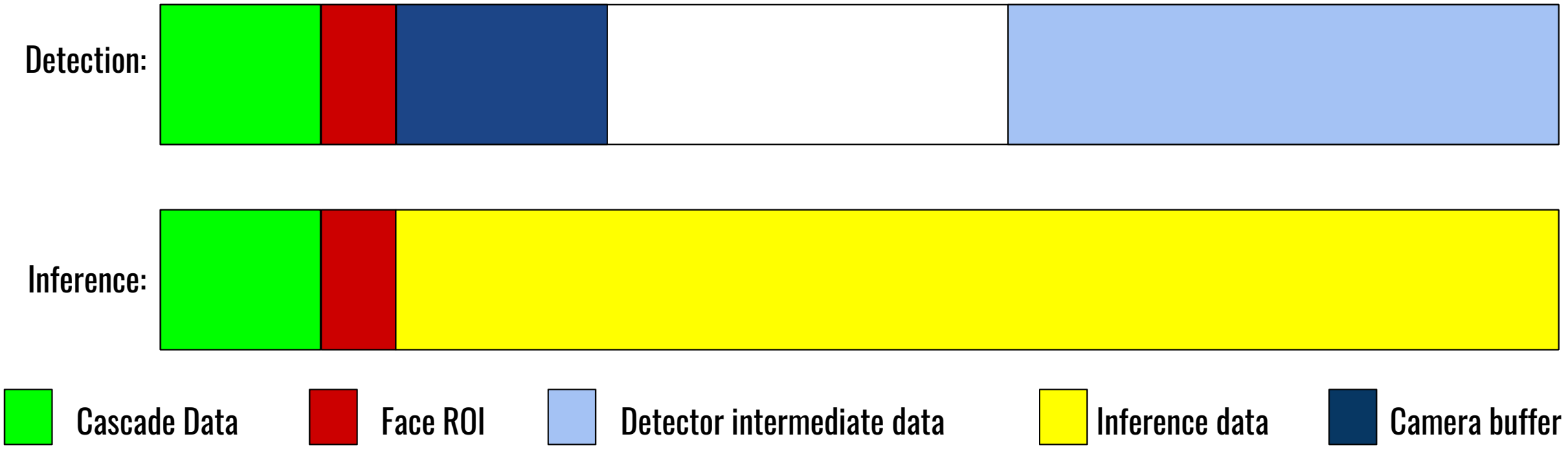
$ riscv32-unknown-elf-nm -S --size-sort ./test

xperience/ai

Оптимизация памяти

- Статическая схема распределения памяти
- Минимум глобальных переменных, переиспользование буферов
- Фон-Неймановская архитектура => *-Os* и другие опции
- Стек FC в L1. Всё, что возможно — в стек!

Переиспользование памяти



xperience/ai

Результат в числах

- Качество на LFW: **0.9638**
- Падение качества после квантизации (ПК): **0.0005**
- Падение качества на GAP: **0.0023**
- Время выполнения сети:
  - С синхронной загрузкой весов **780 мс**
  - С асинхронной загрузкой весов **380 мс**
- Пиковое потребление L2 памяти для сети **344 KiB**

xperience/ai

Результат

- ICCV 2019 submission: “512 KiB RAM is enough! Live camera face re-identification DNN on MCU”
- GAP квантизация в PyTorch: https://github.com/xperience-ai/gap_quantization/
- ReID Demo -> Gap SDK на Github

Что дальше?

Оптимизация времени сети и потребления энергии:

- Квантизация в 8 бит
- HWCE
- общий профиль и оптимизация

Новые сети:

- Autotiler + внешняя память
- Детектор SSD, YOLO, etc

xperience/ai

Ссылки

- 512KiB RAM Is Enough! Live Camera Face Recognition DNN on MCU: http://openaccess.thecvf.com/content_ICCVW_2019/html/LPCV/Zemlyanikin_512KiB_RAM_Is_Enough_Live_Camera_Face_Recognition_DNN_on_ICCVW_2019_paper.html
- GAP Quantization (PyTorch): https://github.com/xperience-ai/gap_quantization/
- GAPuino User Manual: https://gwt-website-files.s3.amazonaws.com/gapuino_um.pdf
- GAP8 Hardware Reference Manual: https://gwt-website-files.s3.amazonaws.com/gap8_datasheet.pdf
- GAP SDK: https://greenwaves-technologies.com/manuals/BUILD/HOME/html/index.html

xperience/ai

Вопросы? Предложения!

xperience/ai