Реализация сервиса работы с медиа-файлами на лету на основе FPGA Intel

Аким Волков

Заместитель директора по эксплуатации ВК

ВК в цифрах

97 млн. активных пользователей в месяц

10 млрд. сообщений в день

1 млрд. лайков в день.

9 млрд. просмотров постов в день

650 млн. просмотров видео в день

89 языков

ВК в цифрах

19 тысяч серверов

3 ЦОД объединеных в единую сеть

30 узлов CDN и точек присутствия по миру

1.1 экзабайта данных пользователей

3.5 Терабита/с скачиваемых пользователями данных

Как нам успешно вести бизнес?

Оптимизация всех процессов – ключ к успеху.

Мы оптимизируем ИТ для улучшения ТСО на всех уровнях, во всех наших ЦОД.

И сегодня: еще одна история про работу с данными от ВК.
И немного про FPGA. ☺

Итак, мы разделили данные. Что можно еще улучшить?

Задача: Разделить максимально эффективно данные на несколько уровней: Горячий, теплый, холодный для улучшения TCO и увеличения производительности каждого слоя

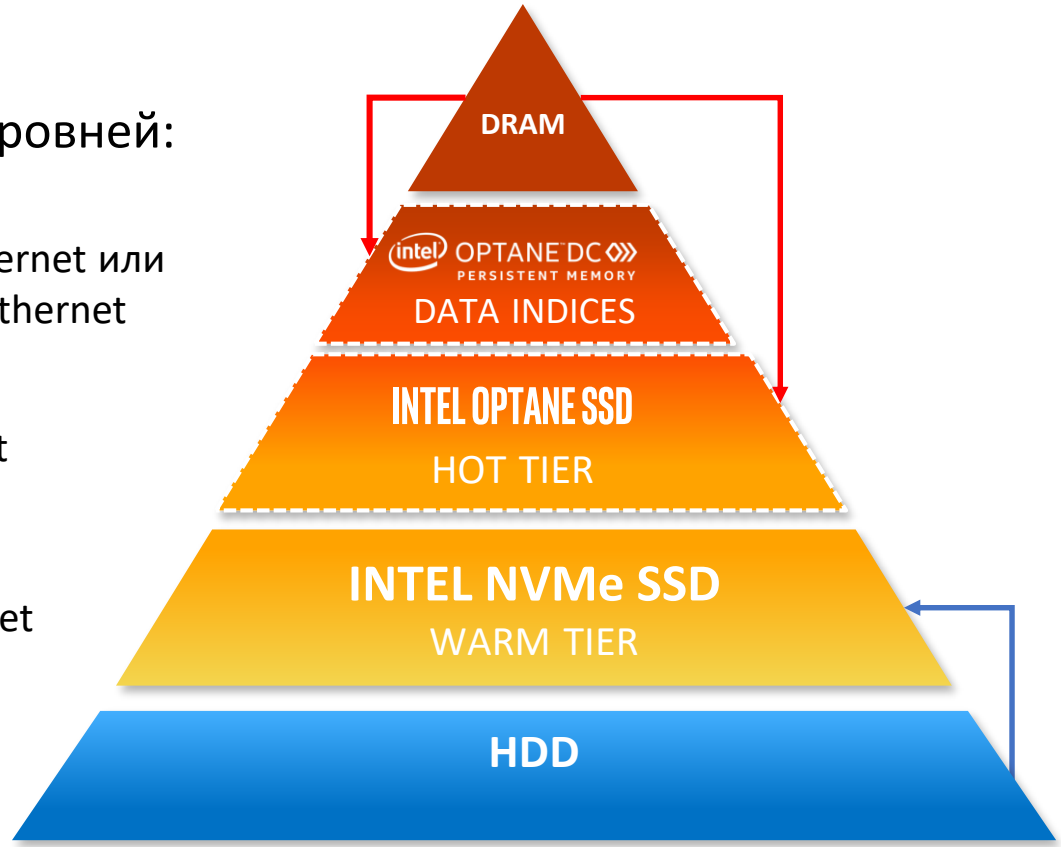
Результат: Комплексное решение включает в себя сервера уровней:

Горячий:
2xIntel Xeon 6230/8x**128GB Intel Optane DC Memory**/10Gbs Ethernet или
2xIntel Xeon 6230/8x**128GB Intel Optane DC Memory**/2x25Gbs Ethernet

Горячий:
2xIntel Xeon 6230/**Intel Optane P4800X 750GB**/2x25Gbs Ethernet

Теплый:
2xIntel Xeon 6230/6x**Intel NVMe SSD P4320 8TB**/4x25Gbs Ethernet

Холодный:
2xIntel Xeon 6230/100x14TB HDD/10Gbs Ethernet



Состояние: Улучшили TCO на 65-90% в различных слоях и сэкономили миллионы долларов в год.

Куда двигаться дальше?

Даже поделенный на слои данных **1.1 экзабайт** стоит недешево. Для дальнейшей оптимизации распределенной системы хранения нужно более точное знание **о типах данных**.

Цель – оптимизация хранения изображений. У нас их несколько **сотен Петабайт**. При этом мы часто храним несколько копий в различных разрешениях и форматах для выдачи на различные типы устройств пользователя.

А если их привести к единому формату и конвертировать “на лету”?

Это может дать нам **заметное** сокращение объема данных.

Преобразование изображений на лету

ЗА:

Храним одну копию данных вместо n копий:

- сокращение издержек на систему хранения (сервера, диски, стойки, сеть, электричество)
- Простота расширения списка поддерживаемых к конвертации форматов
- Гарантия консистентности данных
- Меньшая нагрузка на систему при перемещении между слоями данных

ПРОТИВ:

- Мы должны заплатить за это работой серверов по преобразованию данных в различные форматы
- Необходимо купить эти серверы и обеспечить их работоспособность в ЦОД (сеть, стойки, энергия).

Как поступить?

Варианты решения задачи

Варианты реализации преобразования:

- С помощью серверов на стандартных x86 процессорах
- Серверы с ускорителями GP GPU nVidia T4
- Серверы с ускорителями Intel FPGA Arria 10

По результатам проведенных пилотов стало понятно: наиболее перспективным вариантом является сервер x86 с ускорителями (чем больше тем лучше) на борту.

JPEG to JPEG преобразование

```
static $file_sizes = [
  's' => ['width' => 75, 'height' => 75],
  'm' => ['width' => 130, 'height' => 130],
  'x' => ['width' => 604, 'height' => 604],
  'y' => ['width' => 807, 'height' => 807],
  'z' => ['width' => 1280, 'height' => 1080],
  'w' => ['width' => 2560, 'height' => 2160],
];
static $crop_sizes = [
  'o' => ['width' => 130, 'min_height' => 87, 'max_height' => 390],
  'p' => ['width' => 200, 'min_height' => 133, 'max_height' => 600],
  'q' => ['width' => 320, 'min_height' => 213, 'max_height' => 900],
  'r' => ['width' => 510, 'min_height' => 340, 'max_height' => 900],
];
```

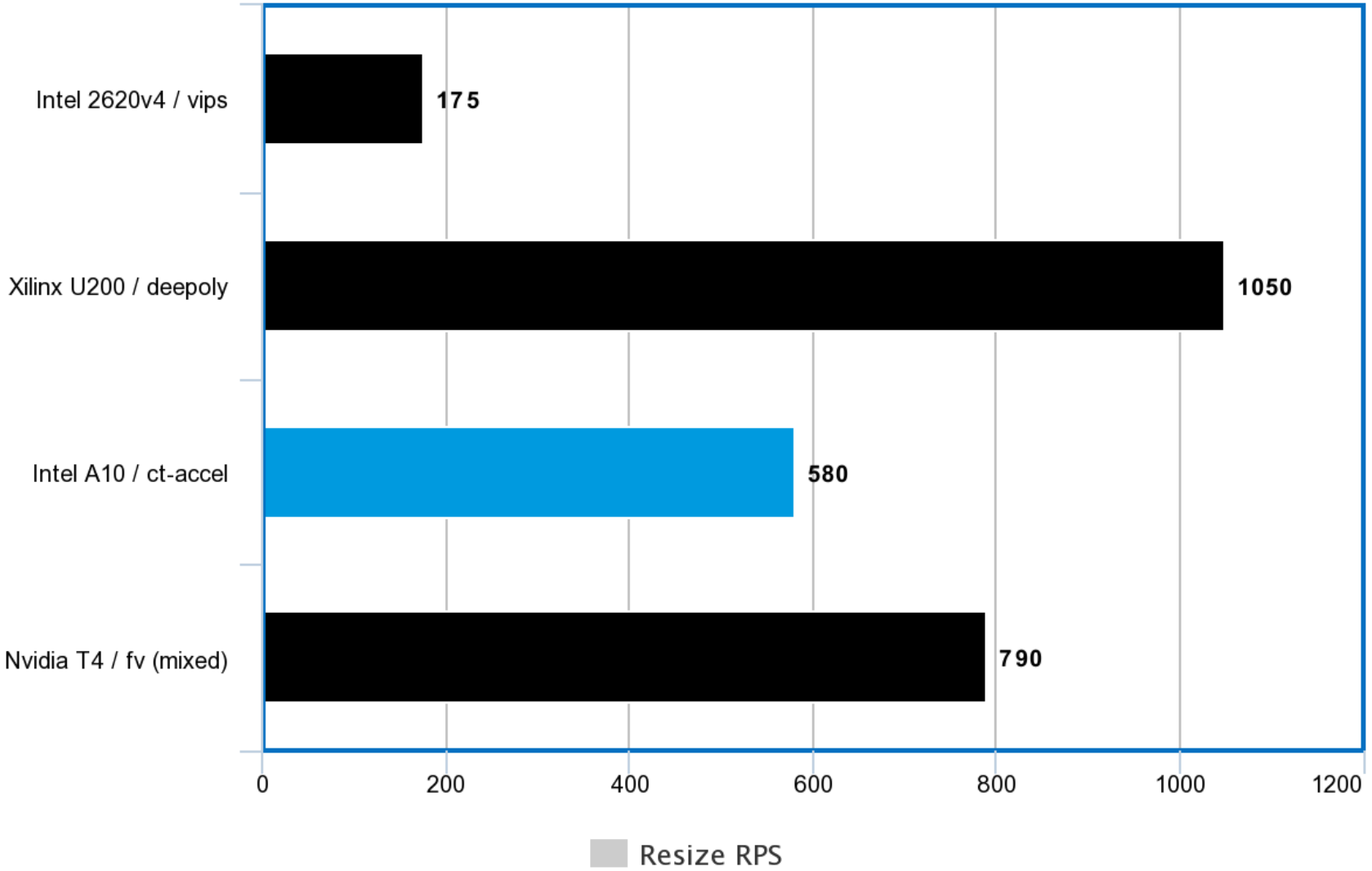
Исходные данные:

- фиксированная сетка размеров
- >3M RPS к кеширующим фронтам
- ~500K RPS требуемая производительность фермы ресайза

Test Environment:
CPU: 2*Intel® Xeon® CPU E5-2620v4
RAM: 128GB
OS: Debian Jessie / Debian Stretch

JPEG to JPEG : Бенчмарк (начало)

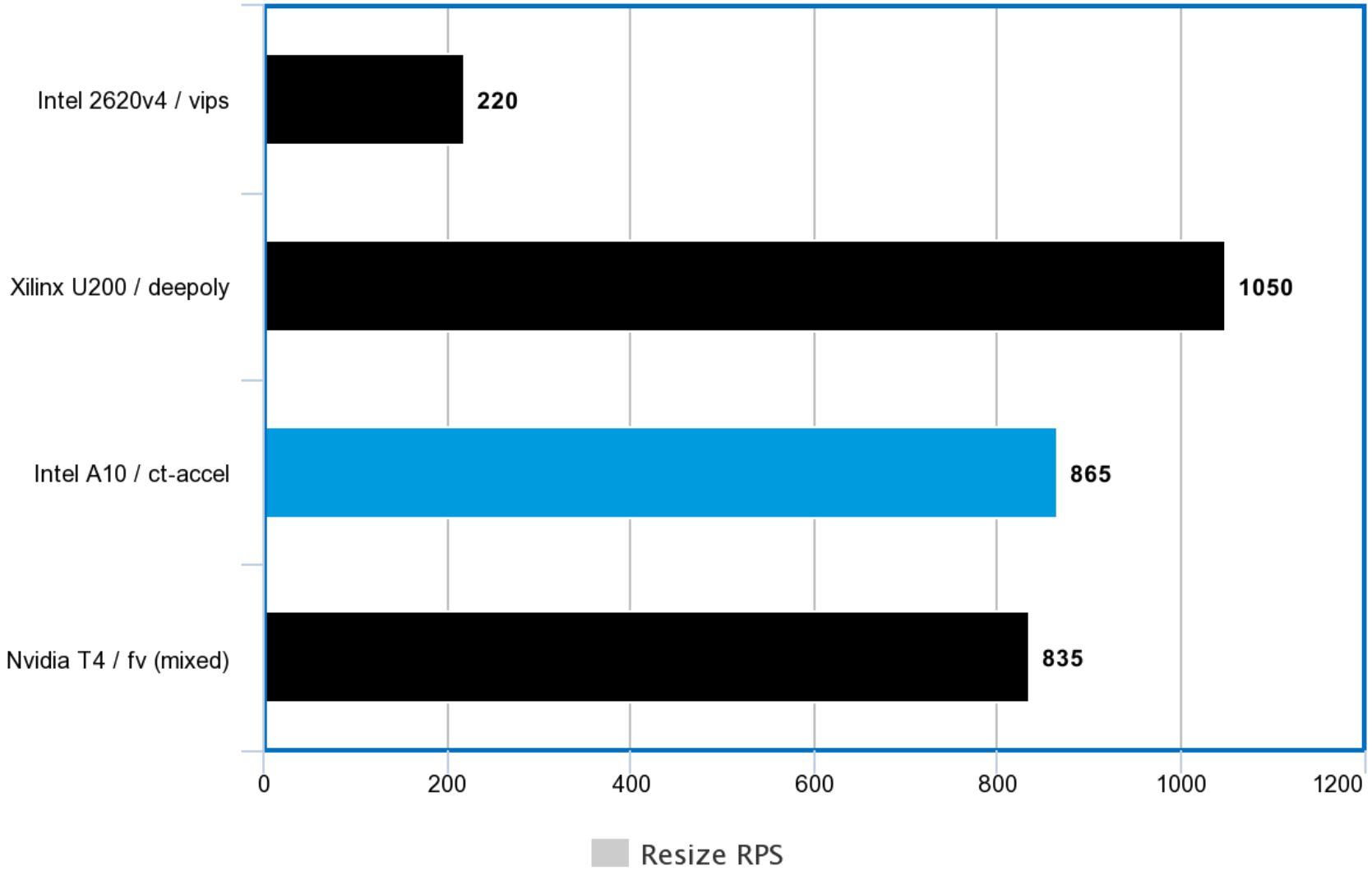
Workload:
12% 2048x1536 -> ...x510
27% 2048x1536 -> ...x320
6% 2048x1536 -> ...x807
23% 2048x1536 -> ...x200
1% 2048x1536 -> ...x1280
21% 2048x1536 -> ...x130
10% 2048x1536 -> ...x604



Test Environment:
CPU: 2*Intel® Xeon® CPU E5-2660v4
RAM: 128GB
OS: Debian Jessie / Debian Stretch

JPEG to JPEG : Бенчмарк (тюнинг)

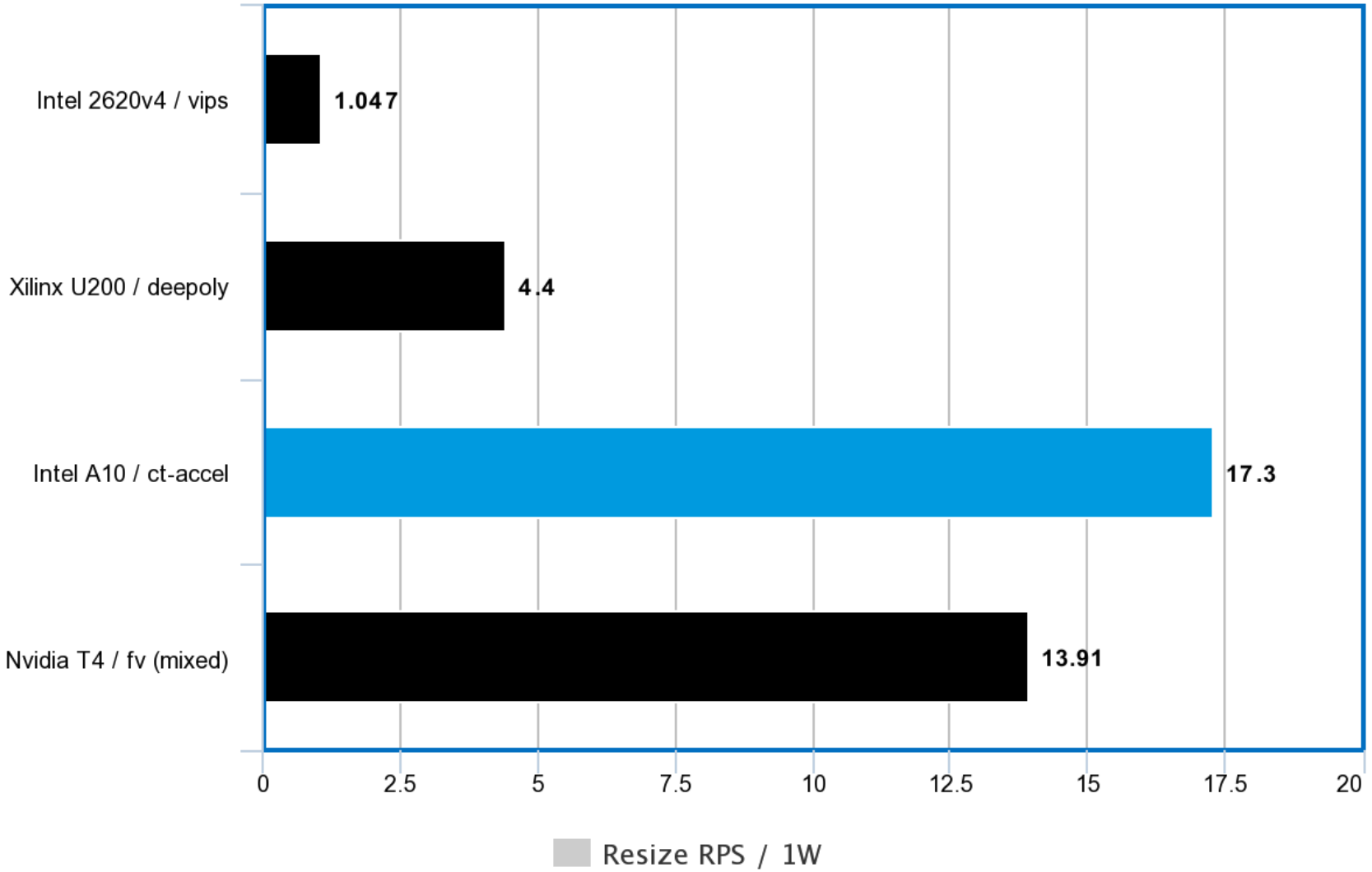
Workload:
12% 2048x1536 -> ...x510
27% 2048x1536 -> ...x320
6% 2048x1536 -> ...x807
23% 2048x1536 -> ...x200
1% 2048x1536 -> ...x1280
21% 2048x1536 -> ...x130
10% 2048x1536 -> ...x604



Test Environment:
CPU: 2*Intel® Xeon® CPU E5-2660v4
RAM: 128GB
OS: Debian Jessie / Debian Stretch

JPEG to JPEG : Бенчмарк (RPS/1W)

Workload:
12% 2048x1536 -> ...x510
27% 2048x1536 -> ...x320
6% 2048x1536 -> ...x807
23% 2048x1536 -> ...x200
1% 2048x1536 -> ...x1280
21% 2048x1536 -> ...x130
10% 2048x1536 -> ...x604



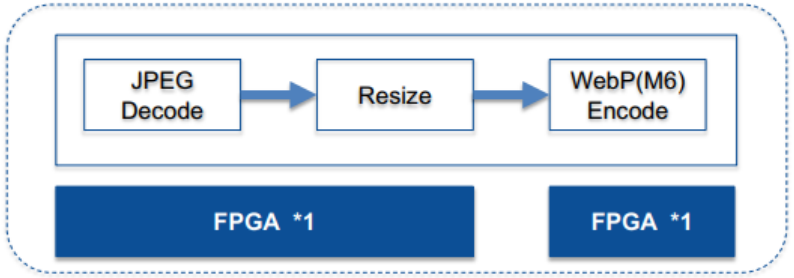
Test Environment:
CPU: 2*Intel® Xeon® CPU E5-2660v4
RAM: 128GB
OS: Debian Jessie / Debian Stretch

Результаты пилотных проектов

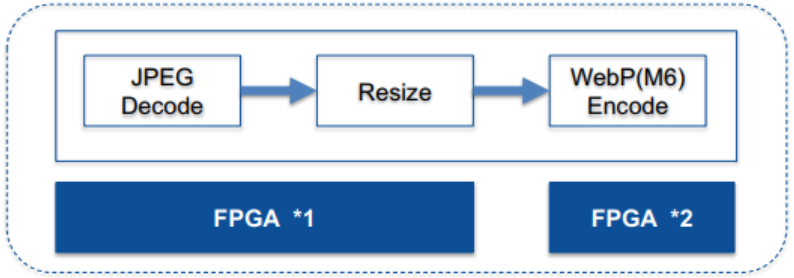
- Производительность сервера с ускорителями масштабируется линейно с ростом числа ускорителей в сервере не зависимо от мощности центрального процессора.
- Производительность ускорителя Intel FPGA Arria 10 GX не хуже чем nVidia T4 на преобразовании Jpeg->Jpeg.
- Помимо Jpeg->Jpeg ускоритель Intel FPGA Arria 10 GX умеет производить преобразования Jpeg<->Heif, Jpeg<->Webp.

JPEG to WebP преобразование

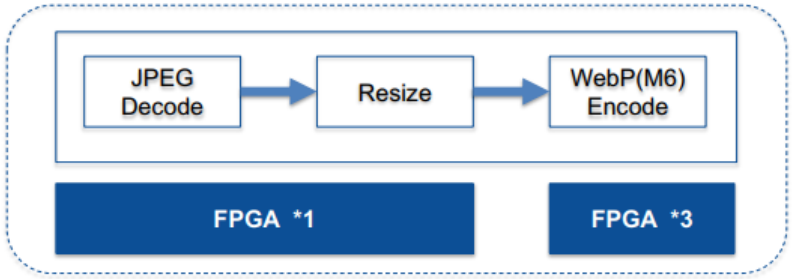
Configuration-1:FPGA (1+1)



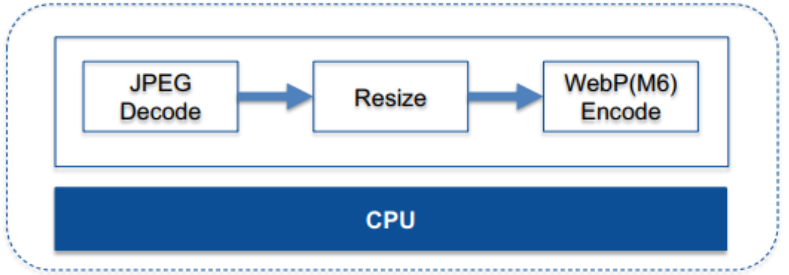
Configuration-2:FPGA (1+2)



Configuration-3:FPGA (1+3)



Configuration-4:CPU

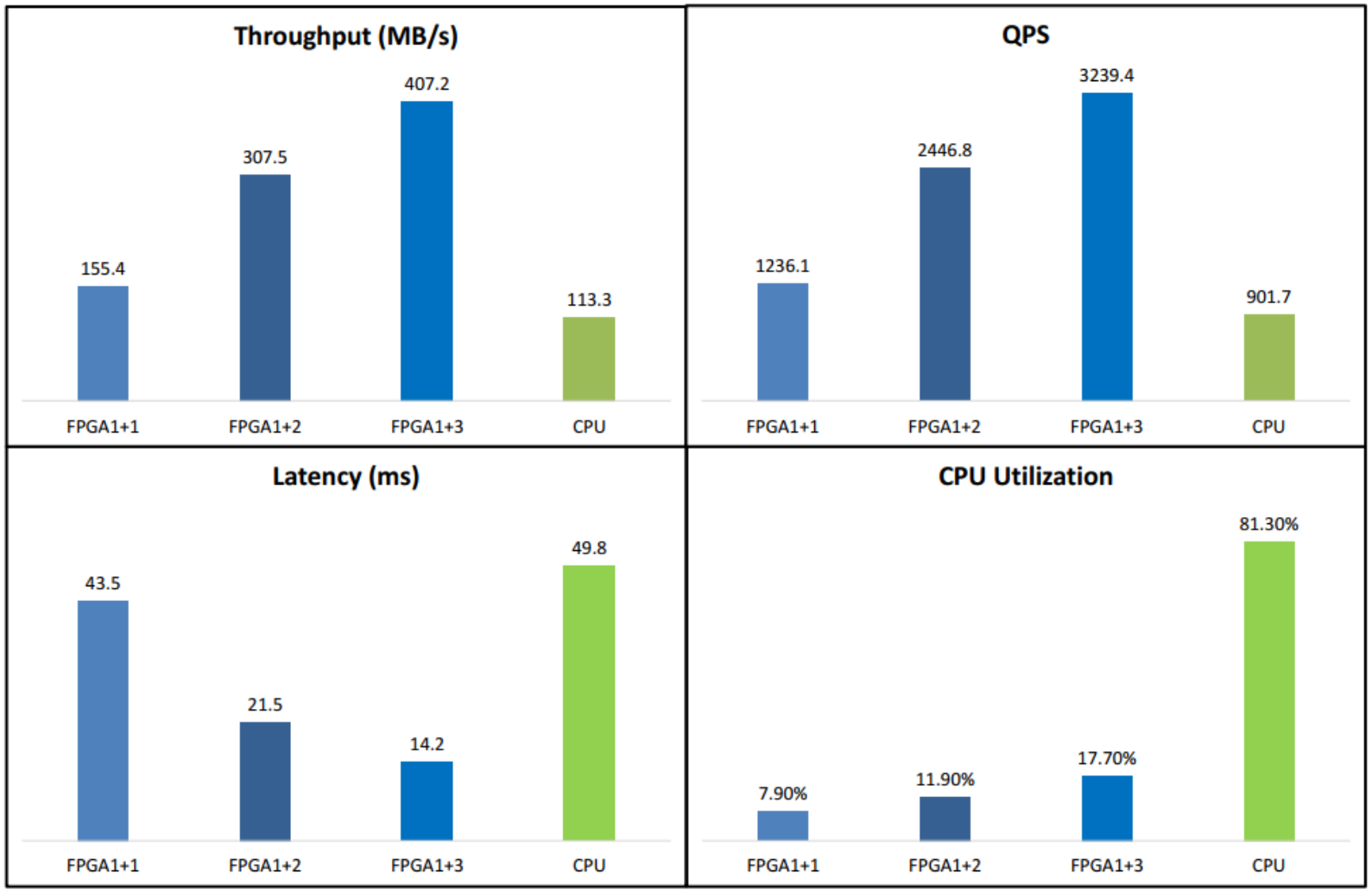


JPEG to WebP маленькие картинки

- **QPS:**
  - FPGA1+1 is 1.4 times that of CPU
  - FPGA1+2 is 2.7 times that of CPU
  - FPGA1+3 is 3.6 times that of CPU
- **Latency:**
  - FPGA1+1 is reduced to 87% of CPU
  - FPGA1+2 is reduced to 43% of CPU
  - FPGA1+3 is reduced to 29% of CPU

Input:
- Average file size=130K
- Resolution: 1024x768
- Total file number=10000
- Total files size=1256.97MB
- Format=JPEG

Output:
- Resolution:240x180
- Format=WebP(M6)



JPEG to Heif преобразование

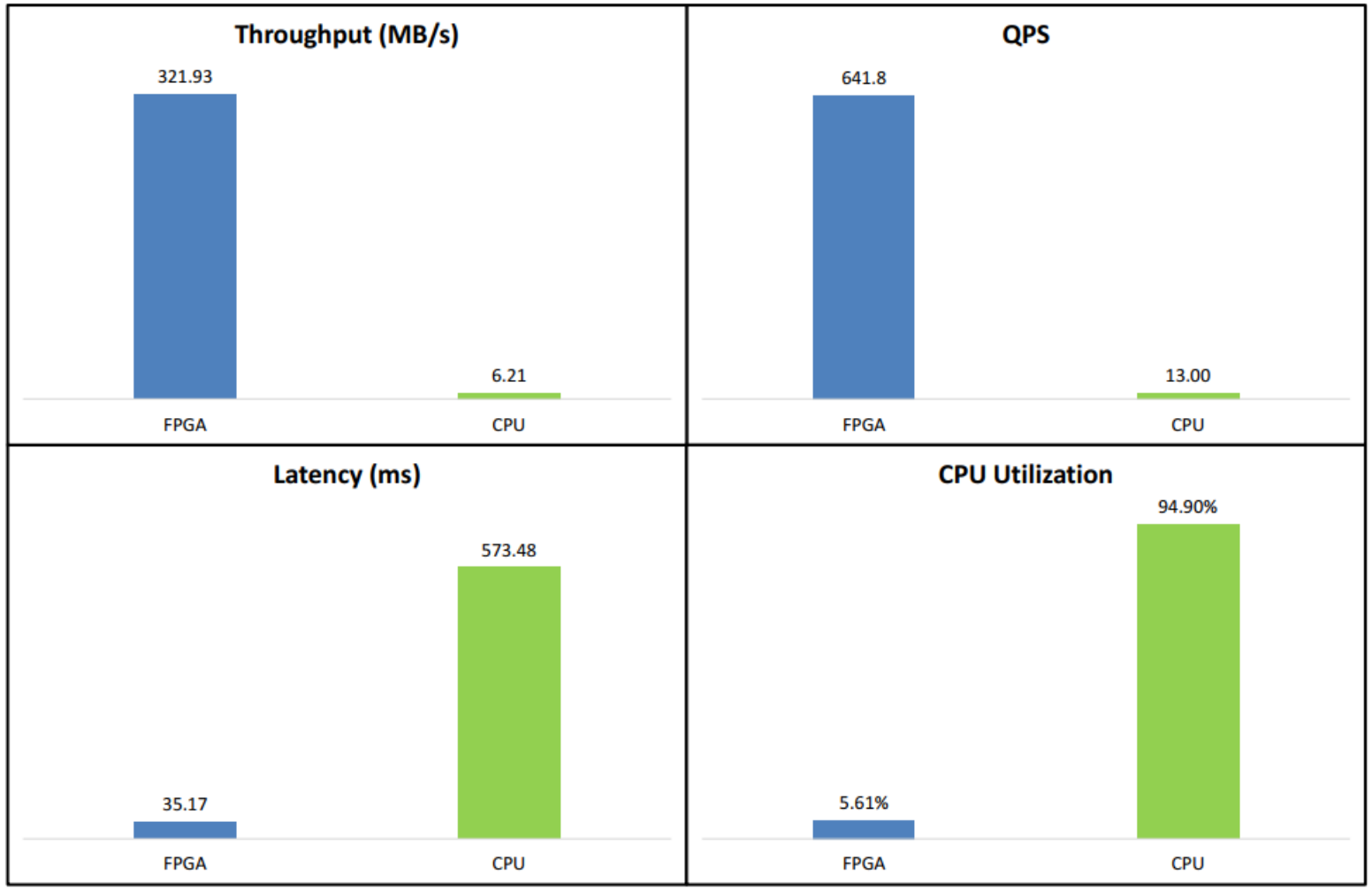
- **QPS :**
  - FPGA is 48 times that of CPU
- **Latency:**
- FPGA is reduced to 6% of CPU

Input:

- Average file size=130k
- Resolution=1024x768
- Total file number=100
- Total files size=51MB
- Format=JPEG

Output:

- Resolution=1024x768
- PSNR=47
- Format=HEIF



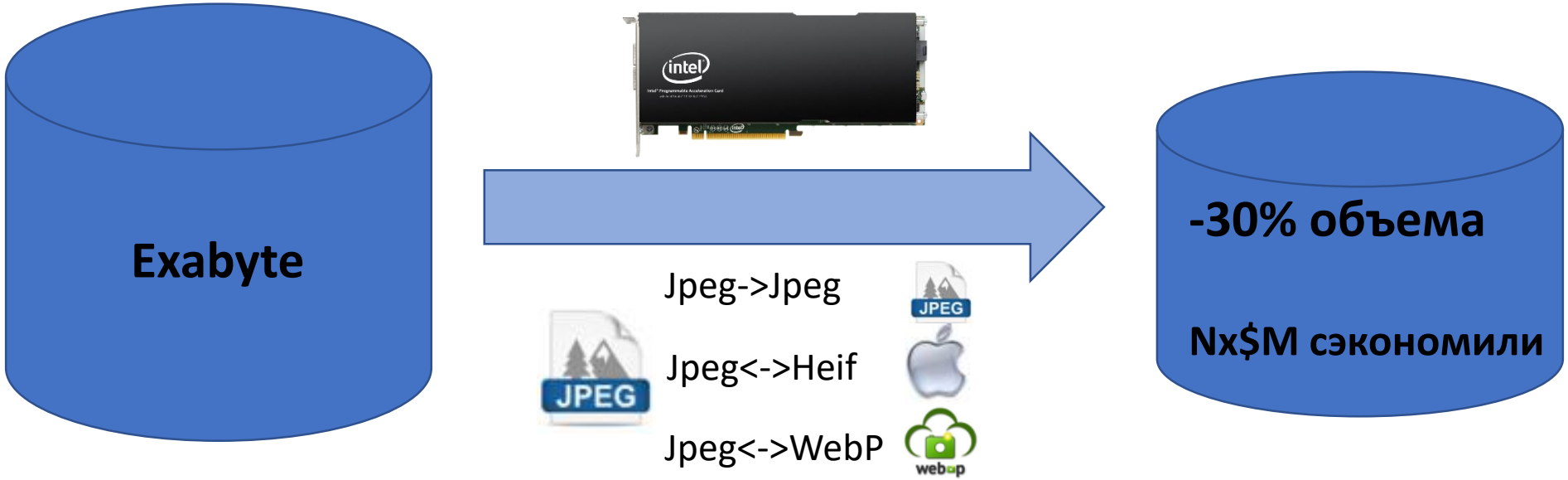
Преобразование изображений на лету

Задача: Сохранить не менее 30% от объема хранения изображений.

Решение: Сервера с ускорителями Intel FPGA Arria A10:
2xIntel Xeon 6230/8xIntel FPGA Arria 10 GX/2x25Gbs Ethernet

Состояние: Удалось уменьшить объем системы хранения на несколько десятков Петабайт
Экономия составила несколько миллионов долларов.

Интересно: Рассматриваем другие типы нагрузок для вынесения их обработки на FPGA



Спасибо!