ЦЕНТР
ПРИКЛАДНЫХ
ИССЛЕДОВАНИЙ
КОМПЬЮТЕРНЫХ
СЕТЕЙ

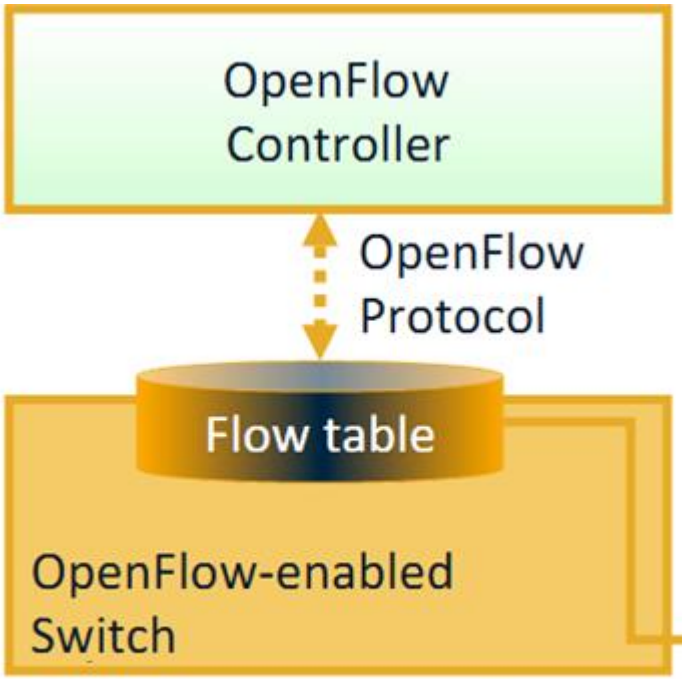
# Разработка OpenFlow коммутатора.

Вячеслав Васин: vvasin@arccn.ru

# Текущая ситуация с сетевым оборудованием:

# OpenFlow:

- **Детальный контроль потоков.**
- **Гибкость в реализации политик.**
- **Поддержка производителей оборудования, исследовательских центров, академических кругов делает его в текущее время стандартом de-facto.**




- Коммутатор может поддерживать только OpenFLow или быть гибридным.
- Режим работы OpenFlow может быть реактивный или проактивный
- Коммутатор может иметь закрытую архитектуру, но обеспечивать открытый интерфейс для управления своей таблицей коммутации.
- Коммутатор может быть сравнительно простым устройством, а значит более надежным и дешевым

Поиск по заголовку пакета

| Switch Port | MAC src | MAC dst | Eth type | VLAN ID | IP Src | IP Dst | IP Prot | TCP sport | TCP dport |
|---|---|---|---|---|---|---|---|---|---|

# Предпосылки:

Ситуация на рынке OpenFlow и гибридных коммутаторов.

Отличия в скорости реализации нового функционала в классическом и OpenFLow коммутаторе.

Отсутствие в России коммутаторов, созданных на российской элементной базе.

## Рынок сбыта

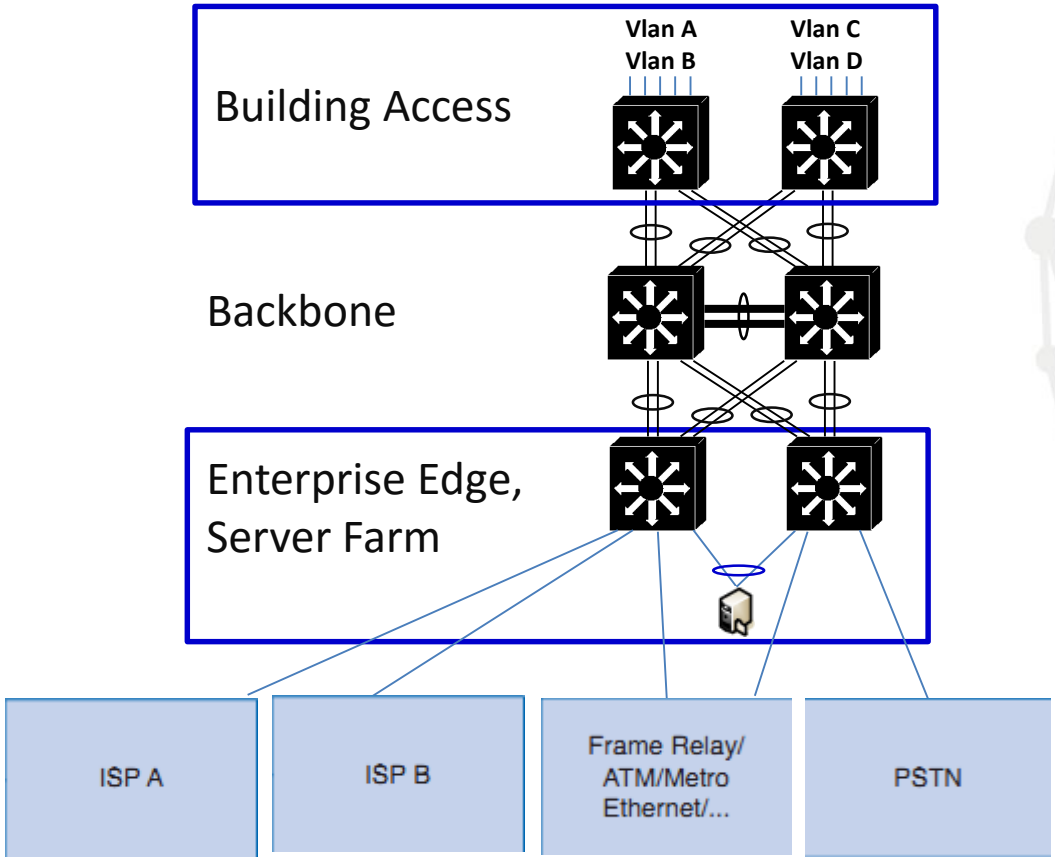
- Государственные компании.
- Компании работающие с кредитными историями граждан.
- Коммерческие компании работающие в области медицины, образования, транспорта.
- Системы связанные с обработкой персональных данных.

# Каким должен быть коммутатор для небольшой корпоративной сети:

- Подключение рабочих мест на скорости 100Mbps-1Gbps.
- Подключение к уровню ядра на скорости 1Gbps-10Gbps.
- Модульность для сохранения инвестиций и гибкости решений.
- Поддержка политик QoS.
- Поддержка механизмов обеспечения безопасности.
- Поддержка инструментов контроля состояния, загрузки, эффективности использования имеющихся линий связи.
- Поддержка возможности виртуализации сетевой инфраструктуры.
- Возможность передачи широковещательного трафика.
- Возможность организации балансировки нагрузки.
- Быстрая сходимость для достижения минимального времени простоя например за счет перенаправления пакетов на запасные маршруты.
- Динамическое управление сетевой инфраструктурой в зависимости от требований приложений и текущей загрузки.

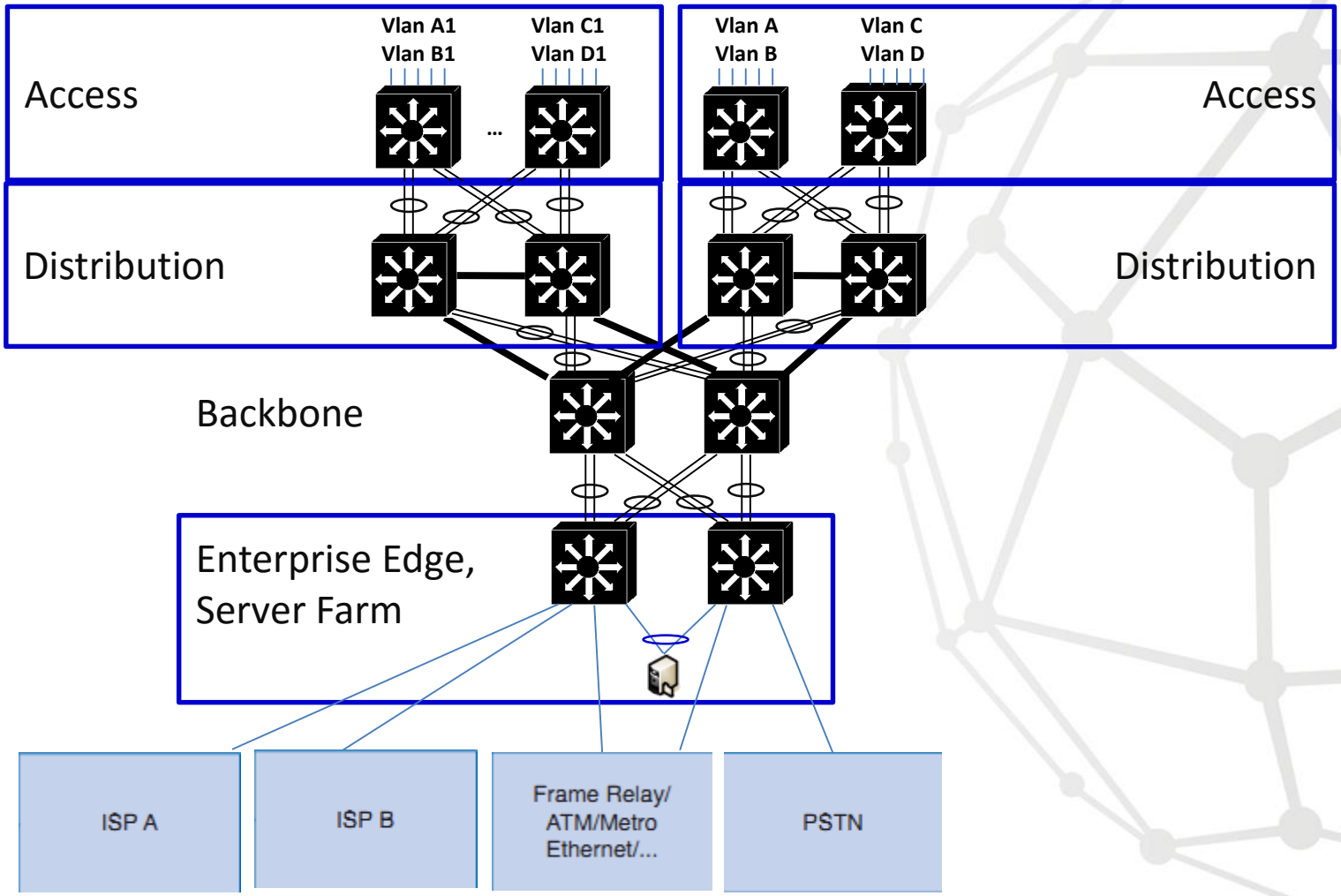
# Архитектура малой корпоративной сети:

- Небольшая корпоративная сеть – до 150 пользователей
- Уровни ядра и распределения объединены.



# Архитектура средней корпоративной сети:

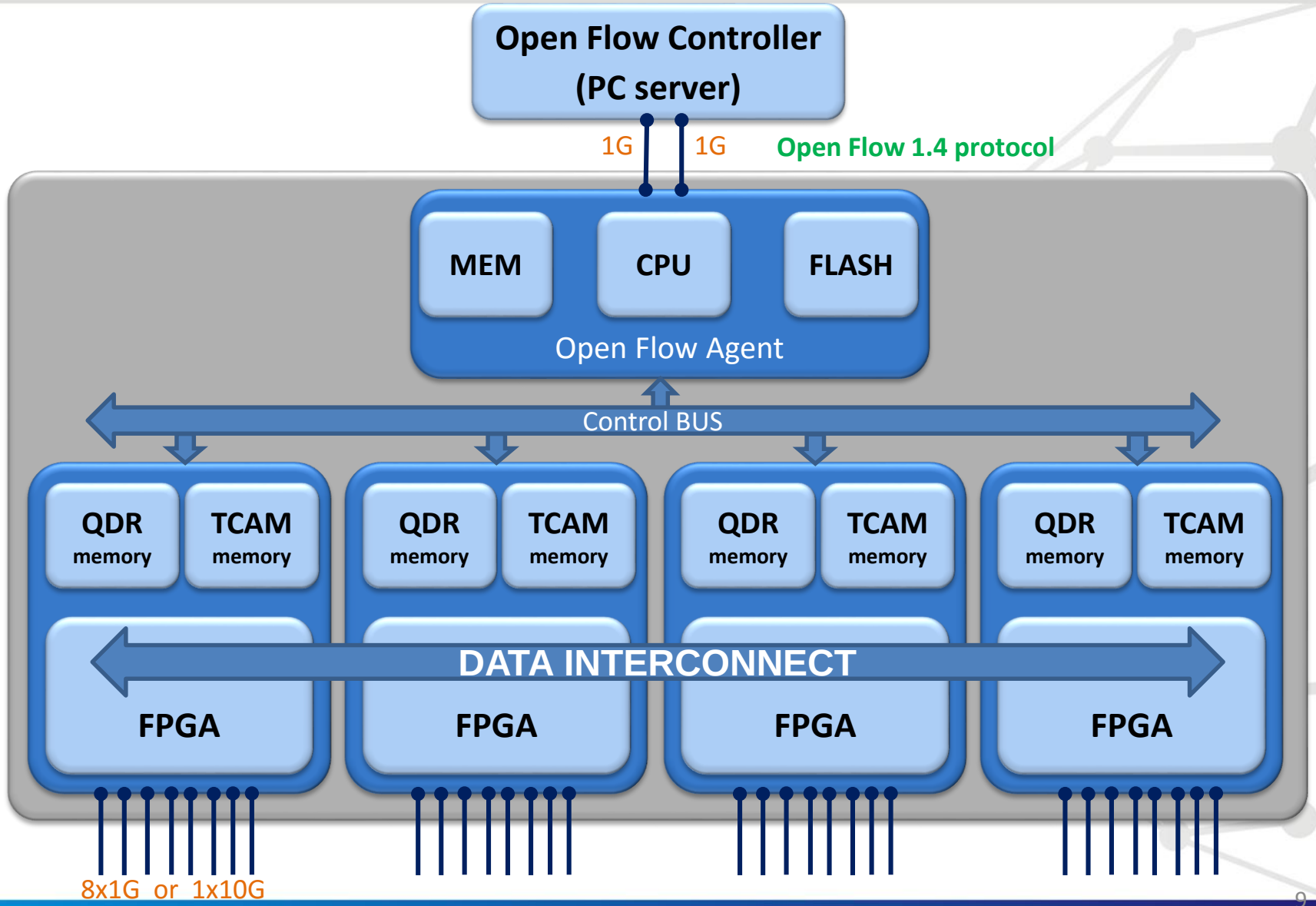
Корпоративная сеть – до 500 пользователей



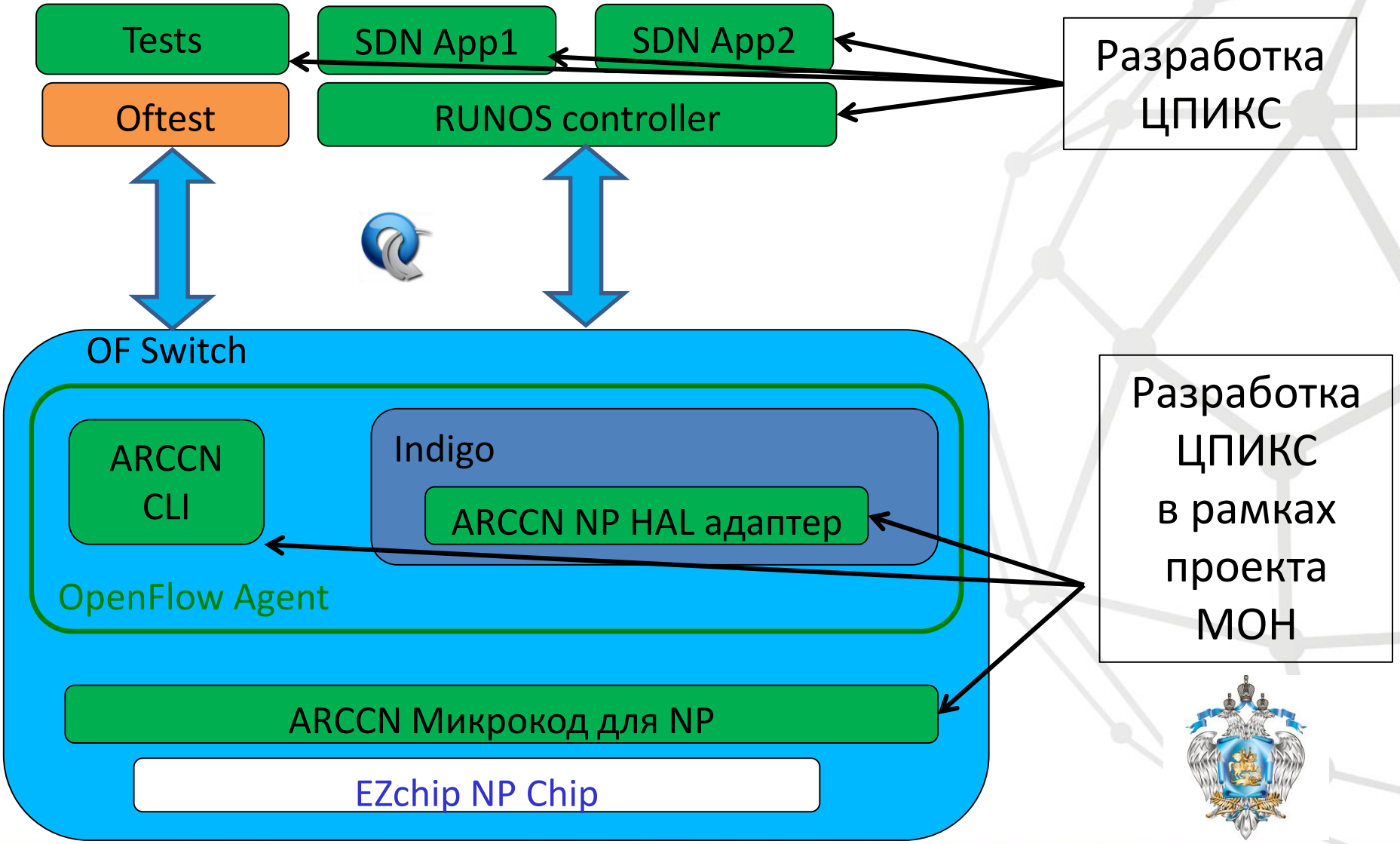
# ARCCN OpenFlow Switch:

- Enterprise-class Multigigabit Ethernet Switch.
- Модульная платформа на 4 слота.
- 2 типа модулей (8x1Gbps портов и 1x10Gbps порт).
- Высокоскоростное взаимодействие между модулями на скорости 10 Gbps
- Распределенная архитектура.
- Прототип на на базе Xilinx Virtex-7 family FPGA (XC7V485T-2).
- Оперативная память QDR II SRAM Cypress (CY7C1515KV18)
- Использование TCAM и/или SRAM для хранения записей Flow Table.
- 4 Flow Table различной длины (до 16000 записей).
- 4 входящих и 4 исходящих очереди на каждый порт для реализации политик качества обслуживания.
- Частичная поддержка спецификации OpenFlow 1.4 (полная поддержка в будущих релизах ПО).
- 2x1Gbps порта для OOB соединения с OpenFlow контроллером.
- Высокоскоростной Control Plane процессор Xilinx Zynq-7030 (1GHz, dual-core ARM Cortex-A9 MPCore) для выполнения функции Open Flow агента.
- Производство будущей коммерческой версии полностью на Российской элементной базе.
- Rack-mount 19’ ; Rack height – 1 rack unit; AC input voltage – 220V.

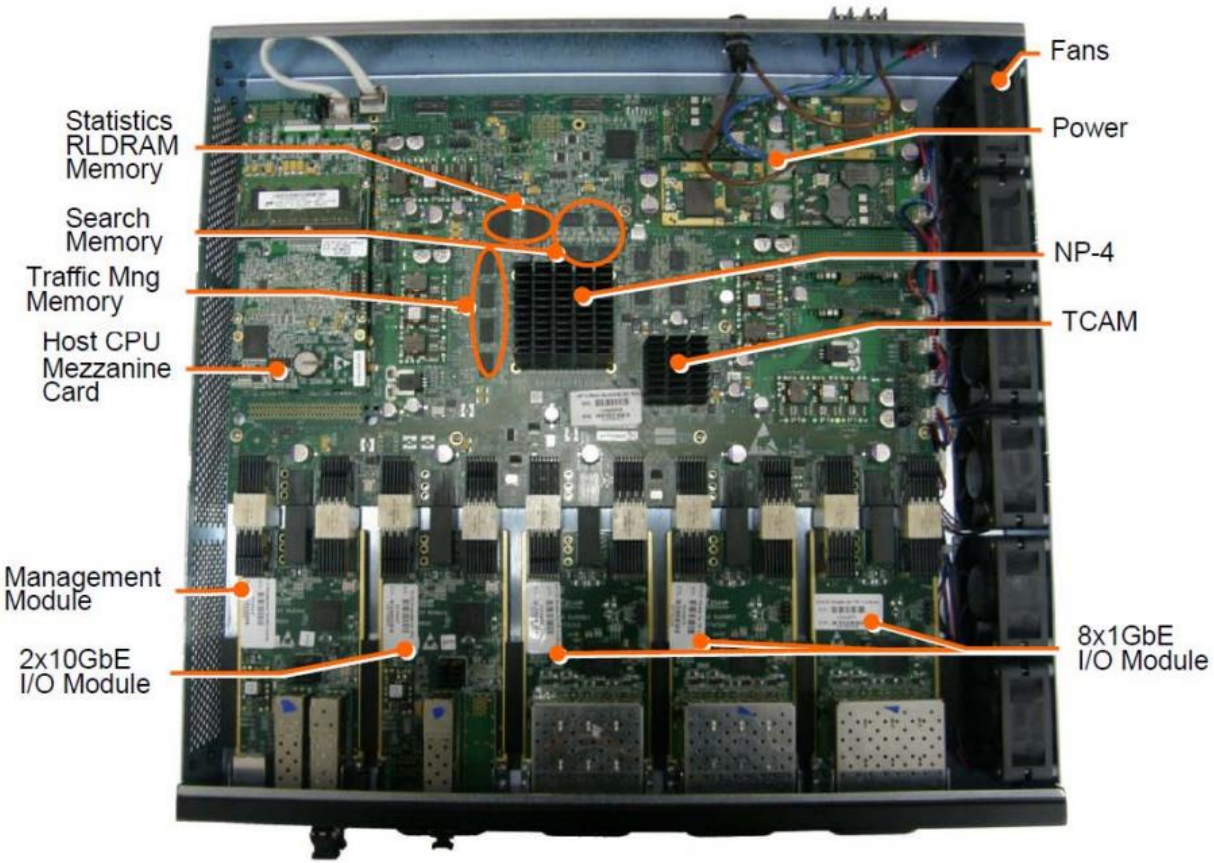
# Архитектура устройства



# Реализация:


Tests
SDN App1
SDN App2
Oftest
RUNOS controller
Разработка ЦПИКС
OF Switch
ARCCN CLI
Indigo
ARCCN NP HAL адаптер
OpenFlow Agent
ARCCN Микрокод для NP
EZchip NP Chip
Разработка ЦПИКС в рамках проекта МОН

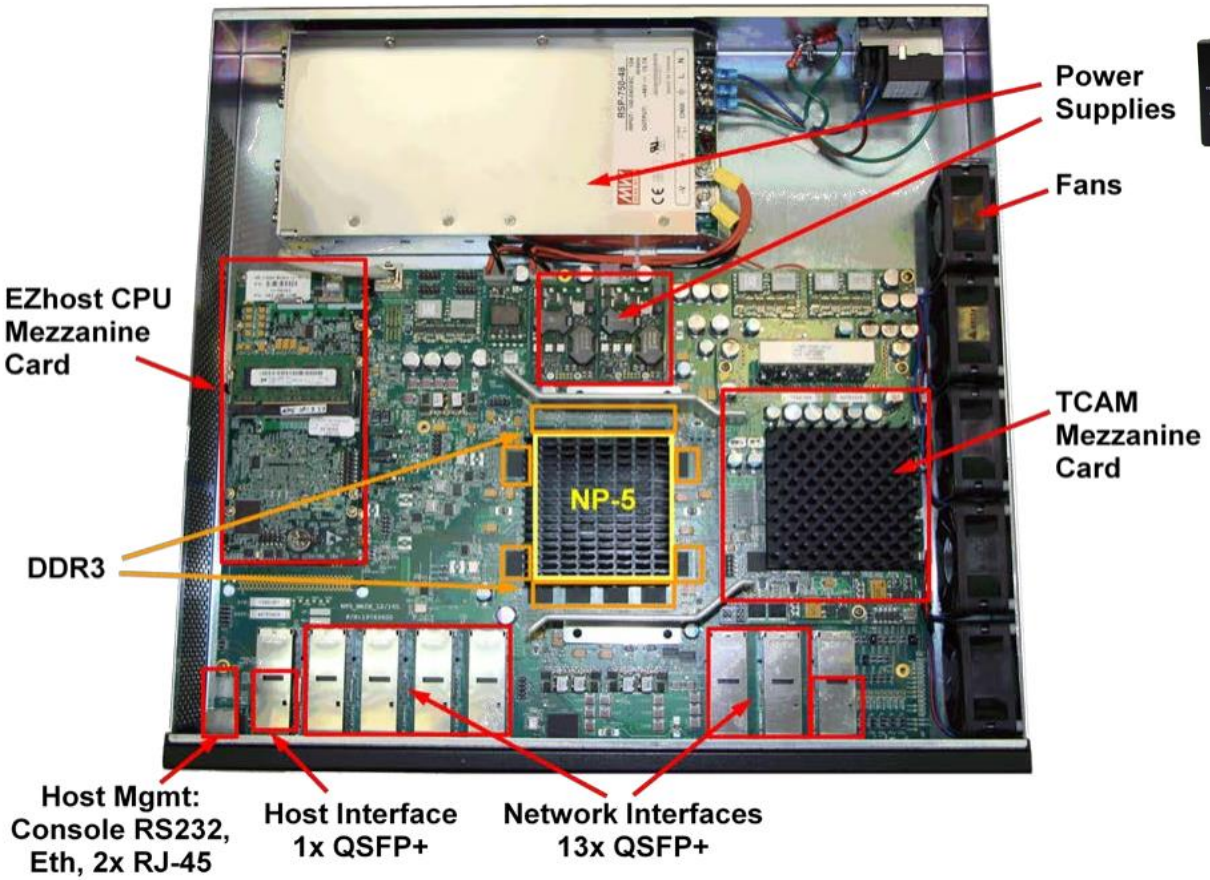
# Характеристики платформы NP-4 (EZchip)



- Модуль управления :
  - ✓ 2x10Гб/с (SFP+)
  - ✓ 2x10/100/1000 Мб/с (SFP)
  - ✓ 1 RS-232 console
- Сетевые интерфейсные модули (4 шт) :
  - ✓ 2x10Гб/с (SFP+) или
  - ✓ 8x1Гб/с (SFP)
- Высота 1.5U
- Монтаж в стойку 19”

- Сетевой процессор производительностью 100 Гб/с, 180 миллионов пакетов в секунду
- Иерархическая система очередизации пакетов. Поддержка WFQ с приоритетной очередью.
- Поддержка WRED, Shaping (CIR, PIR), Per flow metering, marking, policing для миллионов потоков.
- L2-4 switching/routing. Поиск по шаблону вплоть до 80 байт.

# Характеристики платформы NP-5 (EZchip)



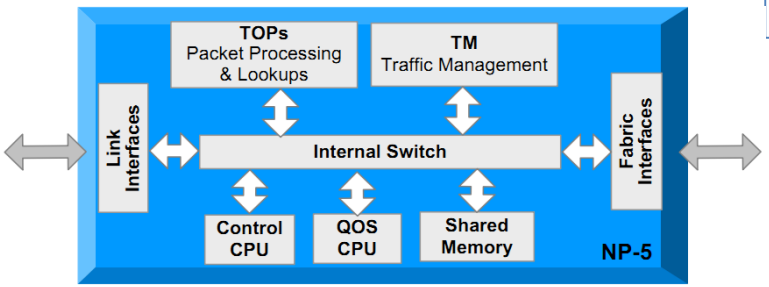
- Модуль управления:
  - ✓ 4x10/1Гб/с
  - ✓ 1x1Гб/с
  - ✓ 1 RS-232 console
- Сетевые интерфейсы:
  - ✓ 44x10Гб/с
    или
  - ✓ 11x40Гб/с QSFP+
    или
  - ✓ 4x100ГБ/с
- Высота 1.5U
- Монтаж в стойку 19”

- Сетевой процессор производительностью до 240 Гб/с, 500 миллионов пакетов в секунду
- 5-уровневая иерархическая система очередизации. Поддержка WFQ с приоритетной очередью.
- Поддержка WRED, Shaping (CIR, PIR), Per flow metering, marking, policing для миллионов потоков.
- L2-4 switching/routing. Сбор статистики по потокам, программируемые пороги. 512M счетчиков.

# Характеристики платформы NP-5 (EZchip)

NP-5 Sub-system

- NP-5 – 500MHz (EZchip)
- Task Optimized Processors (processing/lookup)
- Unified memory 12GB DDR3 800MHz (Micron)
  - Frame size from 16B to 16KB
  - Up to 16GB frame memory (for 24GB DRAM)
- Internal TCAM – 2.5Mbit
  - 365M lookups per second
- External TCAM - 40Mbit, 300MHz (Broadcom)
  - 20B/40B/60B/80B key lookups
- Integrated Search Engine (table, hash, tree, TCAM)
  - 48 bytes keys for hash
  - 18 bytes keys for trees
  - 80 bytes keys for TCAM
  - Up to 16GB lookup tables headroom



# Аппаратные особенности Host CPU NP-4 (NP-5)

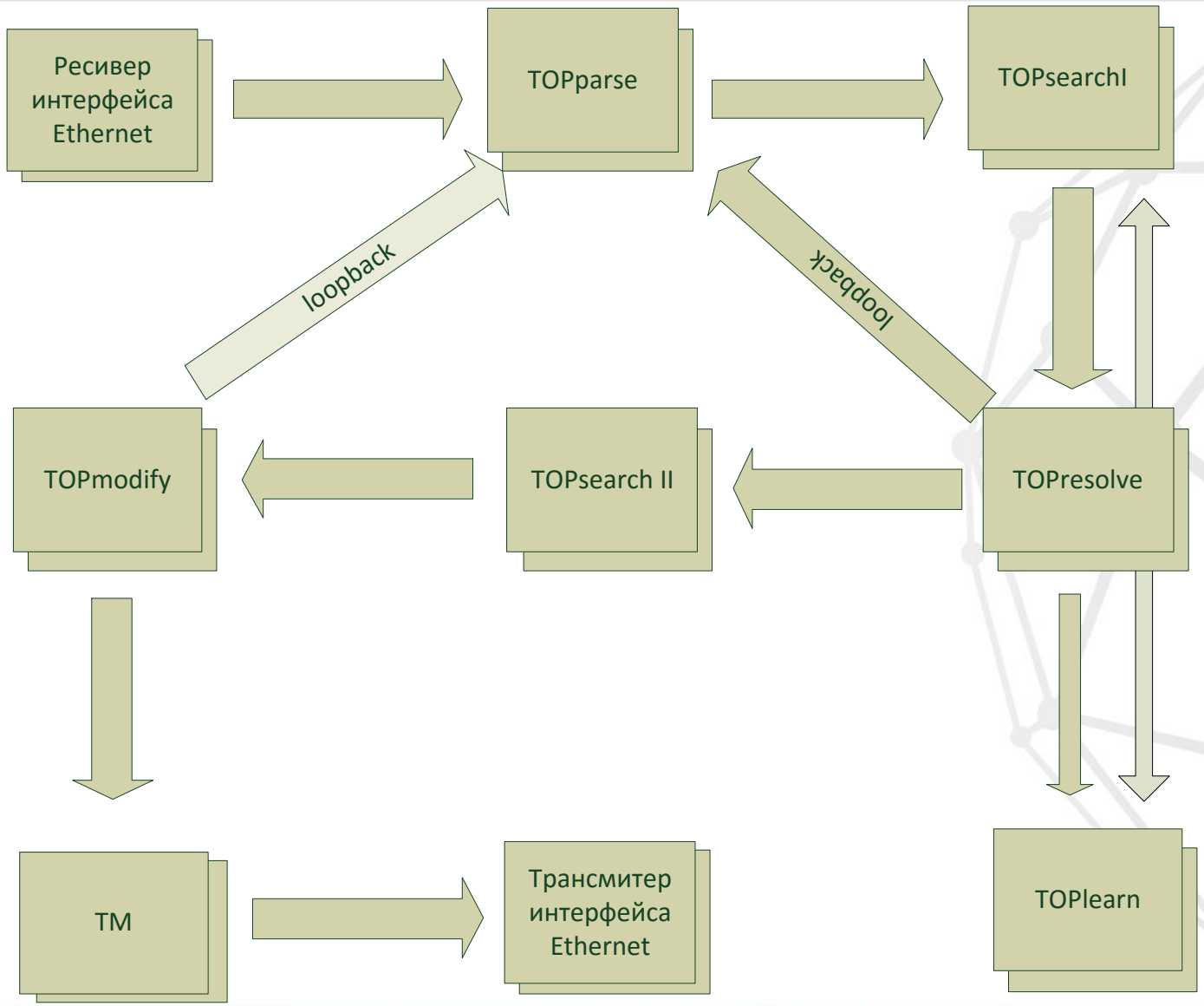
Карта расширения, выполняющая функции блока управления

Решаемые задачи:
- взаимодействие с контроллером
- интерфейс администратора
- изменение настроек коммутатора
- мониторинг и управление

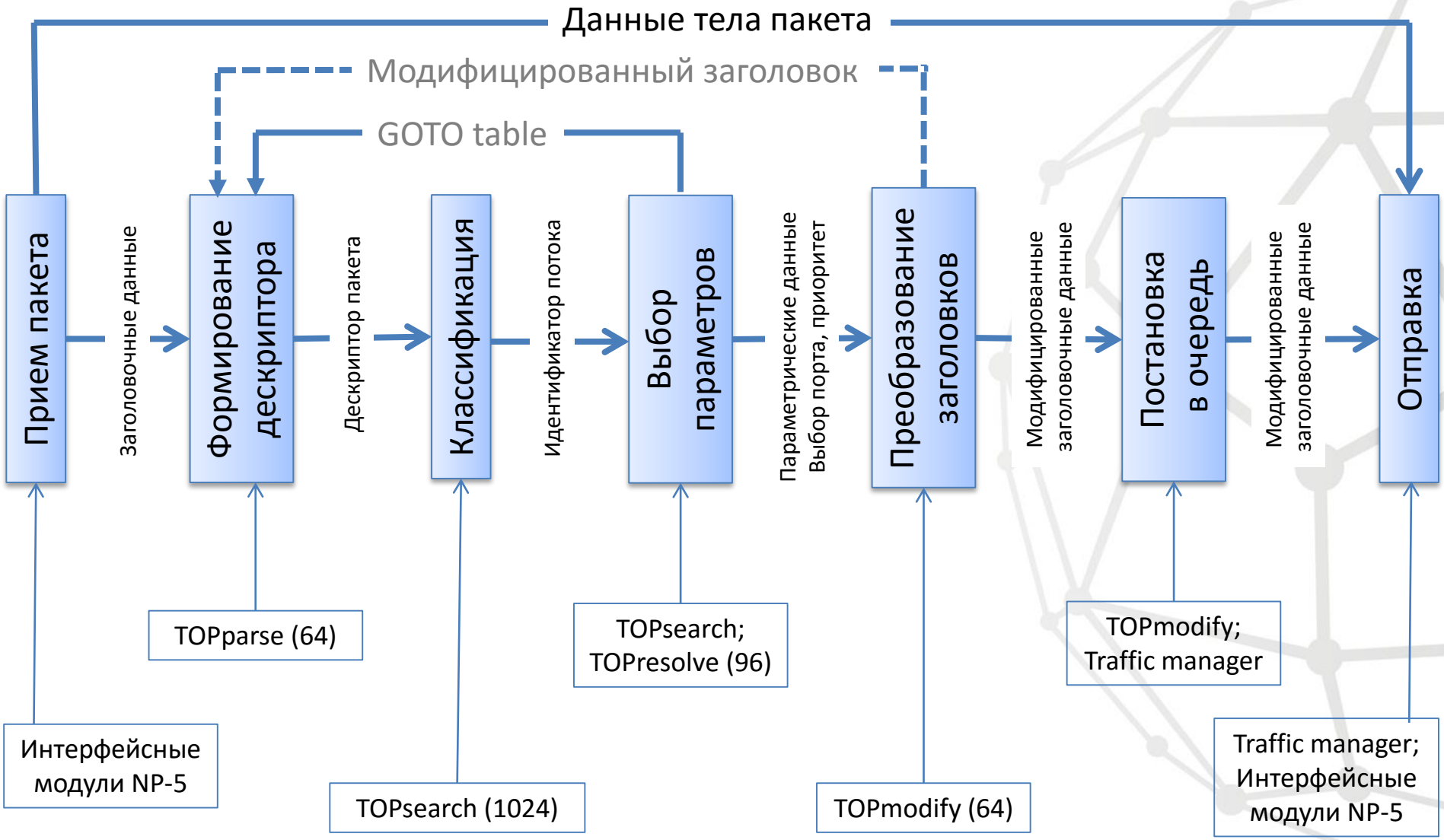
Характеристики:
OS: Linux 2.6 GCC ver. 4.2.1;
CPU: Freescale MPC8543 (792 МГц)
Mem: DDR2 667MHz 512 МБ;
HDD: SDD 256 МБ;
EEPROM: 2 КБ.

Интерфейсы:
- мост PCI (66 МГц) – PCI-express;
- Ethernet порт – 10/100/1000 Мб/с.
- RS-232 (для управления)

# Стадии обработки данных в NP:



# Тракт обработки данных



# Проблематика выбора и организации таблиц OpenFlow коммутатора

Несколько последовательно опрашиваемых таблиц, реализованных с помощью TCAM

Возможность маскирования всех нужных полей

Возможность реализации любой политики приложений клиентов

TCAM - дорогостоящий ресурс.

TCAM имеет высокое энергопотребление.

# Проблематика выбора и организации таблиц OpenFlow коммутатора

Несколько разнородных таблиц оптимизированных под выполнение конкретной задачи в сети клиента.

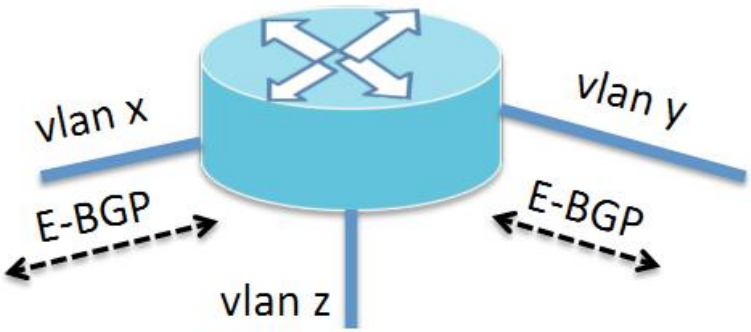
Возможность реализовывать таблицы с миллионами записей

Сложность в адаптации приложений под текущую структуру и архитектуру взаимодействия таблиц

Сложности с потоками требующими проверку маскированных полей

# Проект Atrium

OPEN NETWORKING FOUNDATION




Open Source Routing

ONOS

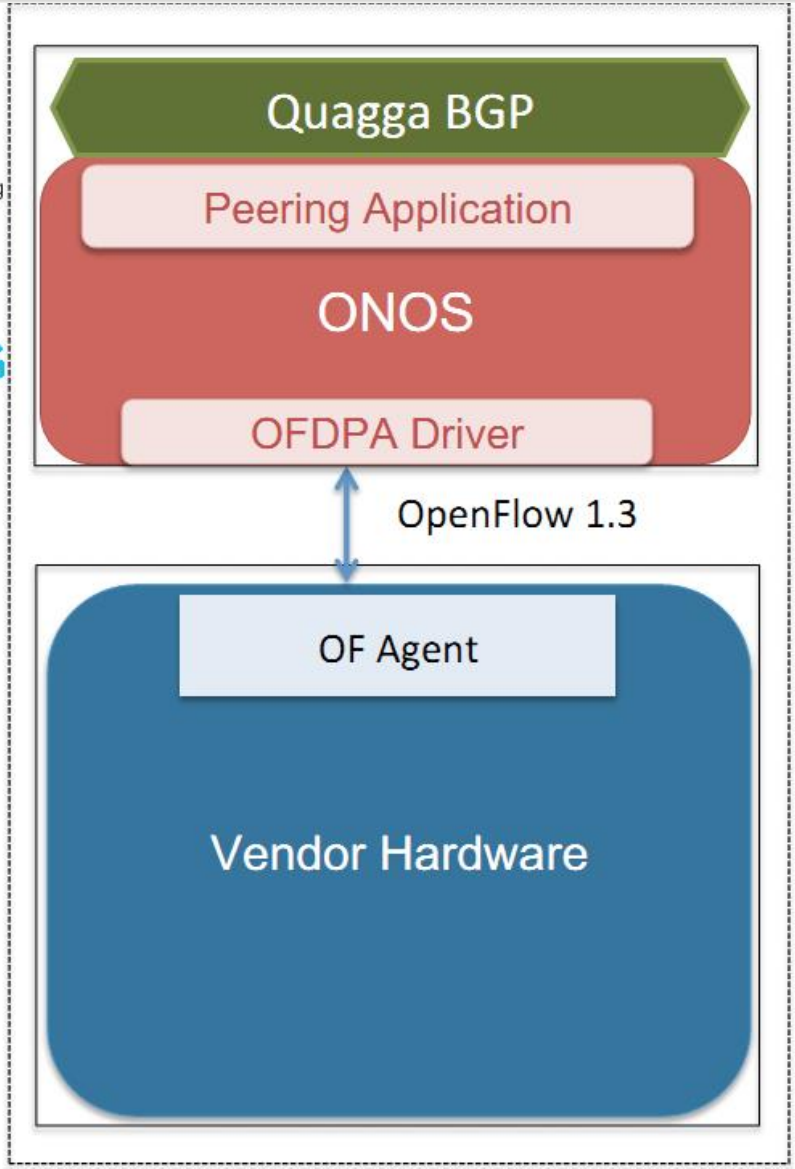
PICA8 OPEN NETWORKING

NoviFlow switching made smarter

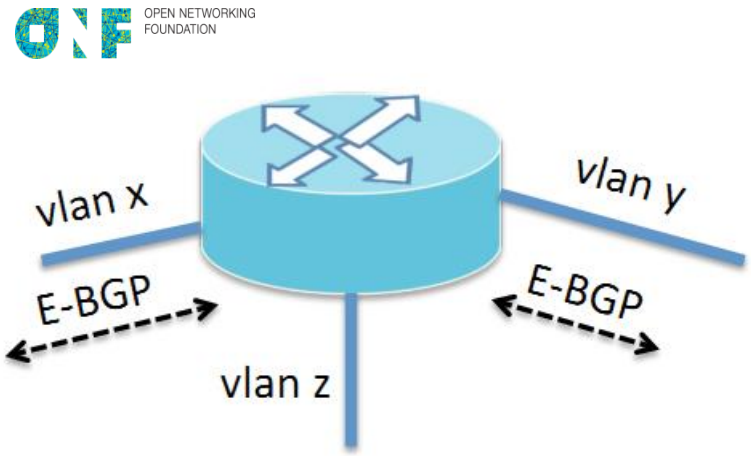
NETRONOME

CORSA

centec networks



# Проект Atrium




Open Source Routing

onos

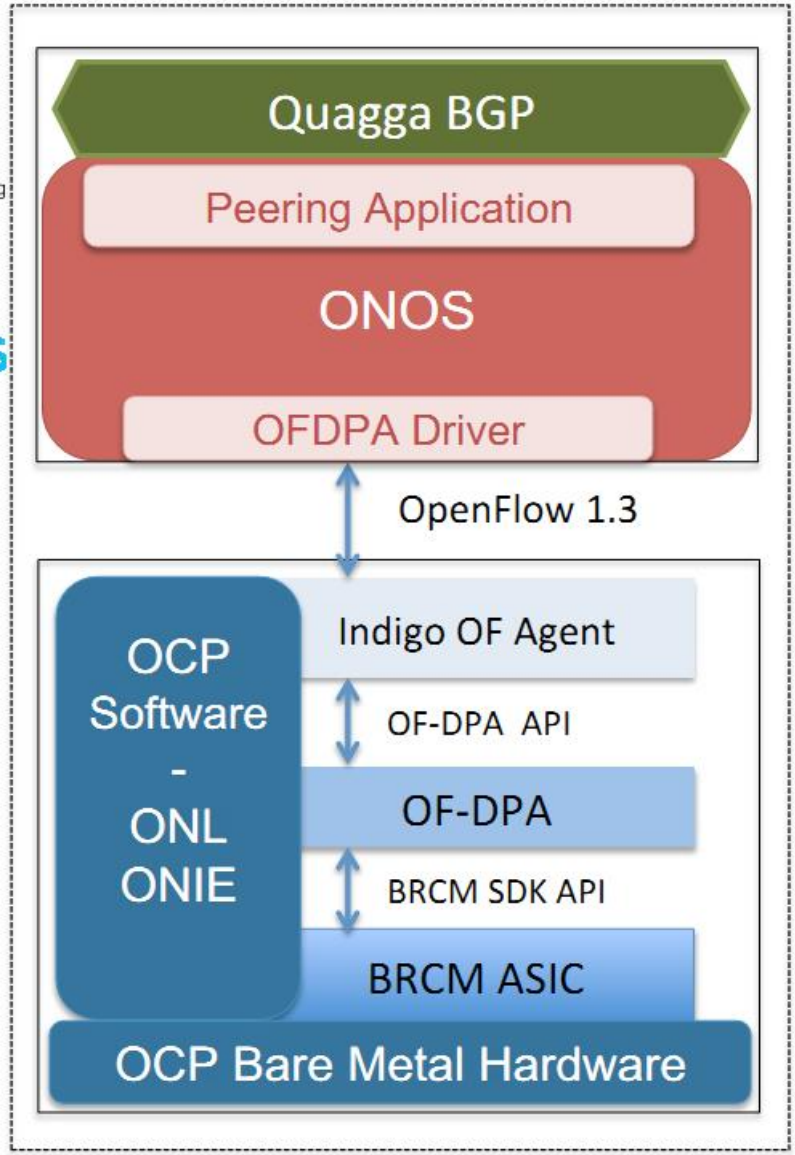
big switch networks

Accton
Making Partnership Work

BROADCOM

Quanta

OPEN Compute Project



# Проект Atrium

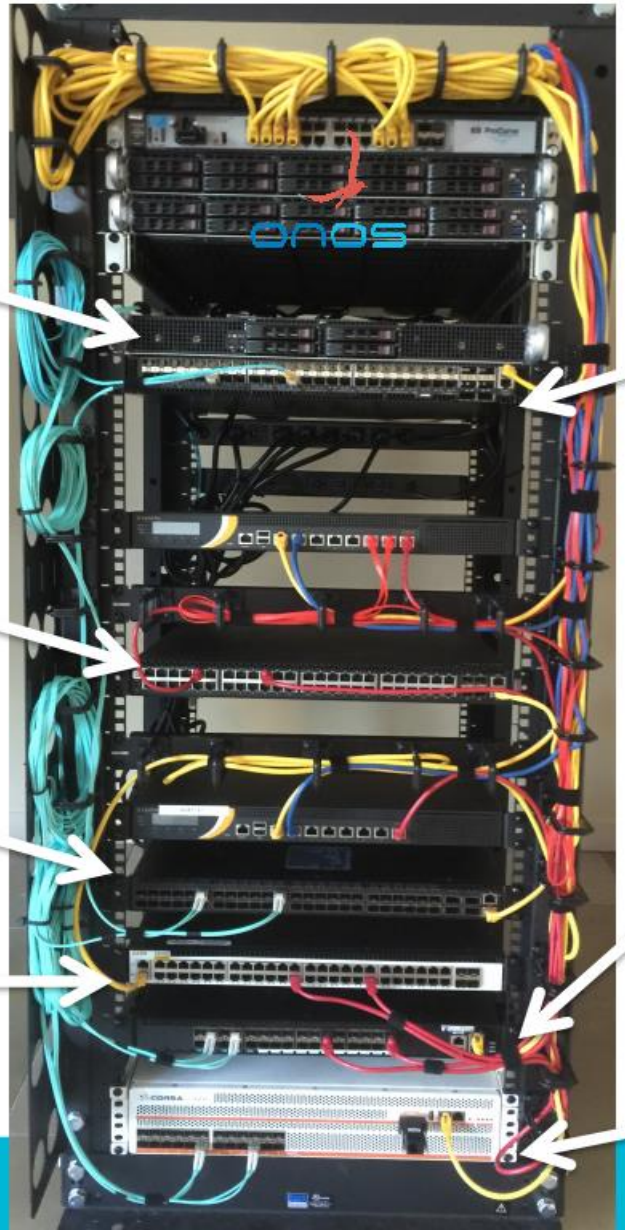
OPEN NETWORKING FOUNDATION

NETRONOME

PICA8 OPEN NETWORKING

Quanta

centec networks



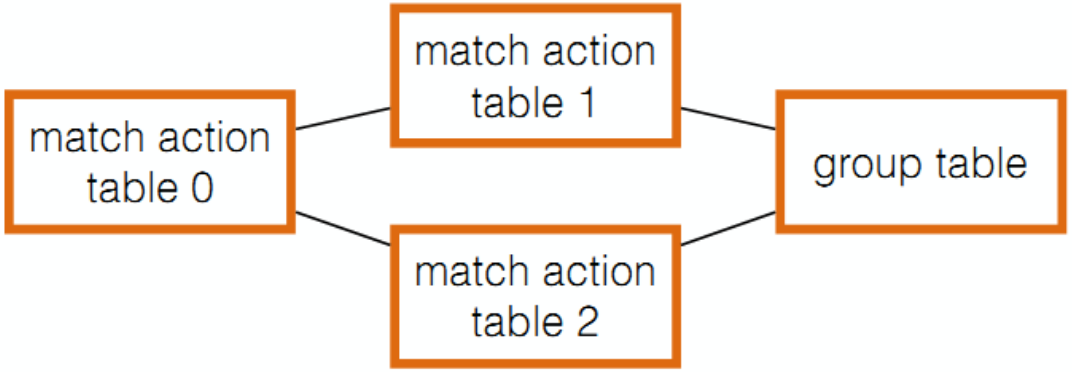

Accton
Making Partnership Work

NoviFlow
switching made smarter

CORSA

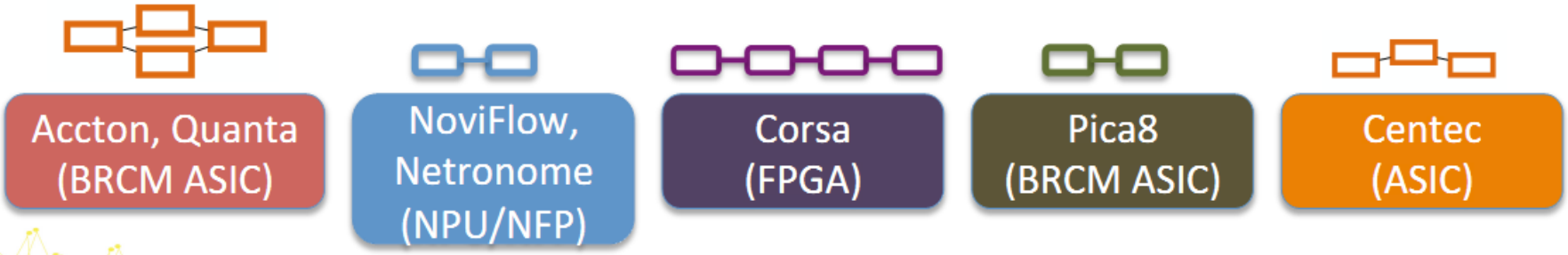
# Проблема взаимозаменяемости

Different combinations of Match-Action Tables form Different OF 1.3 pipelines

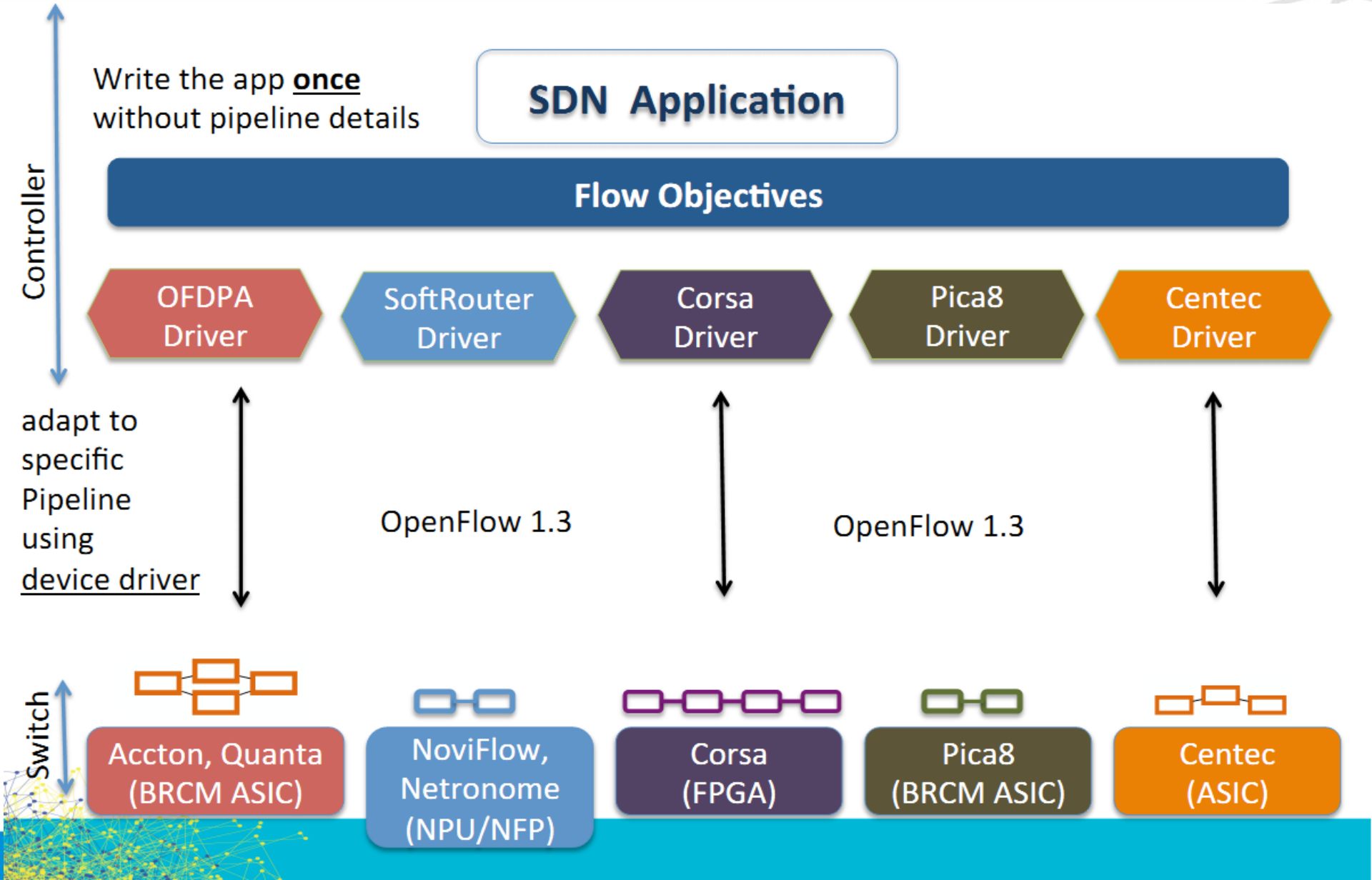



How can a controller manage these differences?

How can an application work across these differences?

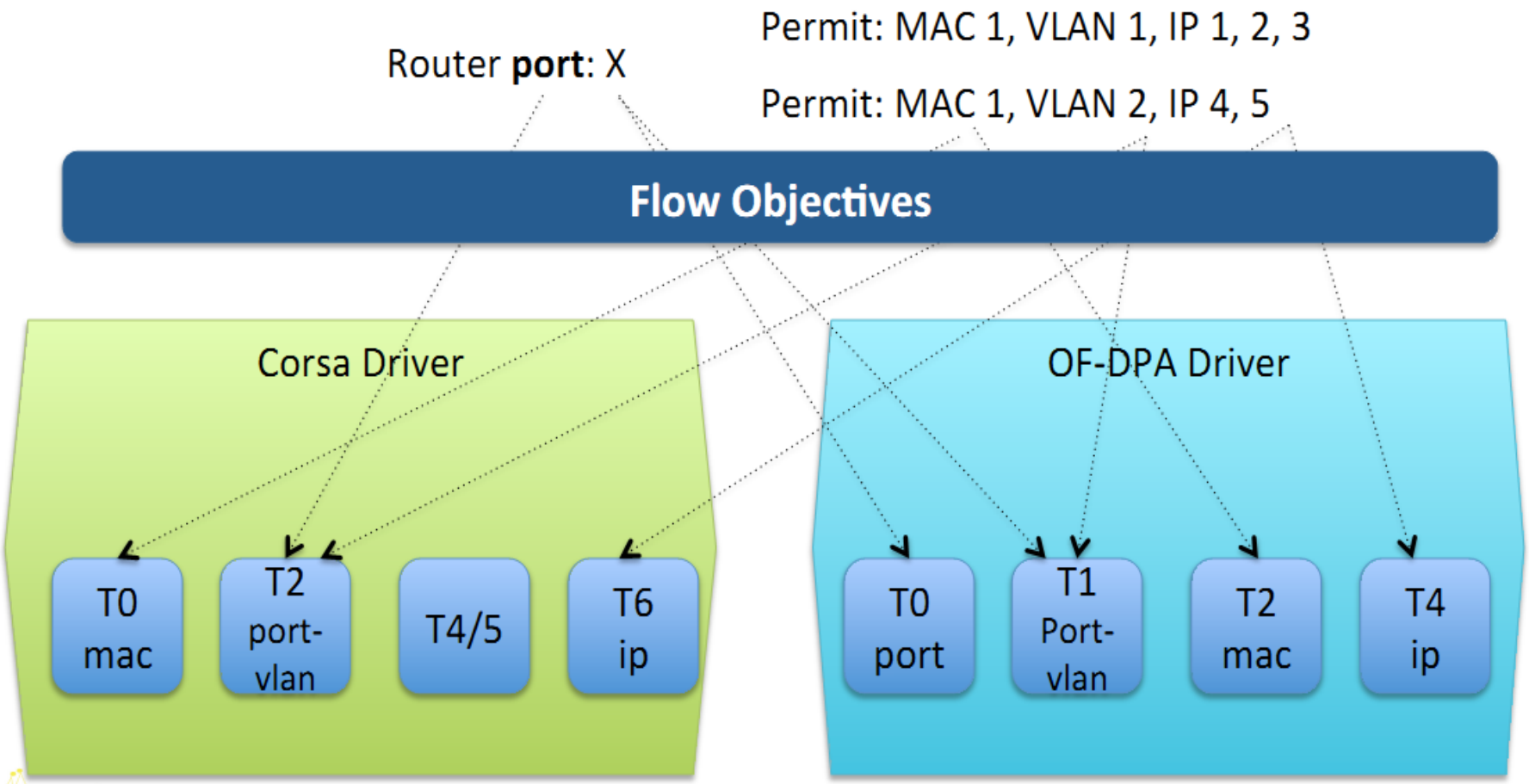


# Предлагаемое решение – скрыть от приложений детали реализации

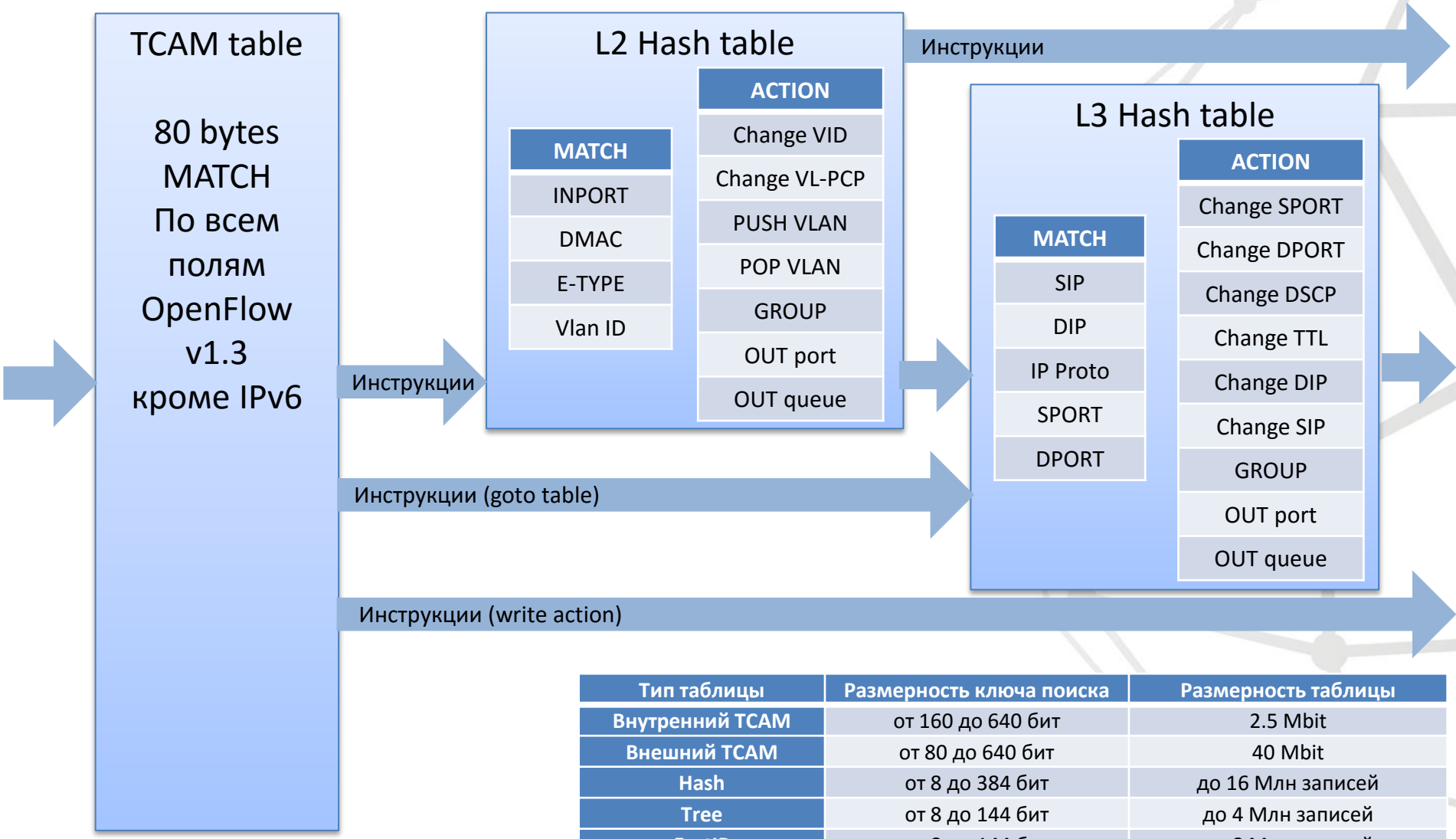


# Пример реализации фильтрации.

- Filter ➔ only Permit or Deny options
- On match fields of packet header

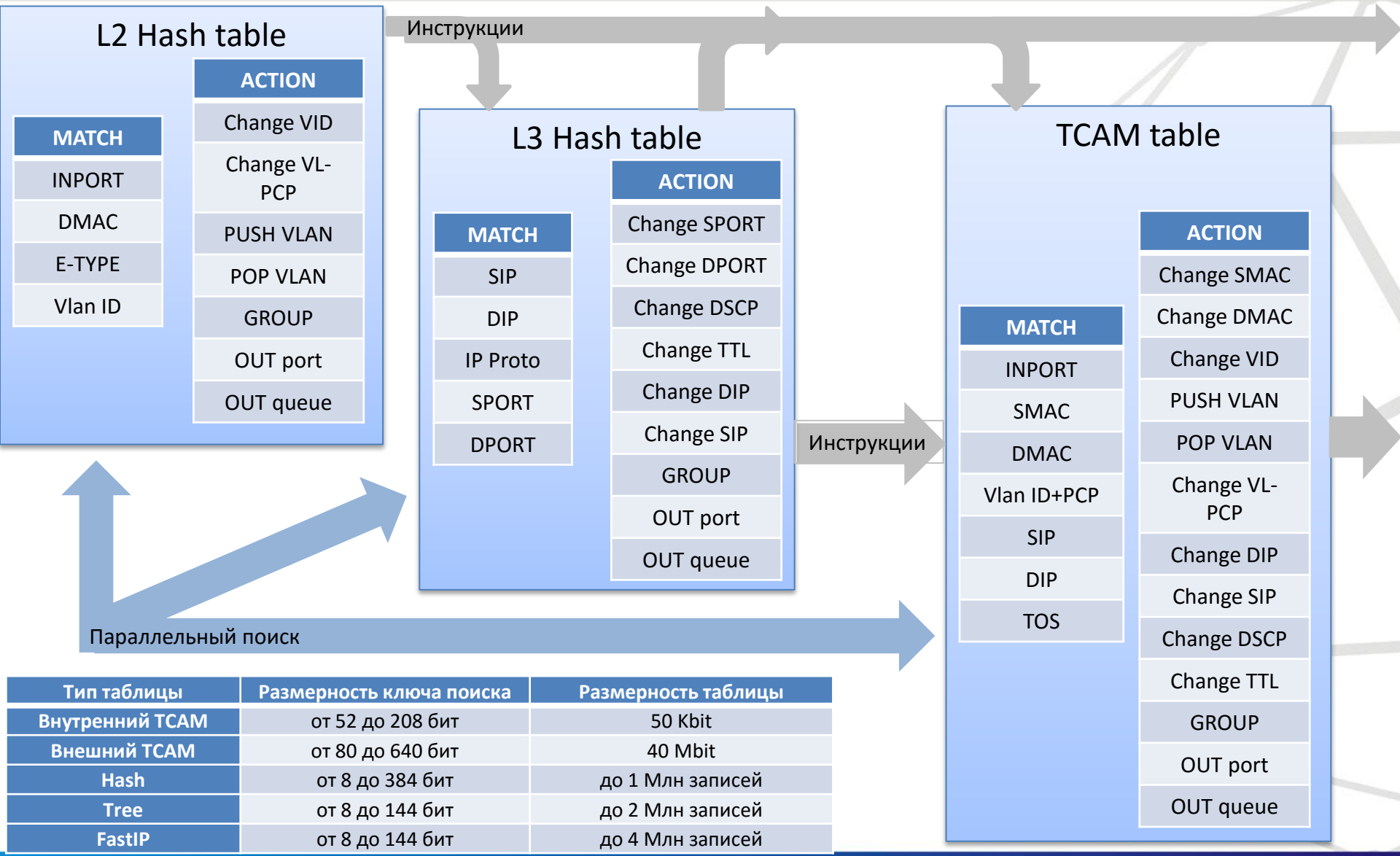


# Последовательный тракт обработки данных




| Тип таблицы | Размерность ключа поиска | Размерность таблицы |
|---|---|---|
| Внутренний TCAM | от 160 до 640 бит | 2.5 Mbit |
| Внешний TCAM | от 80 до 640 бит | 40 Mbit |
| Hash | от 8 до 384 бит | до 16 Млн записей |
| Tree | от 8 до 144 бит | до 4 Млн записей |
| FastIP | от 8 до 144 бит | до 8 Млн записей |

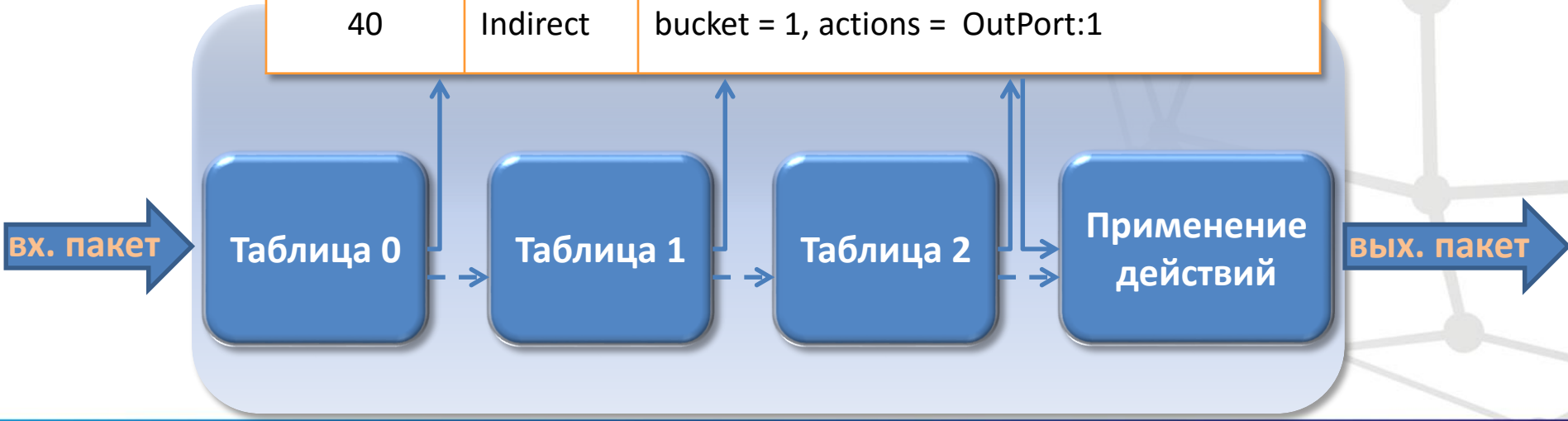
# Параллельный тракт обработки данных




| Тип таблицы | Размерность ключа поиска | Размерность таблицы |
|---|---|---|
| Внутренний TCAM | от 52 до 208 бит | 50 Kbit |
| Внешний TCAM | от 80 до 640 бит | 40 Mbit |
| Hash | от 8 до 384 бит | до 1 Млн записей |
| Tree | от 8 до 144 бит | до 2 Млн записей |
| FastIP | от 8 до 144 бит | до 4 Млн записей |

# Групповые действия

| Group ID | Type | Action Bucket(s) |
|---|---|---|
| 10 | All | bucket = 1, actions = OutPort:1<br>bucket = 2, actions = OutPort:2<br>bucket = 3, actions = OutPort:3<br>bucket = 4, actions = OutPort:4 |
| 20 | Select | bucket = 1, actions = OutPort:1<br>bucket = 2, actions = OutPort:2 |
| 30 | Fast Failover | bucket = watch_port:1, actions = OutPort:1<br>bucket = watch_port:2, actions = OutPort:2 |
| 40 | Indirect | bucket = 1, actions = OutPort:1 |



# Примеры

При использовании параллельного тракта обработки получаем:

- минимизацию задержки на обработку пакета для задач L2 коммутации
- минимизацию задержки на обработку пакета для задач сетевой трансляции (NAT)
- более эффективное использование ресурсов платформы ввиду большего количества TOPsearch (1024) по сравнению с TOPresolve (96) и TOPmodify(64)
- Возможность снизить требование к количеству записей в TCAM путем организации достаточного количества таблиц разных типов, покрывающих потребности наиболее часто используемых в сети приложений

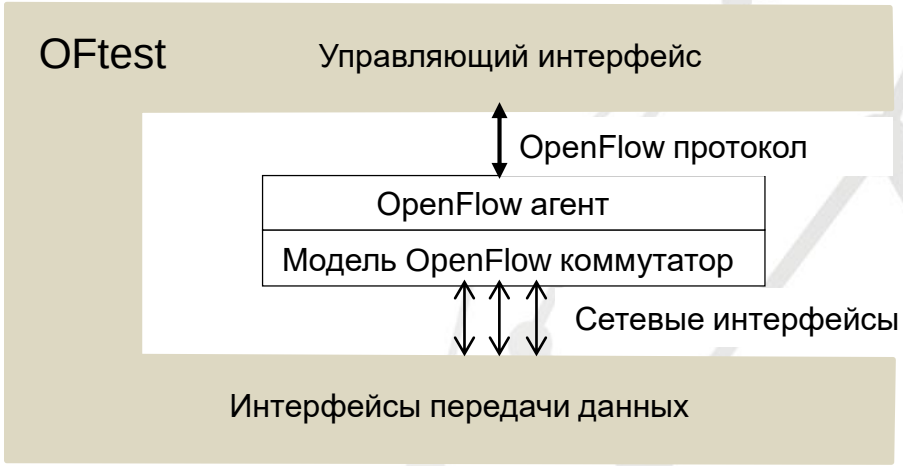
При использовании последовательного тракта обработки получаем:

- Более эффективный поиск при использовании маскируемых полей, а также за счет просмотра единственной таблицы
- Более эффективное использование ресурсов при решении задач требующих анализа большого количества полей, так как при этом результат достигается за счет просмотра только 1 таблицы
- Более очевидный и понятный  алгоритм обработки потоков с помощью последовательности OF таблиц.
- Возможность реализации Apply Action и использование результатов обработки пакета предыдущей таблицей.

# Тесты

**OFTEST v1.3**

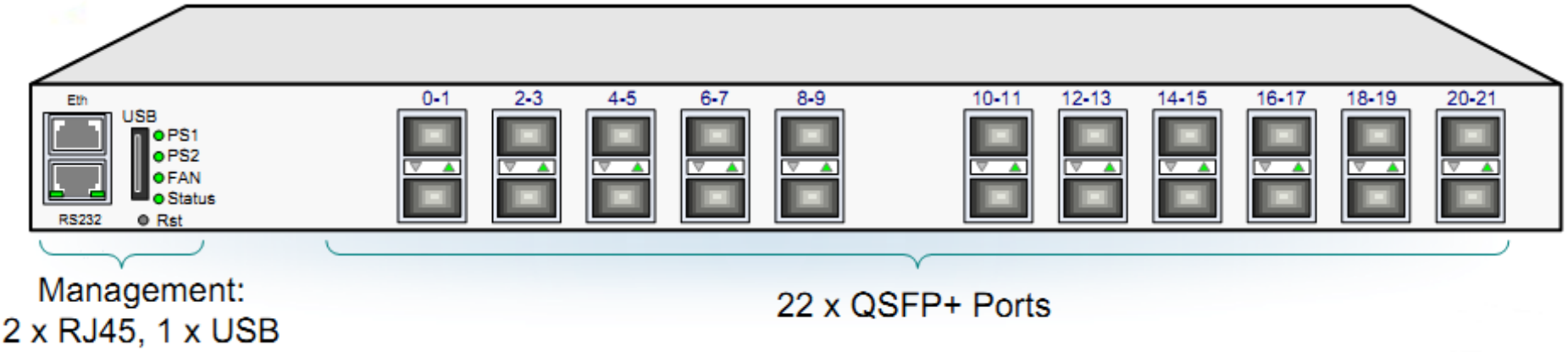
- Basic
  -echo
  -feature request
  -default drop
  -output exact
  -packet in
  -packet out
- Matching fields
  -ipv4
  -arp
  -ethernet
  -vlan
  -mpls
  -in port
- Group
  -groups op.
  -buckets (indirect/select/all)
- Action sets
  -change fields
  -pop or push Vlan / MPLS
  -instruction set
- Role request
  -Permissions for
  slave / equal / master controllers




**Нагрузочное тестирование**

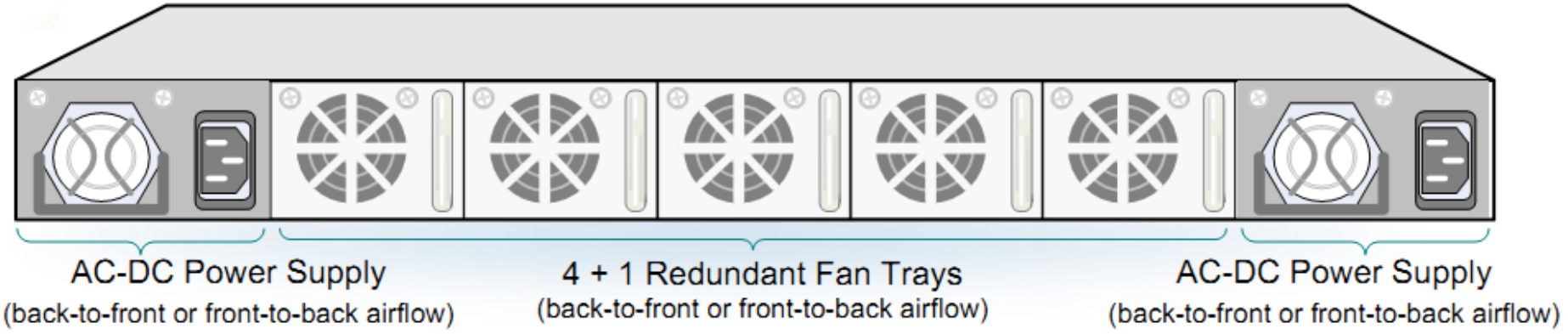
- Подтверждается 10G производительность на пакетах от 64 до 1500 байт, в том числе при поиске по нескольким таблицам, в том числе при изменении IP/TCP/UDP (свыше 10 Mpps)
- До 530 Flowmod в секунду для TCAM таблицы (зависит от последовательности приоритетов добавляемых записей)
- Порядка 4000 Flowmod в секунду для Hash таблиц (независимо от количества записей в таблице)
- Порядка 1000 PACKET_IN в секунду.
- Порядка 2500 PACKET_OUT в секунду.

# NPS Smart NFV TOR/Appliance




- Высота 1U
- Монтаж в стойку 19”
- Сетевые интерфейсы:
  - ✓ 88 x 10GBASE-SR/LR/CR с помощью кабеля
  - ✓ 22 x 40GBASE-SR4/LR4/CR4 QSFP+
  - ✓ 8 x 100GBASE-SR10/LR10 с помощью кабеля
- Модуль управления:
  - ✓ 1 x 10/100/1000 Mbps
  - ✓ 1 RS-232 console RJ-45
- Индикаторы:
  - ✓ Системные: Блок питания - 1, Блок питания - 2, Охлаждение, Статус
  - ✓ Сетевые: Наличие линка, активность линка

# NPS Smart NFV Appliance




- Блок питания:
  - ✓ Вольтаж AC: 100-240V AC, 50-60Hz
  - ✓ Вольтаж DC: 40-72V DC
  - ✓ 1+1 (резервирование)
- Блок вентиляторов
  - ✓ 4+1 (резервирование)
- Вентиляция
  - ✓ Спереди назад
  - ✓ Сзади вперед
- Горячая замена
  - ✓ Блоков питания
  - ✓ Блоков вентиляции

# Блок управления устройством:

- COM (Computer On Module)
  - ✓ Процессор Intel Core i7 4700EQ
  - ✓ Память 8GB DDR3L
  - ✓ Диск 32GB SSD

- Высокоскоростное взаимодействие между NPS и блоком управления:
  - ✓ 64 Gbps PCIe

- Конфигурирование с передней панели:
  - ✓ RS232 serial console RJ-45 порт
  - ✓ Ethernet 10/100/1000 Mbps RJ-45 порт
  - ✓ USB Type 3

# Ключевые характеристики NPS-400

- Сетевой процессор обрабатывающий до 400 Gbps.
- Обработка пакетов на скорости до 600 Mpps
- Обработка транзитного трафика до 7 уровня включительно.
- Возможность перепрограммирования поведения процессора, что обеспечит быстрое внедрение новых технологий и сохранения вложенных средств за счет использования имеющегося оборудования.
- Обеспечение политик качества обслуживания.
- Пятиуровневая иерархическая система обработки очередей, обеспечивающая до 1 миллиона очередей для детальноq идентификации пользователей и их приложений.
- Поддержка IPsec и DPI на скорости до 200 Gbps (server offload)
- Интегрированный TCAM с расширением за счет DRAM для хранения записей.
- Высокоскоростной алгоритм поиска по большому количеству записей в таблицах (Flow Table, ACL)
- Поддержка до 24GB DDR4
- До 512 миллионов разнообразных счетчиков.
- До 6.4 миллиардов операций в секунду по сбору статистики.

# Возможные области применения NPS-400

- Коммутатор поддерживающий OpenFlow v1.4.
- Построение наложенных сетей с использованием QinQ, VxLAN, NVGRE.
- IEEE Edge Virtual Bridging, VN-TAG и Virtual Ethernet Port Aggregator (VEPA).
- Классификация и маршрутизация по потокам.
- Высокоскоростное выполнение фильтрации (ACL offload).
- Балансировка HTTP трафика в зависимости от контента.
- Высокоскоростная терминация IPsec (IPsec gateway).
- Терминация SSL (SSL gateway).
- DPI на скорости до 800Gbps с распознаванием более 1500 сигнатур. Создание собственных сигнатур с использованием множества технологий.
- Lawful Intercept – идентификация и перехват интересующих потоков трафика, сбор передаваемой информации и запись ее в файл.
- Open vSwitch data path acceleration.

Благодарю
за
Внимание!