СЕТИ ПЕРЕДАЧИ И ОБРАБОТКИ ДАННЫХ
Лекция 09:
Сетевые процессорные устройства

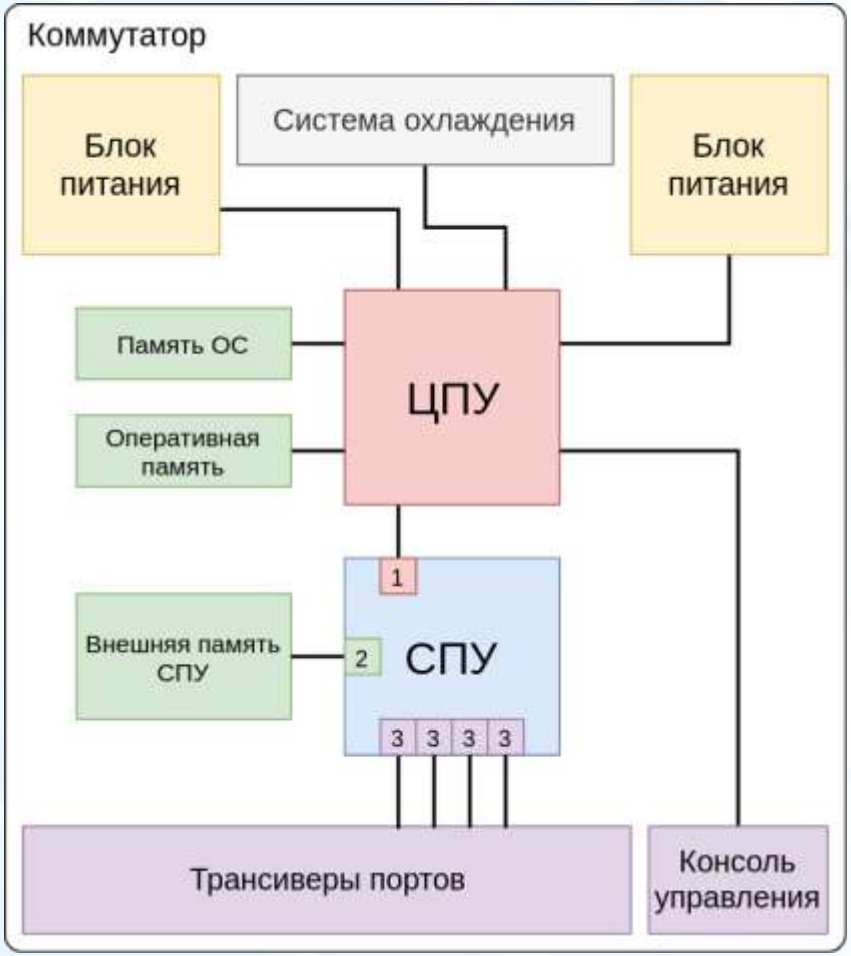
ВМК МГУ им. М.В. Ломоносова, Кафедра АСВК
Доцент, к.ф.-м.н. Волканов Д.Ю.

План лекции

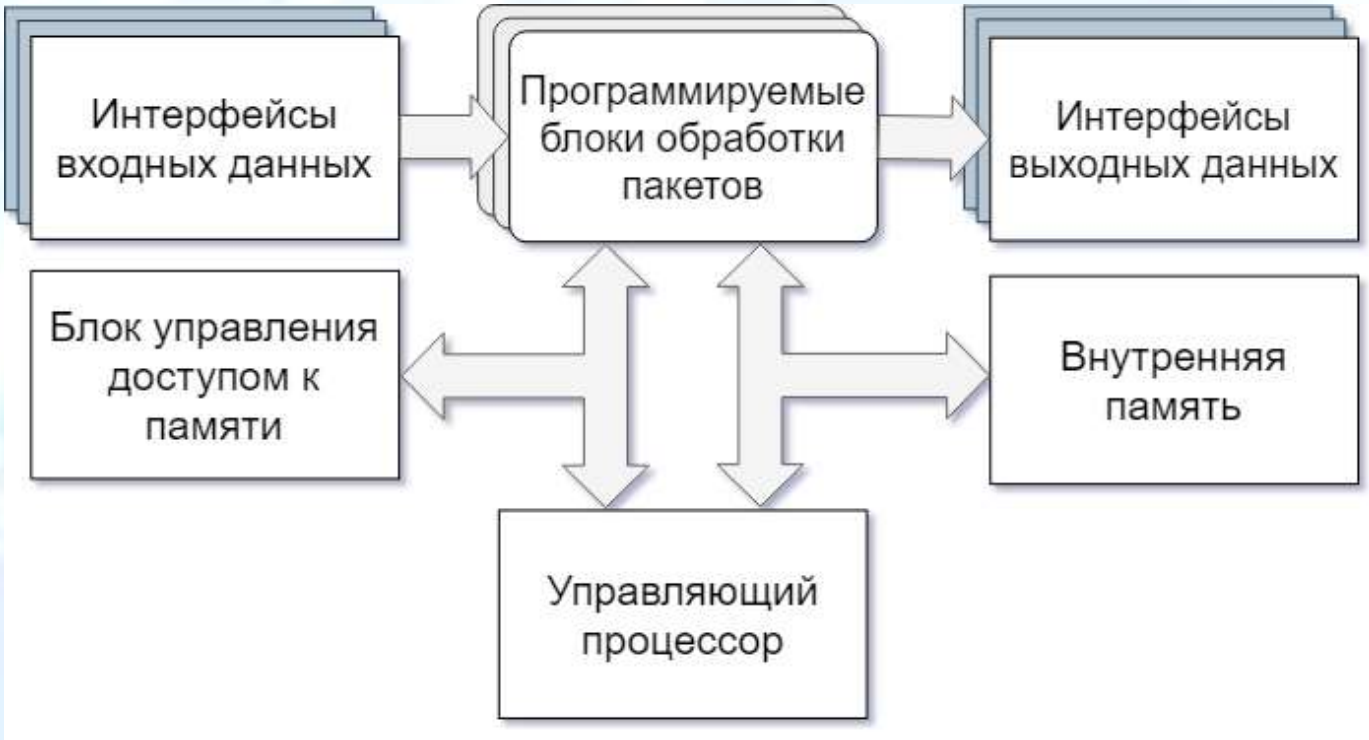
- Общая схема коммутатора
- Основные функции сетевого процессорного устройства (СПУ)
- Жизненный цикл сетевого пакета в СПУ
- Обзор существующих сетевых процессоров

Место СПУ в коммутаторе

- **Сетевое процессорное устройство (СПУ)** – встроенная полупроводниковая система, оптимизированная для выполнения операций передачи данных
- Функции СПУ:
  - получение пакета;
  - выделение заголовка из пакета;
  - классификация пакета;
  - модификация заголовка и принятие решения о пути следования пакета;
  - управление очередями;
  - передача пакета.

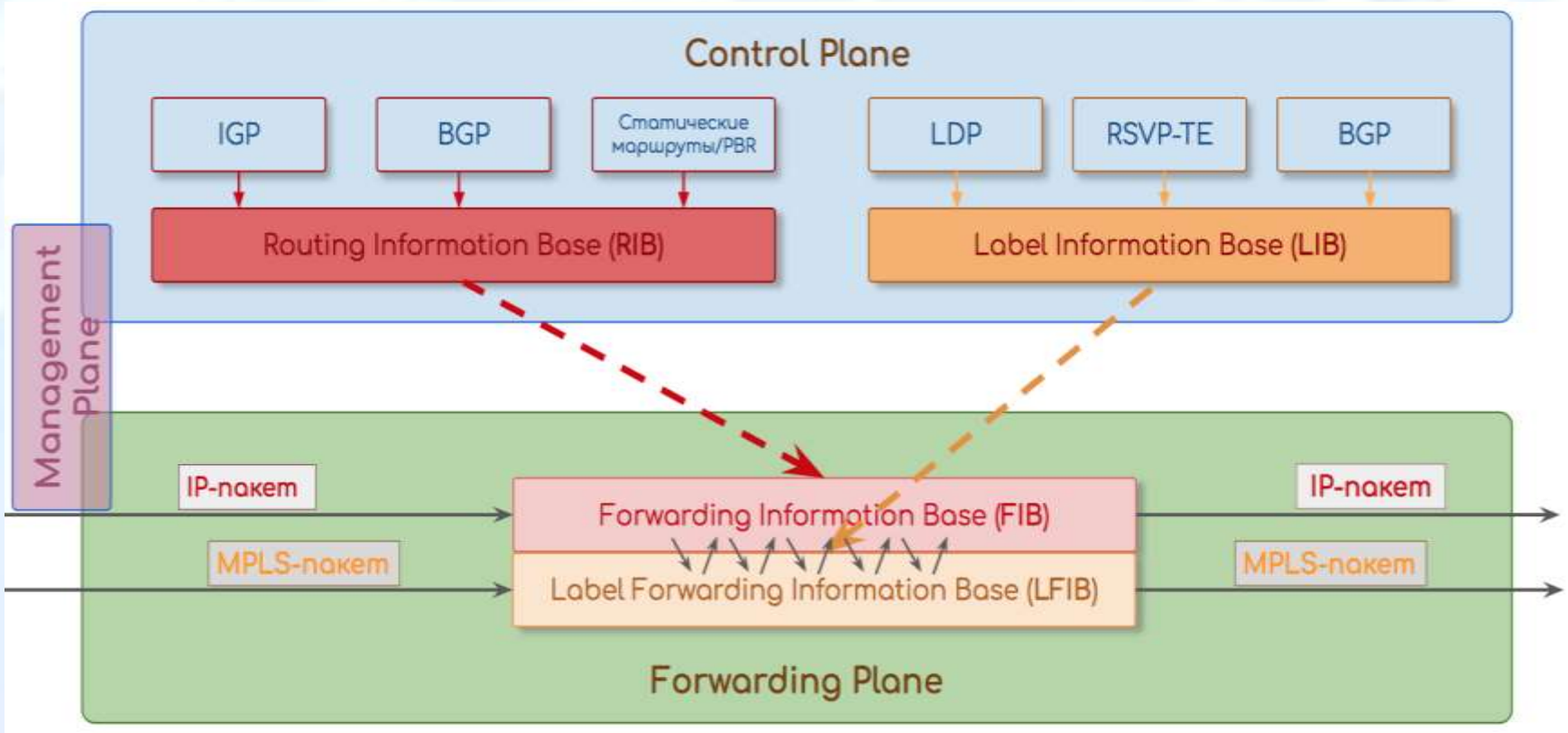


Обобщенная архитектура СПУ

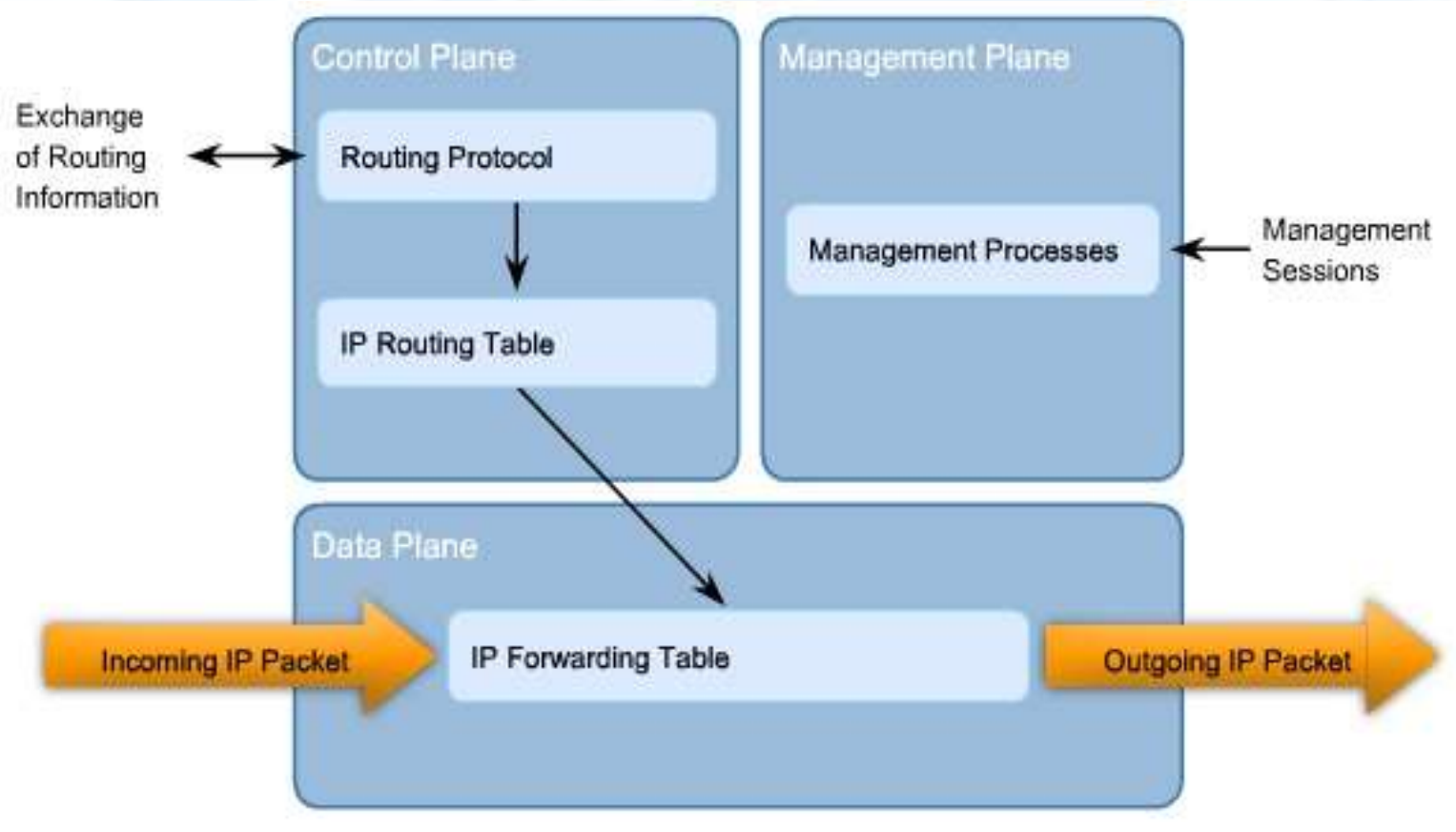


14 байт	20 байт	80 байт	
Ethernet	IP заголовок	TCP заголовок	Полезная нагрузка

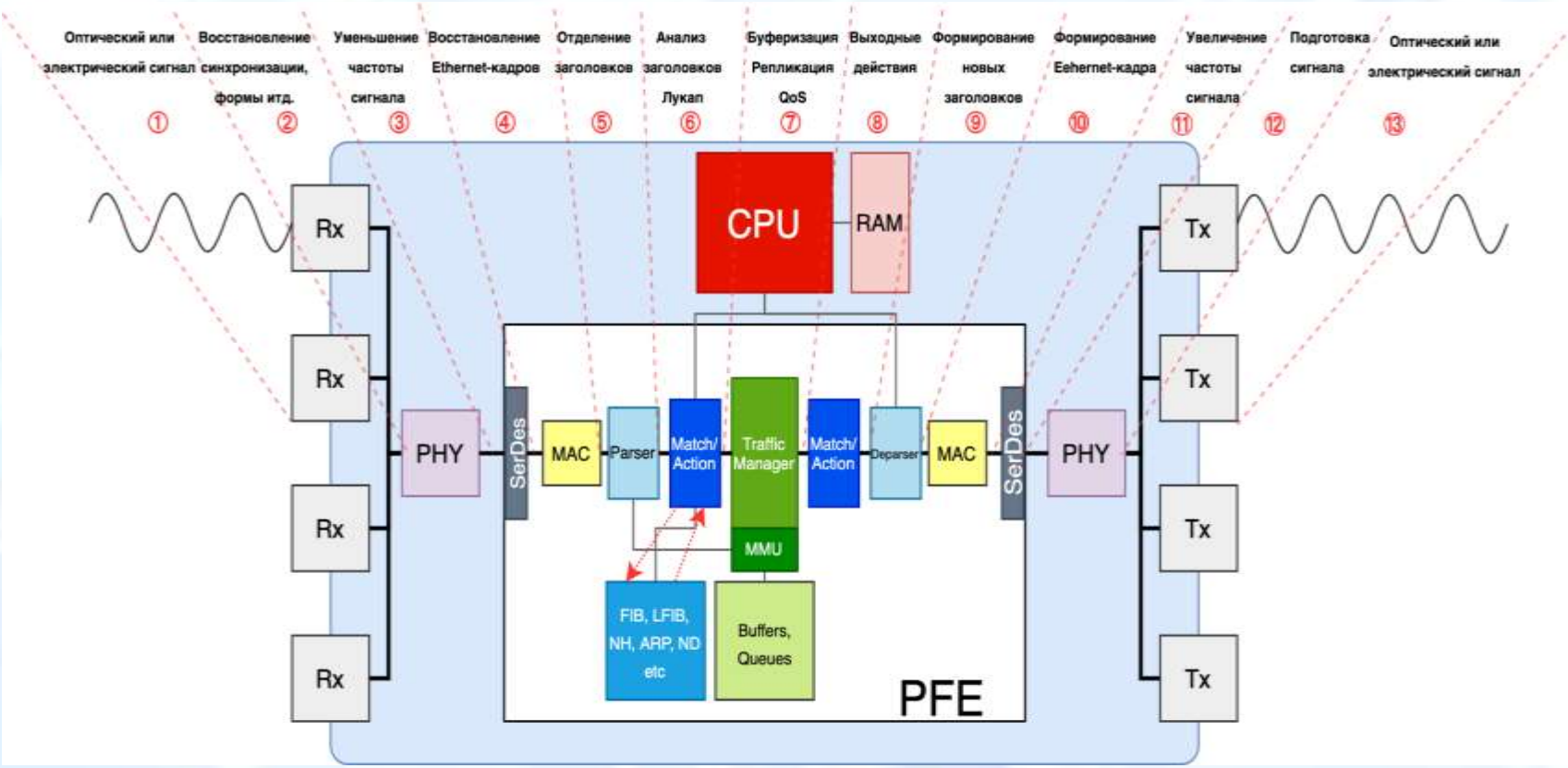
Логические уровни в коммутаторе



Основные блоки



Жизненный цикл пакета в СПУ



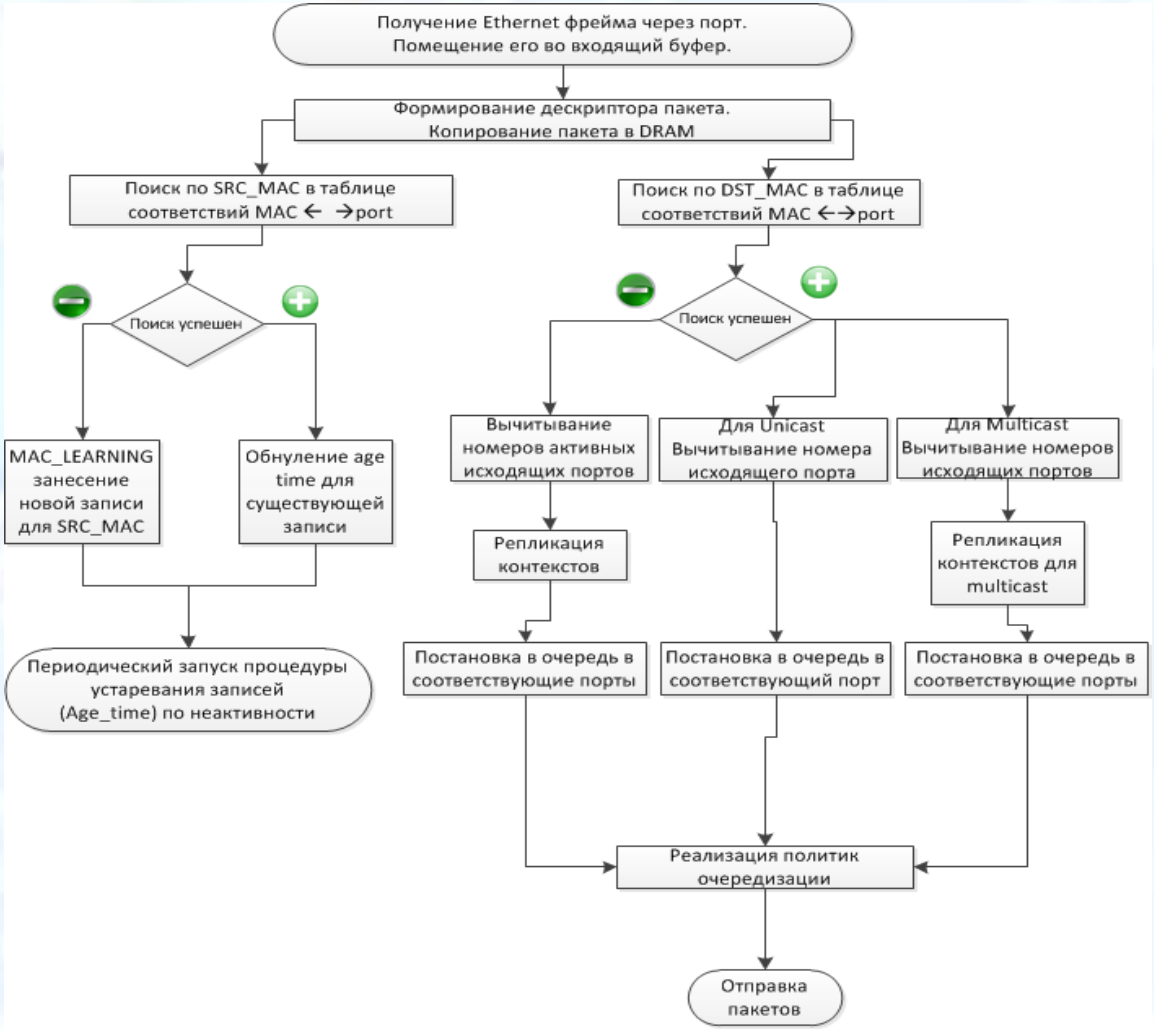
Сценарии обработки пакетов

- Сценарии работы классического L2-коммутатора с обучением
- Сценарии работы L2/L3 коммутатора
- Сценарии агрегирования, очередизации и перенаправления трафика на коммутаторе
- Сценарии обработки трафика в MPLS сетях
- Сценарии реализации протоколов синхронизации времени

Этапы сценариев

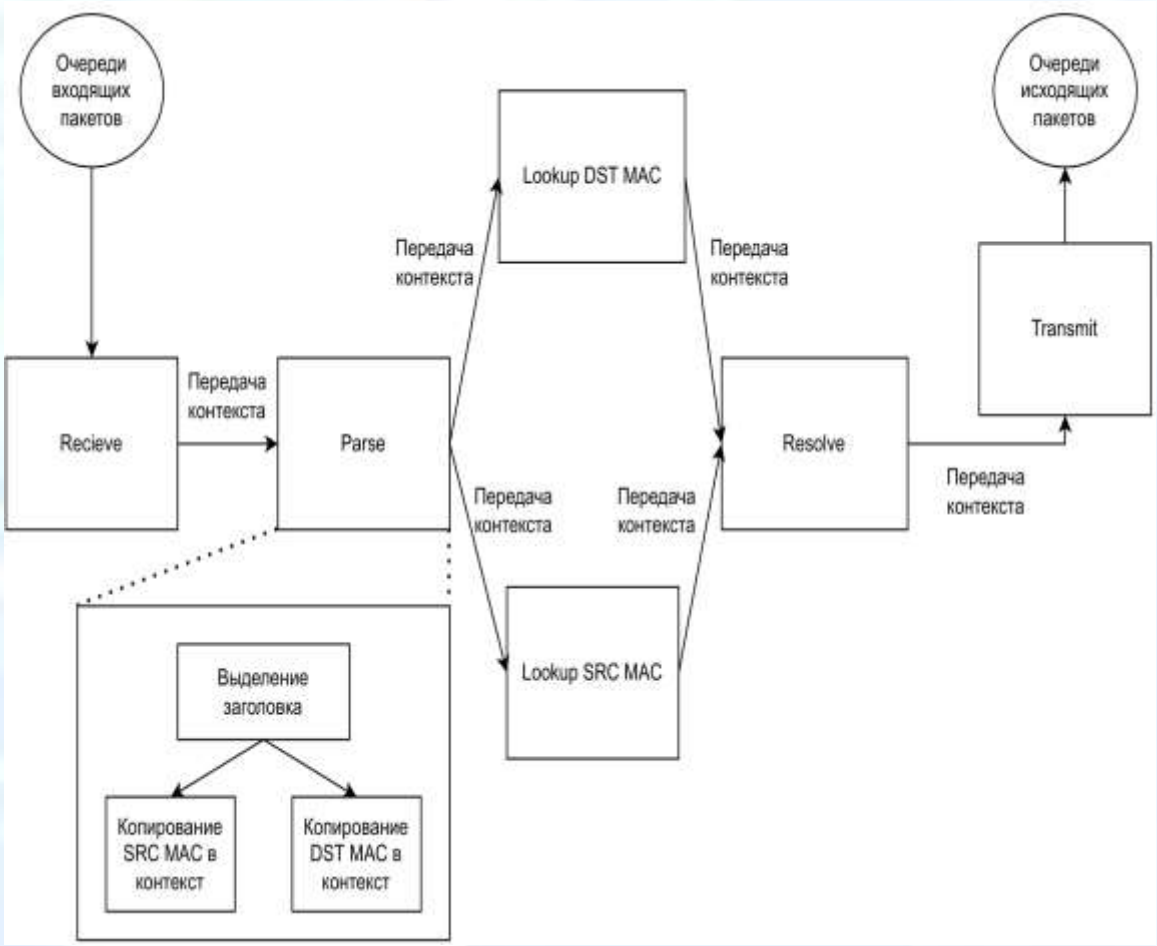
- Получение пакета через порт
- Помещение пакета во входящий буфер
- Формирование контекста пакета
- Специфичные для каждого сценария действия, включая классификацию пакета.
- Извлечение тела пакета и объединение с контекстами.
- Постановка в очереди исходящих пакетов в порты, соответствующие маске выходных портов. Реализация политик очередизации
- Коммутация
- Отправка пакетов

Классический L2 коммутатор с MAC обучением



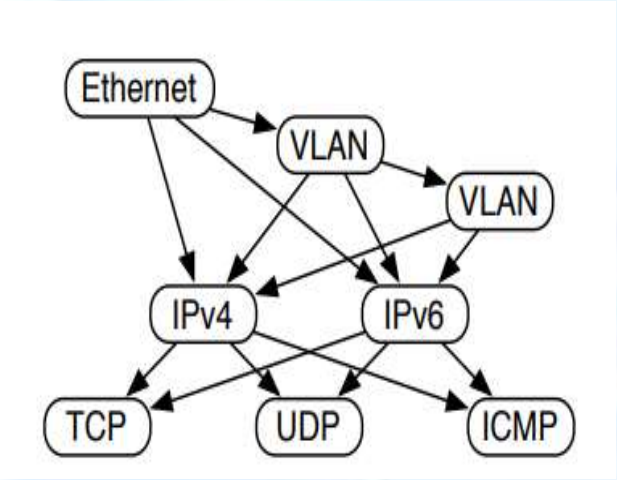
Классический L2 коммутатор с MAC обучением - стадии

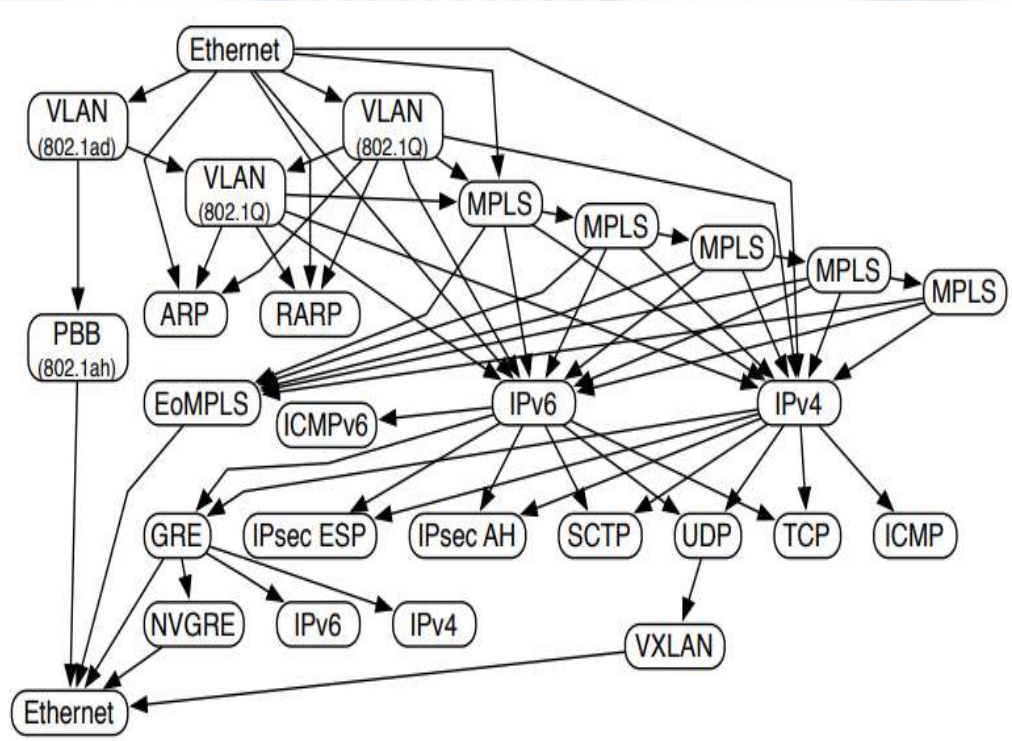
Классический L2 коммутатор с MAC обучением – возможность распараллеливания



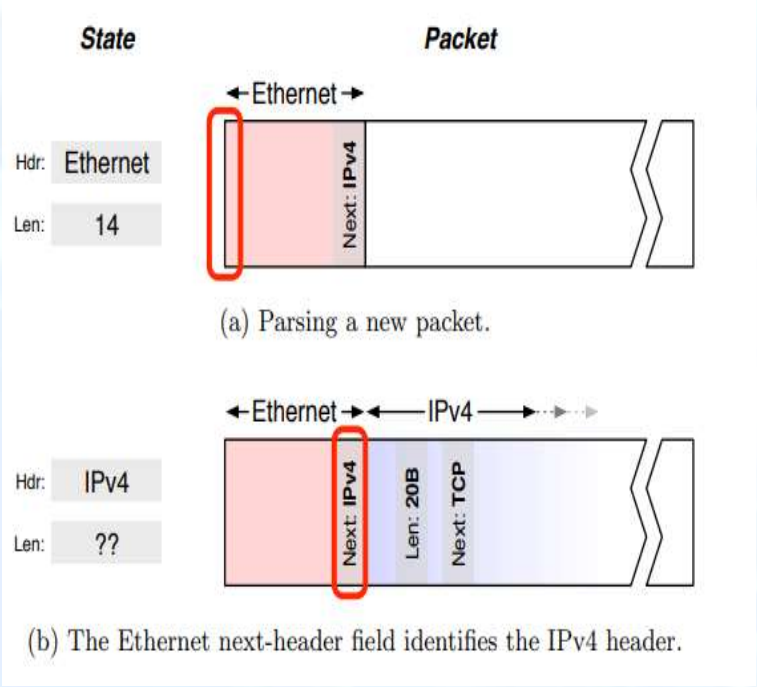
Как разобрать пакет?

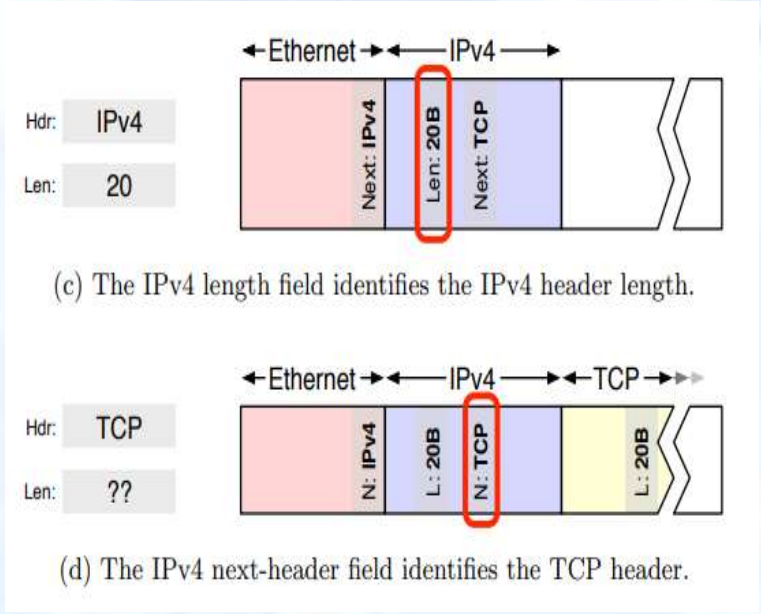
- Граф разбора



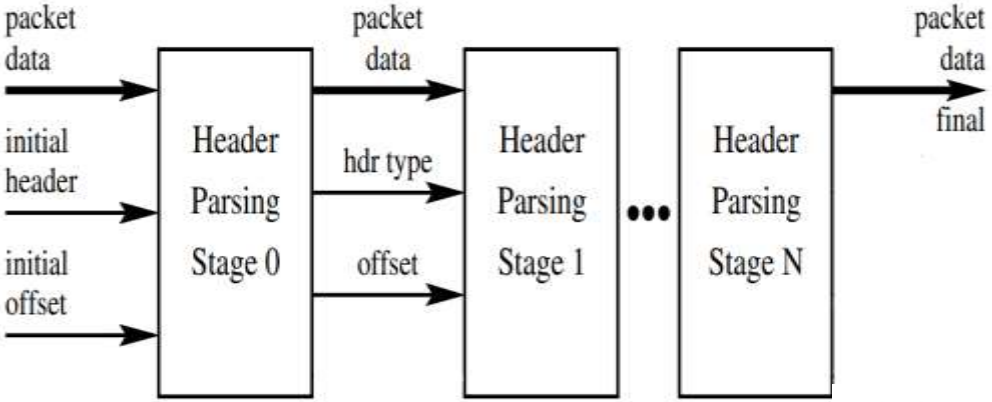


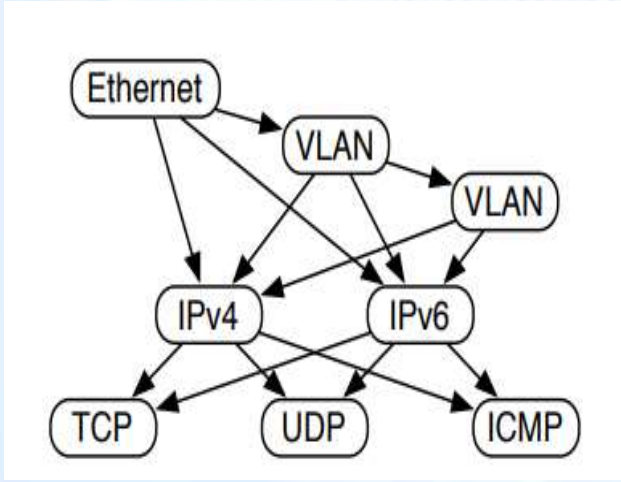
Последовательная (базовая) схема

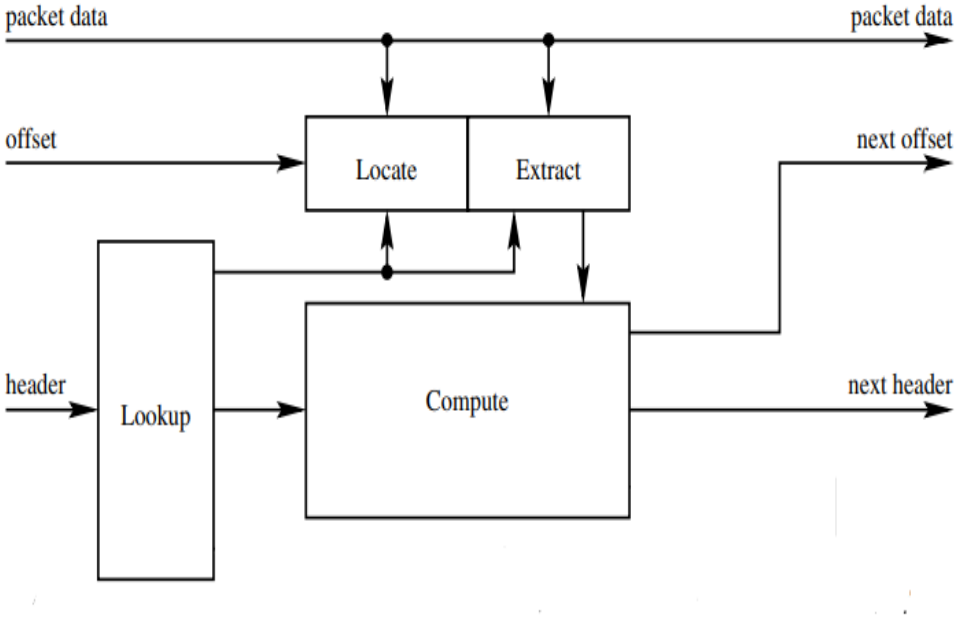




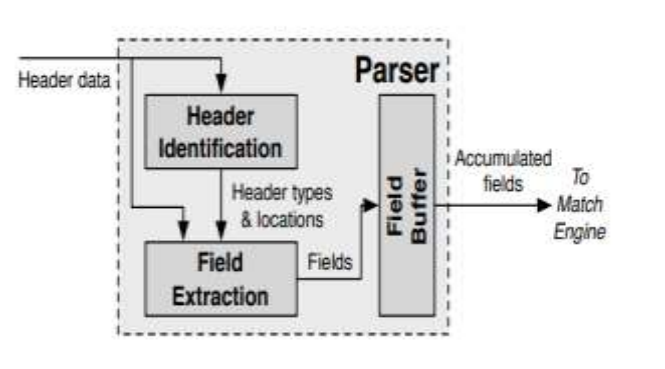
Конвейерная схема

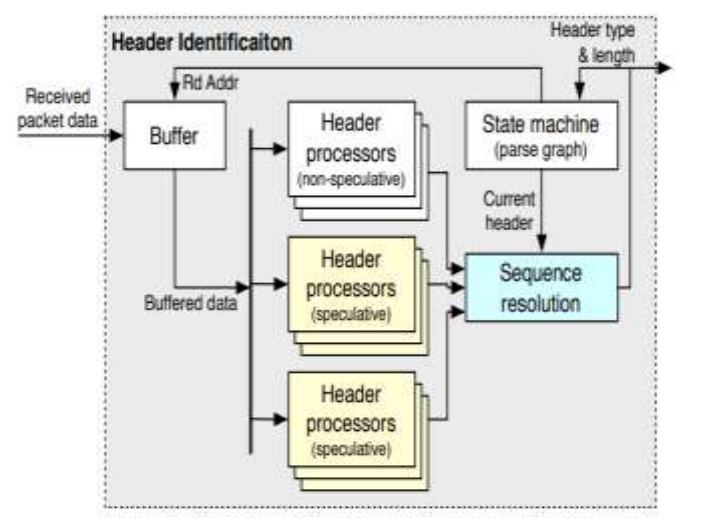


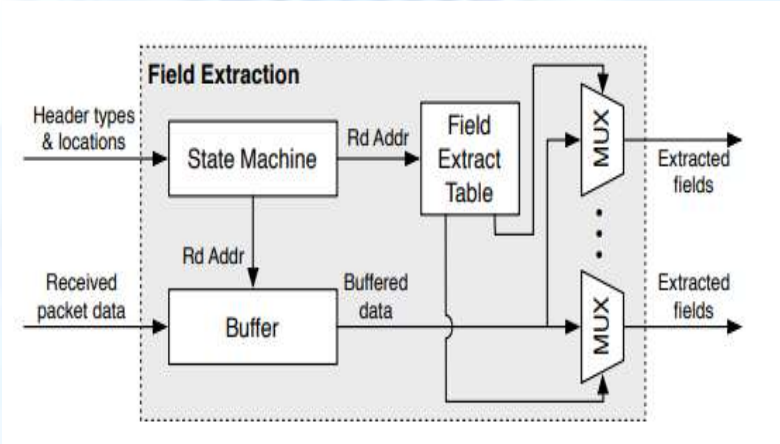




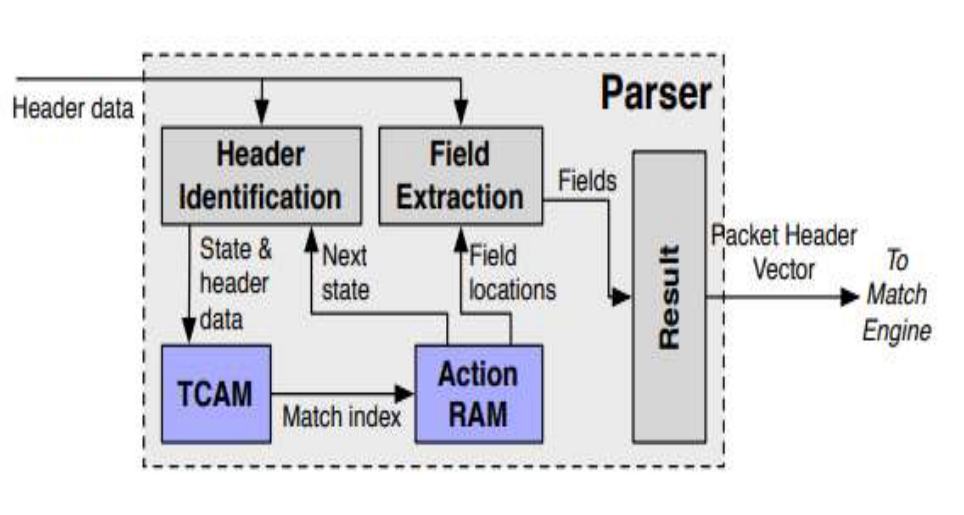
Зафиксированная модульная схема

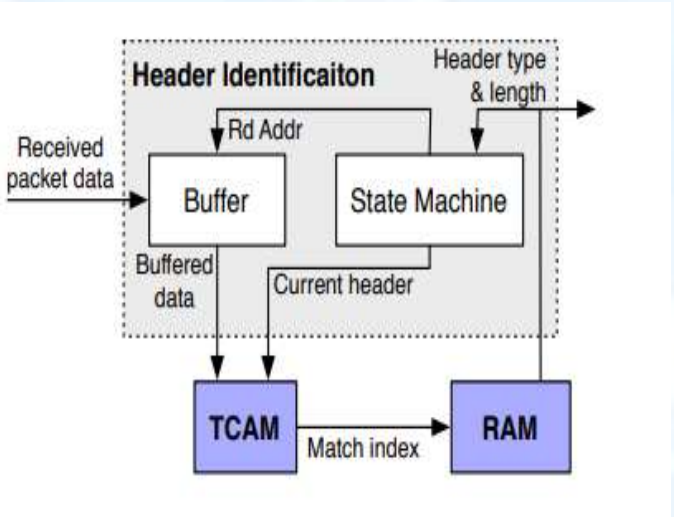




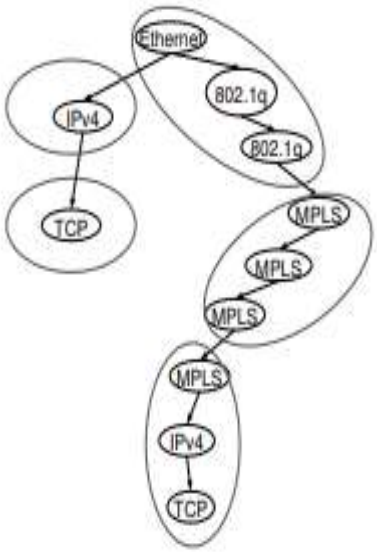


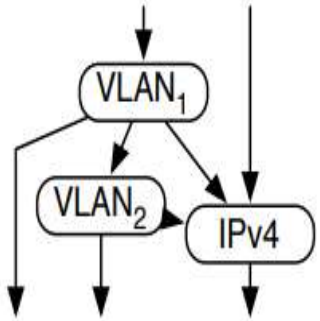
Программируемая модульная схема



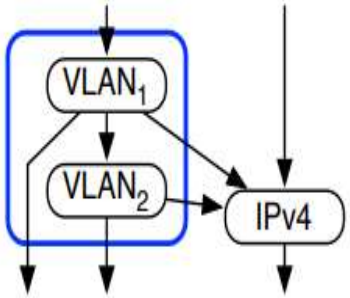


Кенгуру





Curr. Hdr.	Lookup Val.	Next Hdr.	Hdr. Len	Next lookup offset
$VLAN_1$	0x0800	IPv4	4	0
$VLAN_1$	0x8100	$VLAN_2$	4	2
$VLAN_2$	0x0800	IPv4	4	0



Curr. Hdr.	Lookup Vals.		Next Hdr.	Hdrs	Next lookup offset
$VLAN_1$	0x8100	0x0800	VLAN, IPv4	4	0, 6
$VLAN_1$	0x0800	--	IPv4	IPv4	0, 6

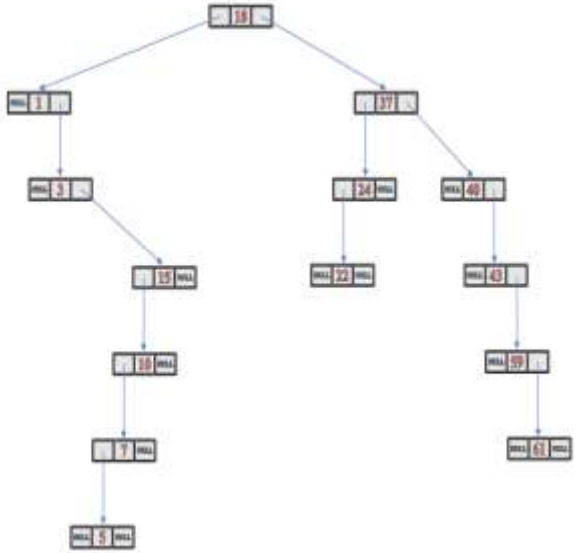
Таблицы классификации

	Название таблицы	Количество записей (согласно ТЗ)	Тип поиска	Предварительная оценка размера*, байт
1	Интерфейсов (iface)	предположительно ≤ 32	по индексу в массиве	260 (SRAM)
2	MAC–VLAN	64 000	по полному соответствию	712 000 (SRAM)
3	Маршрутизации (Forwarding Information Base, FIB)	128 000	по наибольшему префиксу	7 168 000 (SRAM), 5 000 000 (HP–TCAM+SRAM, Z–TCAM+SRAM), 4 500 000 (TCAM+SRAM)
4	ARP	32 000	по полному соответствию	416 000 (SRAM)
5	Multicast (mcast)	2 000	по полному соответствию	32 000 (SRAM)
6	Label Forwarding Informati on Base (LFIB)	64 000	по полному соответствию	1 088 000 (SRAM)
7	Egress (group)	2 000	по полному соответствию	494 000 (SRAM)
8	VLAN	4 096	по индексу в массиве	16 000 (SRAM)
9	Mirror	≤ 8 224	по индексу в массиве	33 896 (SRAM)
10	Ethertype	предположительно ≤ 10	по полному соответствию	50 (SRAM)
11	LAG	предположительно ≤ 32	по индексу в массиве	128 (SRAM)
12	Таблицы ACL (*будут добавлены на следующем этапе)*	-	-	-
	Сумма			9 960 334 (SRAM)

Варианты представления таблиц классификации в SRAM-памяти с помощью деревьев поиска

- **несбалансированное дерево бинарного поиска (базовая структура);**
- **рандомизированное дерево бинарного поиска (место вставки узла выбирается случайно);**
- **АВЛ-дерево (для каждого узла которого высоты двух его поддеревьев отличаются не более чем на один);**
- **красно-черное дерево (логически является идеально сбалансированным 2-3-4-деревом);**
- **B-дерево (у каждого узла может быть до M ≥ 2 дочерних узлов).**

- в таблицах классификации, для которых необходимо проводить поиск по полному соответствию (большинство таблиц).



Варианты представления таблиц классификации в SRAM-памяти с помощью деревьев поиска

Структура	Высота дерева	Размер памяти (в битах)	Время поиска/модификации	Хранение узлов в нескольких независимых областях памяти
Несбал. ДБП	$\leq n - 1$	$(K + V + 2P) \times N$	$O(\log_2(n))$ в среднем, $O(n)$ в худшем	неэффективно по памяти
Ранд. ДБП	$\leq n - 1$	$(K + V + 2P) \times N$	$O(\log_2(n))$ в среднем, $O(n)$ в худшем	неэффективно по памяти
АВЛ–дерево	$\leq 1.45 \times \log_2(n + 2)$	$(K + V + 2P + 2) \times N$	$O(\log_2(n))$	Неэффективно по времени модификации и размеру обновляемой области памяти при модификации
Красно–черное дерево	$\leq 2 \times \log_2(n)$	$(K + V + 2P + 1) \times N$	$O(\log_2(n))$	Возможно разбиение памяти на примерно $\log_2(n)$ областей различных размеров, при этом суммарный размер памяти увеличивается в 2-3 раза
B–дерево порядка M	$\leq \log_{M/2}(n)$	$\leq (K + V + P) \times (2N / (1 - 2 / M)) + M$	поиск: $O(\log_2(n))$ при быстрой загрузке узла из памяти, иначе $M \times O(\log_M(n))$; модификация: $M \times O(\log_M(n))$	Возможно, аналогично красно-черному дереву

Рассматриваемые СПУ

- Barefoot Tofino
- Barefoot Tofino 2
- Mellanox NP-5
- Mellanox SwitchX-2
- Huawei ENP
- Innovium Teralynx 7
- Nokia FP4
- Cisco NPU
- Juniper Q5
- Broadcom Tomahawk 3
- Broadcom Trident 3

Критерии обзора СПУ

- Год выпуска
- **Программируемость СПУ**
- **Тип СПУ**
- **Ключевые особенности архитектуры** (6 критериев)
- Характеристики кристалла (3 критерия)
- Тип интерфейса к ЦПУ
- Управляющий процессор на кристалле (если предусмотрен)
- Производительность
- Допустимые конфигурации сетевых интерфейсов
- Стоимость

23

Программируемость СПУ

- **Устройства с фиксированной функциональностью**
  - Фиксированный стек протоколов и программа обработки пакетов
- **Конфигурируемые устройства**
  - Загрузка программы обработки пакетов в рамках предопределенных протоколов передачи данных

Broadcom Tomahawk

- **Программируемые устройства**
  - Определение новых протоколов передачи данных в загружаемой программе

Barefoot Tofino, Broadcom Trident, Mellanox NP-5, Cisco NPU…

Подходы к построению коммутаторов

Коммутатор на ядрах общего назначения

Достоинства

- Гибкость настройки и модификации функциональности
- Простота внесения изменений

Недостатки

- Плохое соотношение стоимость / производительность
- При скорости выше 10 Гб/сек потеря пакетов 5-6 %
- Высокое энергопотребление

Схемы специального назначения

Достоинства:

- Наилучшее соотношение стоимость/производительность
- Возможность достижения высокой скорости обработки данных
- Низкое энергопотребление

Недостатки:

- Высокая сложность программирования сервисов
- Необходимость наличия глубокой экспертизы в разработке сетевых устройств
- Необходимость полной переделки при переходе на новый стек протоколов

Сетевые процессоры (NPU)

Достоинства:

- Гибкость в программировании новых сервисов
- Промежуточное положение по соотношению стоимость /производительность
- Низкое энергопотребление
- Возможность быстрого развития линейки устройств
- Длительное время нахождения на рынке
- Соответствие имеющемуся опыту разработки в РФ

Недостатки:

- Длительный цикл разработки

Типы программируемых СПУ

- **Многопроцессорная ИС на базе процессоров общего назначения**
  - Гибкость программирования
  - Невысокая скорость обработки пакетов
- **ASIC**
  - Аппаратная реализация основных функций СПУ (низкая гибкость программирования)
  - Высокая скорость обработки пакетов
- **Сетевой процессор**
  - Специализация к задачам обработки пакетов
  - Компромисс по возможностям программирования

Cisco NPU

Barefoot Tofino, Innovium Teralynx, Broadcom Trident, Broadcom Tomahawk

Mellanox NP-5, Huawei ENP, Juniper Q5, …

Сравнение СПУ по общим критериям

СПУ	Программируемость	Производительность	Интерфейсы	Стоимость	Стоимость коммутатора
Mellanox NP-5	+, Си	240 Гбит/с	До 100 GbE	1000$	?
Mellanox SwitchX-2	+, Си	До 2 Тбит/с	До 56 GbE	Нет данных	15000$
Huawei ENP	+	480 Гбит/с	До 100 GbE	Нет в продаже	6000$
Nokia FP4	+	2,4 Тбит/с	До 400 GbE	Нет данных	Нет данных
Barefoot Tofino	+, P4	6,5 Тбит/с	До 100 GbE	Нет данных	8000$

Сравнение характеристик кристалла СПУ

СПУ	Тех. процесс	Интерфейсы с ЦПУ
Mellanox NP-5	28 нм	PCI Express, Ethernet 1×10GbE
Mellanox SwitchX-2	16 нм	PCI Express Gen3
Huawei ENP	16 нм	Нет данных
Nokia FP4	16 нм	Нет данных
Barefoot Tofino	16 нм	4×PCI Express Gen3, 1 или более Ethernet до 100 GbE

Сравнение ключевых особенностей архитектур конвейеров СПУ

СПУ	Состав конвейера	Типы ядер СПУ
Mellanox NP-5	Конвейер из 5 функционально специализированных стадий	Специализированные векторные процессоры
Mellanox SwitchX-2	Нет данных	Нет данных
Huawei ENP	Явной структуры конвейера нет, процессоры объединены в группы, которые могут параллельно выполнять разные задачи	Процессоры со специализированными инструкциями для обработки заголовков Ethernet, IP
Nokia FP4	Нет данных	Нет данных
Barefoot Tofino	Конвейеры входной и выходной обработки, разделенные коммутационной матрицей. Функционально специализированные стадии трех типов	Специализированные процессоры для каждого из типов стадий

Организация конвейера

Два основных подхода:

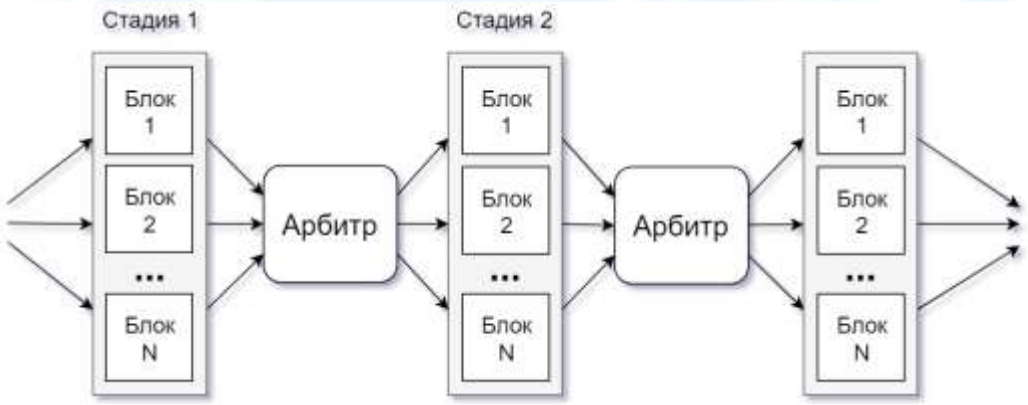
- процессорные ядра общего назначения внутри стадий — Cisco NPU
- специализация ядер к функциям обработки пакетов — Barefoot Tofino, Mellanox NP-5, Huawei ENP, …

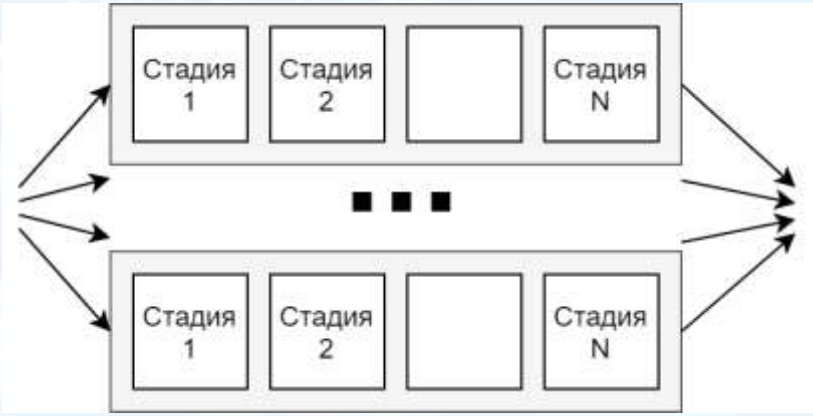
Механизм “разворота” пакетов:

- повторный проход пакета по конвейеру — Barefoot Tofino, Mellanox NP-5, Juniper Q5
- понижает пропускную способность конвейера

Параллелизм СПУ

- Параллелизм на уровне стадий конвейера
- Параллелизм конвейеров
- Комбинированные подходы





Память СПУ

Стратегии размещения данных:

- тела пакетов – внешняя память,
- таблицы классификации – внутренняя память

Broadcom Trident, Broadcom Tomahawk, Barefoot Tofino, Mellanox NP-5

- все данные во внешней памяти

Huawei ENP, Juniper Q5

Память тел пакетов: DDR SDRAM, RL DRAM

Память таблиц классификации: SRAM, TCAM

Основные тенденции

- Программируемость разные производители понимают по-разному
- Наибольшая производительность у устройств ASIC
- Принципы построения конвейеров:
  - разделение на 2 части (ingress, egress);
  - коммутационная матрица и репликатор пакетов между частями конвейера;
  - функциональная специализация стадий;
  - масштабируемая архитектура из однотипных конвейеров.
- Размещение тел пакетов во внешней памяти, таблиц классификации – во внутренней

Спасибо за внимание!