Качество обслуживания

Качество обслуживания - QoS

- Требования разных типов приложений
- Параметры качества обслуживания
- Служба QoS
  - Модель службы QoS
  - Алгоритмы управления очередями
  - Механизмы профилирования и формирования трафика
- Общая характеристика протоколов QoS IP
- Резервирование пропускной способности с помощью RSVP
- Дифференцированное обслуживание DiffServ
- Использование виртуальных каналов MPLS для поддержки QoS
- Качество обслуживания на основе централизованной политики (Policy-based QoS)
  - Общая структура
  - COPS

Требования к сети различных типов трафика

Требования к пропускной способности

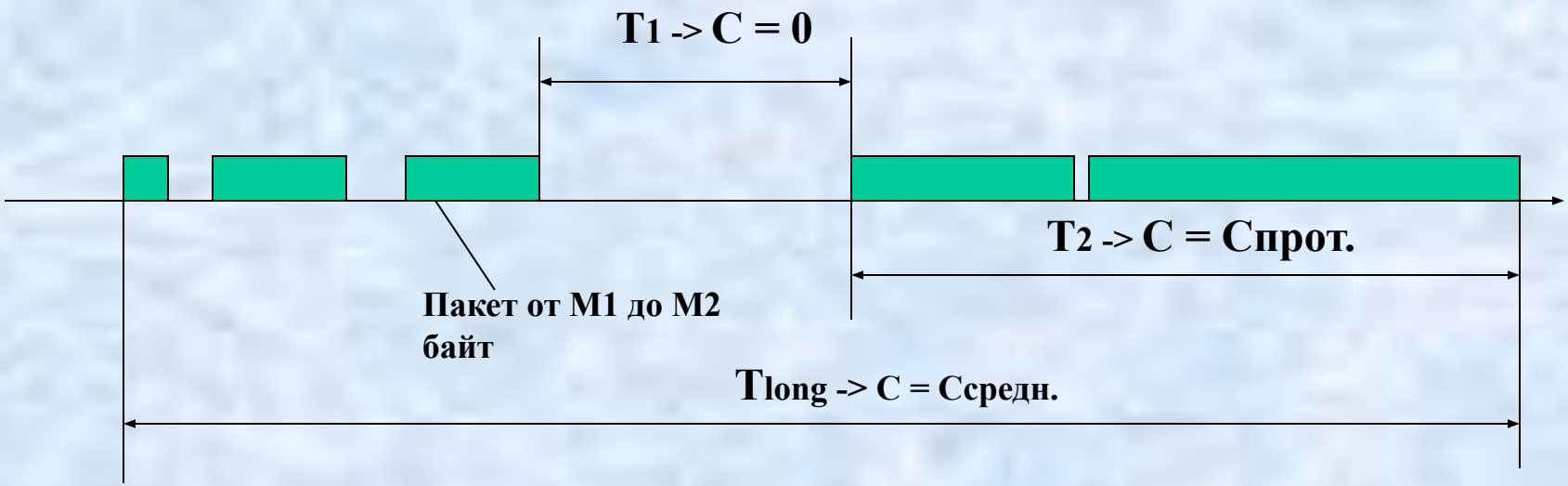
Тип трафика	Описание
Поток (Streem)	Предсказуемая доставка со сравнительно постоянной битовой скоростью (CBR)
Пульсация (Burst)	Непредсказуемая доставка «блоков данных» битовой переменной скоростью (VBR) (нет верхней границы).

Поток (Stream)

Трафик Constant Bit Rate = M / T бит/с

Примеры: оцифрованный голос, цифровое видео, телеметрическая информация

Пульсация (Burst)



Трафик Variable Bit Rate

Скорость меняется от 0 до Спротокола

Пульсация - период T2

Измеряется в:

Сек - длительность пульсации

Байтах (burst size) - объем данных в импульсе

Коэффициент пульсации = C_T /C средн. (например, 50:1)

Примеры: передача файлов, компрессированные голос и видео

Параметры QoS по пропускной способности сети

4Средняя скорость на длительном периоде

- Commited Information Rate у frame relay
- Sustained Cell Rate у ATM

4Максимальная скорость всплеска (пульсации)

- Peak Cell Rate у ATM

4Максимальный объем пульсации

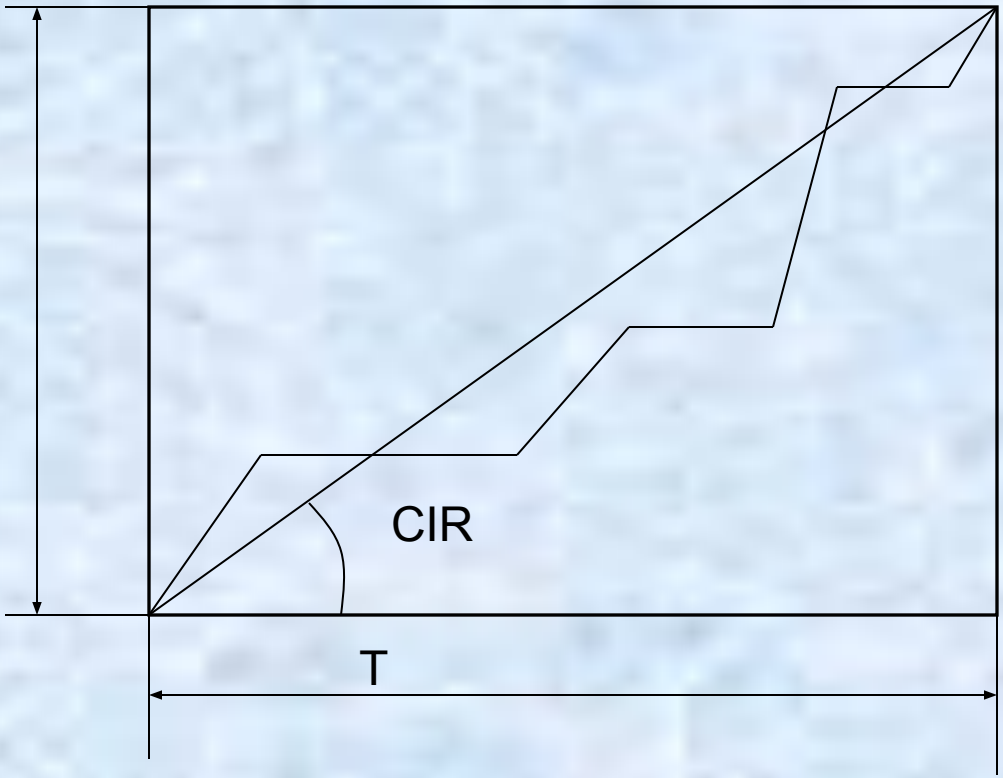
- Bc (Burst commited) у frame relay
- Maximum Burst Size (MBS) у ATM

4Максимальное время пульсации

- T пульсации у frame relay
- Burst Tolerance (BT) у ATM

Взаимосвязь параметров пульсации

Frame relay



Bc = CIR x T

ATM:

BT = (MBS-1) (1/SCR - 1/PCR)

Характеристика приложений в терминах чувствительности к задержкам передачи данных

Терпимость к задержкам	Тип доставки	Описание	Пример
Высокая	Асинхронный	Нет ограничений на время доставки («Эластичный»)	Электронная почта
	Синхронный	Данные чувствительны к задержкам, но допускают их	Компрессированный голос
	Интерактивный	Задержки могут быть замечены пользователями, но они не сказываются негативно на функциональности приложений.	Редактирование удаленного файла
	Изохронный	Имеется порог чувствительности к задержкам, при превышении которого снижается функциональность приложения	Некомпрессированные голос или видео
Низкая	Критически важный	Задержка доставки данных сводит к нулю функциональность.	Управление технологическим объектом

Параметры QoS по задержкам:

- средняя задержка (delay)

- вариация задержки (jitter)

$$Delay = \Sigma(ti)/N$$ – математическое ожидание

$$Jitter = \frac{1/N \cdot \sqrt{\Sigma(ti - delay)^2}}{Delay}$$ -коэффициент вариации

Чувствительность приложений к потерям данных

- Чувствительные к потерям приложения

  Передача дискретных данных - текст, числа, неподвижные изображения -

  при потере пакета данные становятся частично или полностью обесцененными - необходима повторная передача

- Устойчивые к потерям приложения

  Передача аналоговой информации - голос,видео - инерционность процессов позволяет при небольшом проценте потерь восстановить потерянные данные по соседним

Параметры QoS по уровню потерь данных

Процент потерянных пакетов (кадров, ячеек)

- Cell Lost Ratio в ATM

Процент искаженных кадров

Качество обслуживания в сетях с коммутацией каналов

Основной показатель Grade of Service:

вероятность отказа сети в установлении соединения, блокировка (причина - исчерпана коммутационная емкость какого-либо коммутатора вдоль пути)

Пропускная способность и задержки – фиксированные

Формула Эрланга:

$$Pb = \frac{\frac{A^N}{N!}}{\sum \frac{A^X}{X!}}$$

Пример:

A = 3

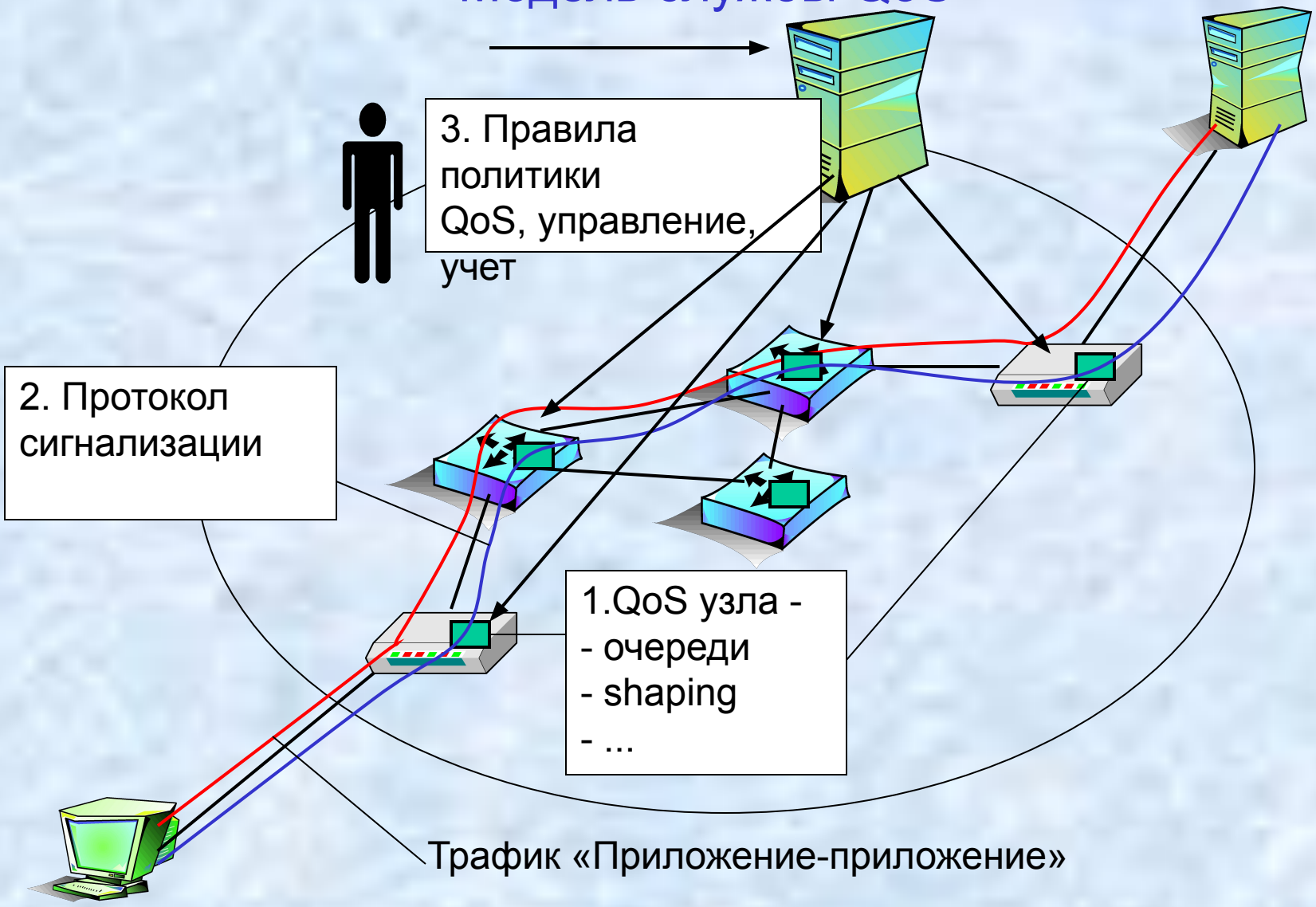
N = 6

Pb = 0.0522 (5%)

A – нагрузка в эрлангах (занятость одной линии)

N – коммутационная емкость (максимальное количество соединений)

Модель службы QoS



Средства QoS узла

1. Механизмы обслуживания очередей:
   - FIFO (первым пришел - первым ушел)
   - Priority – приоритетное обслуживание
   -WFQ – взвешенное обслуживание
   - …
2. Механизмы «кондиционирования» трафика
   - классификация
   - контроль доступа
   - профилирование (policing)
   - формирование (shaping)

Кондиционирование трафика (conditioning)

1. Классификация (classification) трафика на основе:

- IP-адресов Dest и Source
- Протоколу - TCP или UDP
- TCP/UDP ports (по приложениям)
- Метка потока в IPv6
- Признаков в заголовке прикладного протокола
- Имени пользователя

Пример

Класс 2:

- IP Dest = 132.35.100.0/24
- IP Source = 26.0.0.0/8
- TCP/UPD = 80

2. Контроль доступа (Access Control)

- имеет ли право пакет от определенного пользователя обслуживаться в данное время и данной входной точкой сети

Проблема: как определить соответствие «пакет - пользователь»?

При аутентификации пользователя его имя связывают с IP-адресом

Пример: система Meta IP компании Check Point

3.Профилирование (policing)

Проверка соответствия трафика QoS-профилю –
проверка соглашения Service Level Agreement

Профиль:

- средняя скорость
- максимальная скорость
- пульсация
- задержка

При несоответствии пакета профилю - пакет отбрасывается ли помечается как «нарушитель» - его могут отбросить последующие сетевые устройства при перегрузках

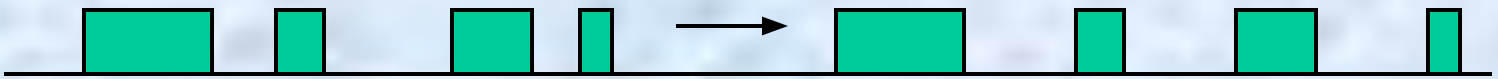
Формирование трафика

(shaping)

Придание потоку пакетов заданных временных характеристик

- равномерность

При равномерном следовании пакетов уменьшаются очереди в маршрутизаторах и, соответственно, времена задержек

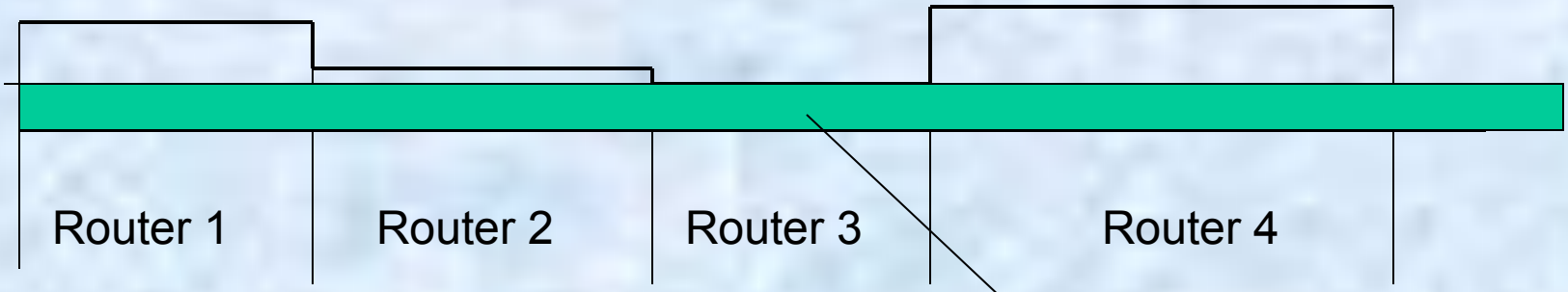


Протоколы сигнализации для QoS

Нужны для распространения вдоль пути следования пакетов данных о требуемых параметрах QoS для трафика

Согласуют усилия сетевых устройств по согласованному обслуживанию определенного потока данных

Без согласованности параметры QoS поддержать нельзя!



Примеры: RSVP, DS-байт

Пропускная способность равна минимальной по всем сетевым устройствам

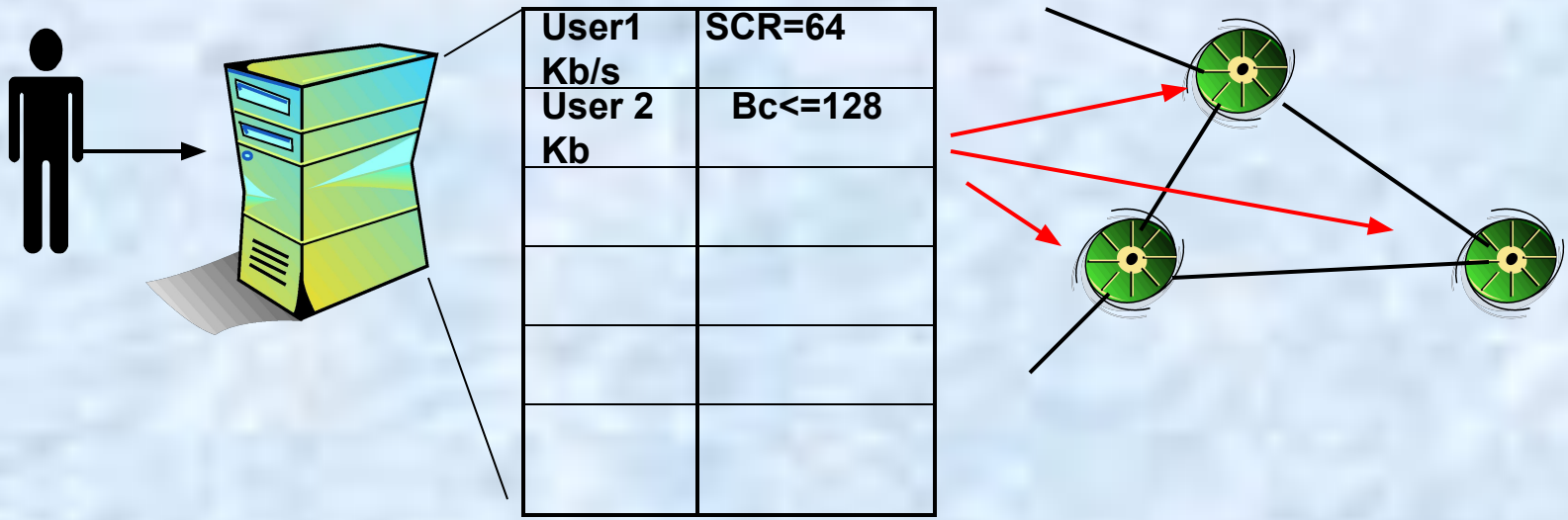
Централизованная политика, управление и учет

Администратор выполняет роль арбитра для пользователей и приложений:

- задает с помощью набора правил условия

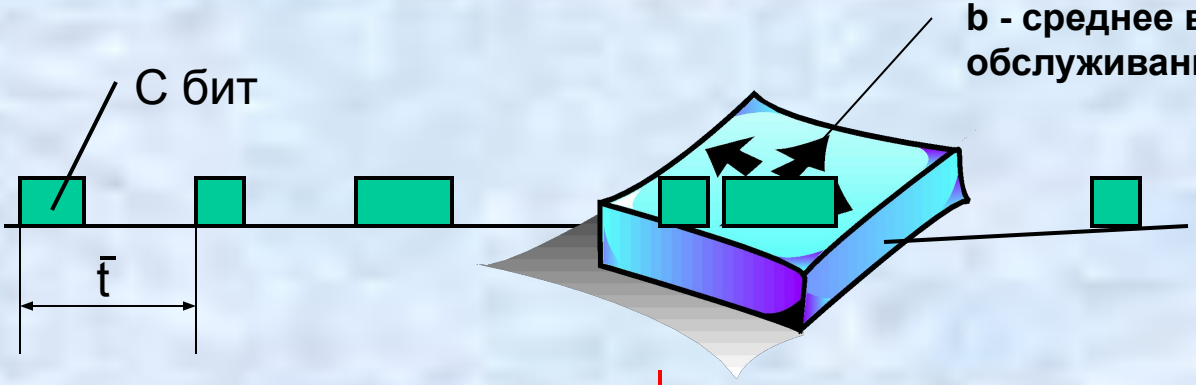
  кому и когда сетевые устройства должны предоставлять услуги QoS и с какими параметрами

Другой способ согласования параметров QoS между устройствами



Алгоритмы управления очередями

Применение методов теории массового обслуживания (Queuing Theory) для анализа очередей в сетях



Модель M|M|1

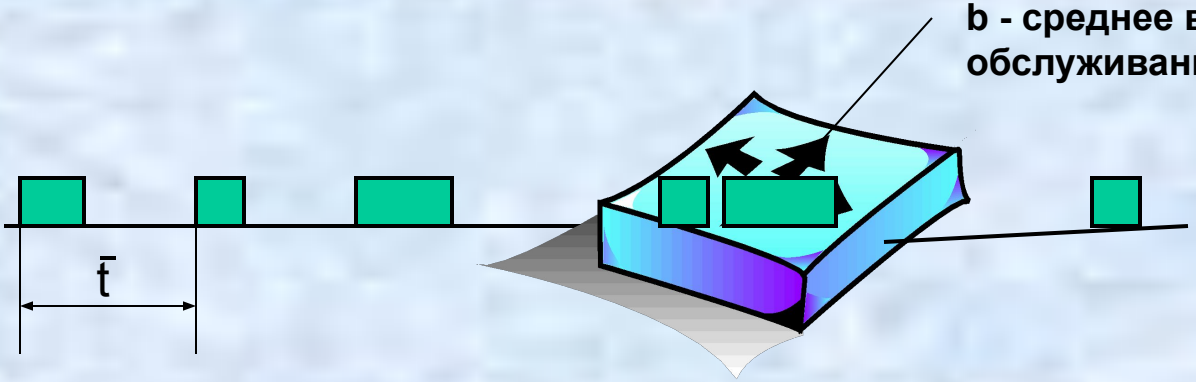
$\lambda = 1/\bar{t}$ - интенсивность поступления заявок-пакетов в обслуживающий прибор,
скорость поступления данных λ x C

$\mu = 1/b$ - интенсивность выхода заявок-пакетов из обслуживающего прибора,
b - среднее время продвижения пакета

$\rho = \lambda/\mu$ - коэффициент загрузки обсл. прибора

Алгоритмы управления очередями

Применение методов теории массового обслуживания (Queuing Theory) для анализа очередей в сетях



При экспоненциальном распределении времен поступления пакетов

$A(t)=1-e^{-\lambda t}$ **- среднее время между пакетами = 1/ λ, коэфф. вар. = 1**

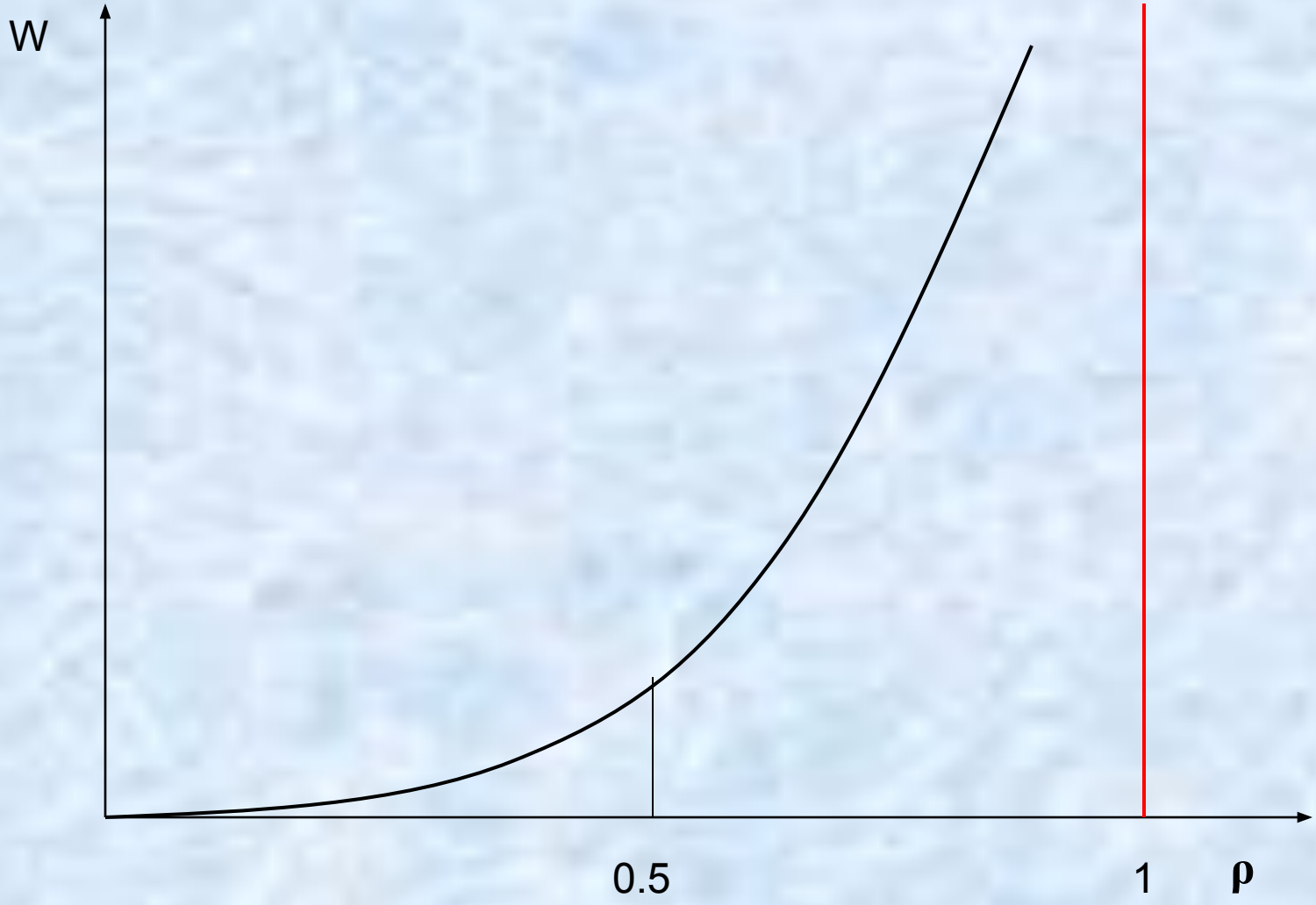
и экспоненциальном распределении времени обслуживания

$B(x)=1-e^{-\mu x}$

среднее время ожидания **W** равно

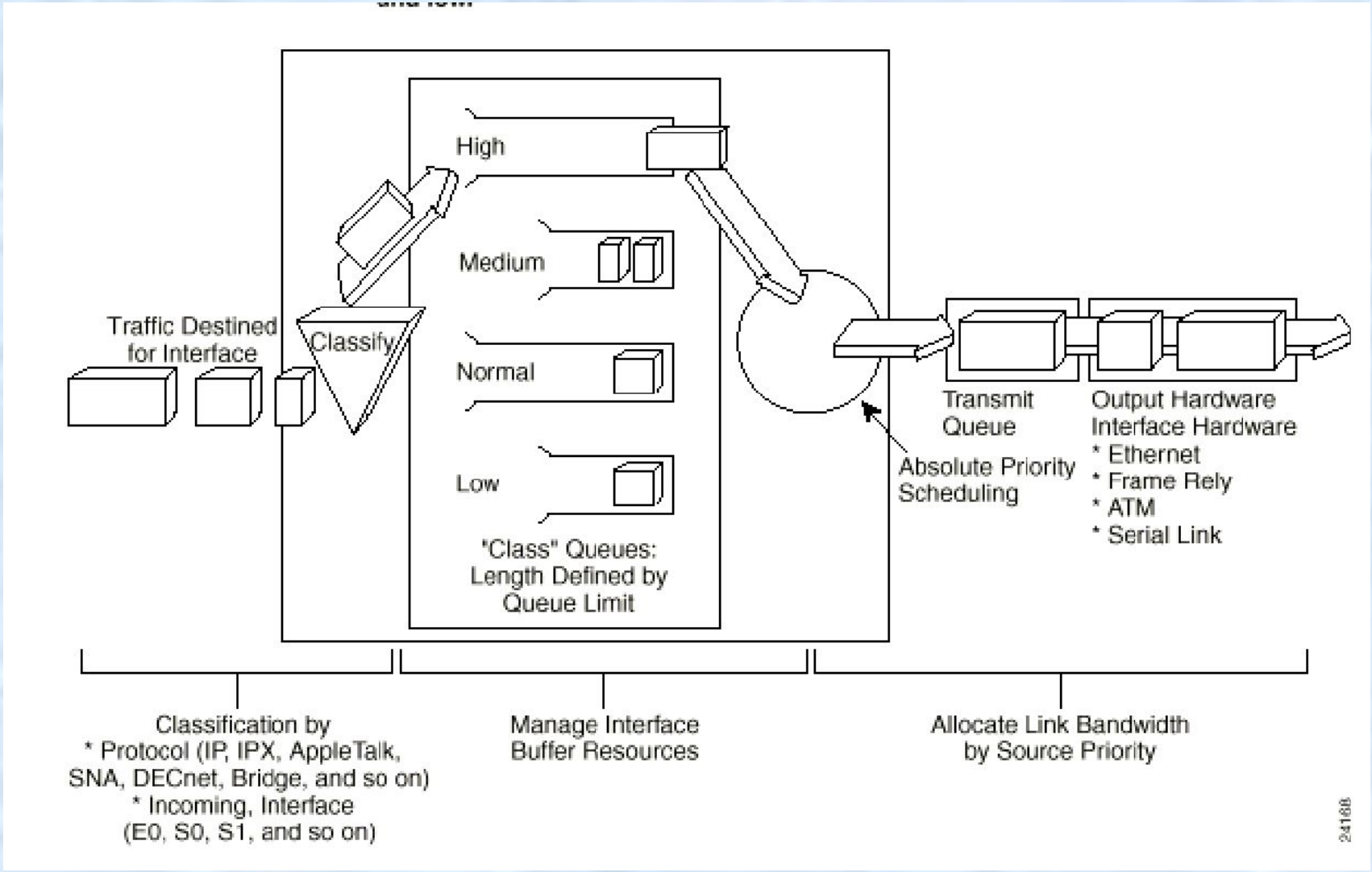
$$W = \rho b/(1-\rho)$$

Среднее время ожидания



При $\rho < 0.5$ задержки незначительны - низкая загрузка сети гарантирует качество обслуживания!

Приоритетное обслуживание очередей



Абсолютный приоритет - пока высокоприоритетная очередь полностью не обслужена, более низкоприоритетные не обслуживаются

Время ожидания в низкоприоритетной очереди может стремится к ∞

Приоритетное обслуживание очередей

High priority

$W = \rho_H b/(1-\rho_H)$

$\rho_H = \lambda_H / \mu$ - например, 200/1000 или 0.2

$\rho_M = \lambda_M / (\mu - \lambda_H)$ - например, 200/(1000 - 200) = 0.4

При равной интенсивности поступления условия обслуживания трафика с приоритетом Medium хуже:

W_H = (0.2 / 1000) / (1 - 0.2) = 0.0002 /0.8 = 0.00025 = 25 мкс

W_M = (0.4 / (1000-200)) / (1 - 0.4) + W_H= 0.0005 /0.6 + 25= 108 мкс

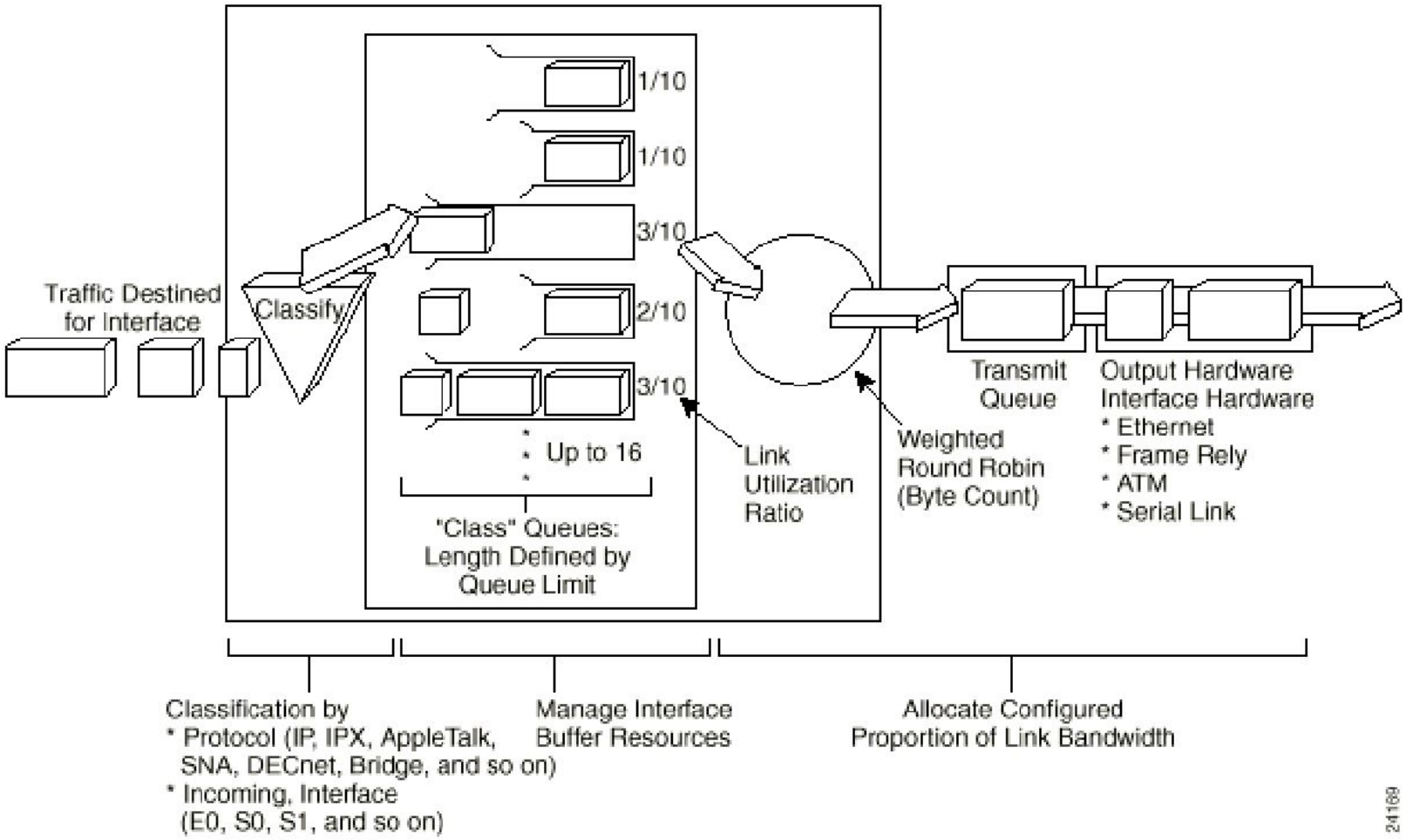
При значительной доле трафика High Priority остальной трафик обслуживается со значительными задержками

Конфигурирование приоритетного обслуживания

Определение списка приоритетов

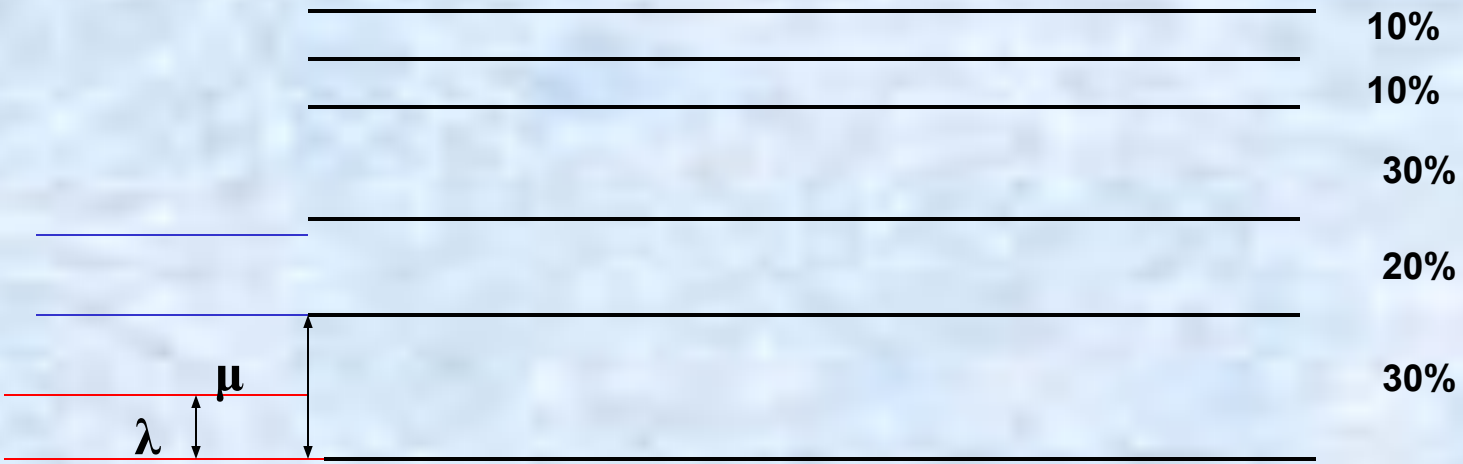
```
priority-list 4 protocol decnet medium lt 200
priority-list 4 protocol ip medium tcp 23
priority-list 4 protocol ip medium udp 53
priority-list 4 protocol ip high

interface serial 0
priority-group 4
```

Взвешенные настраиваемые очереди - Weighted Custom Queuing



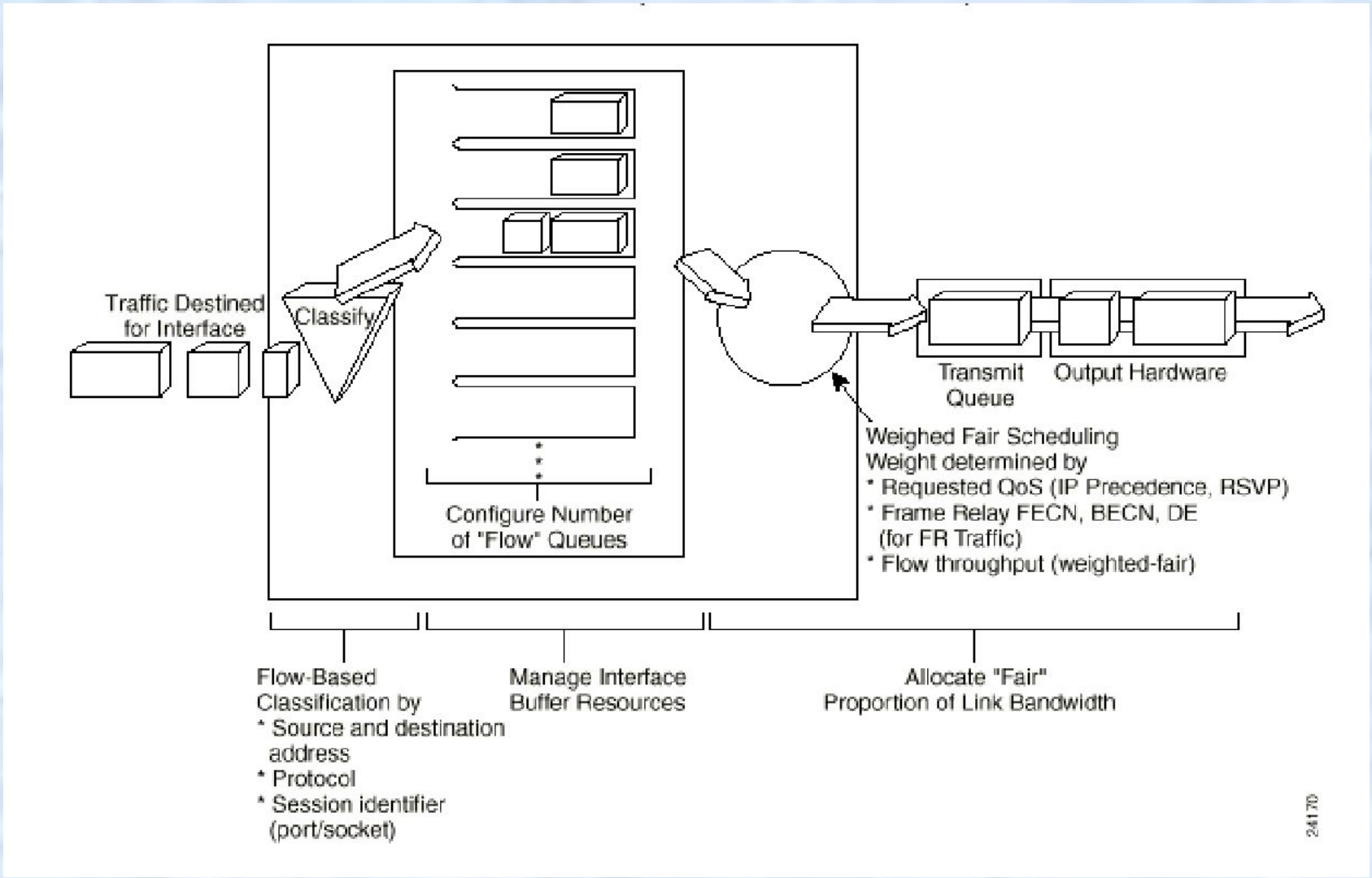
Каждая очередь обслуживается в течение заданной доли времени обработчика очереди - 10%, 10%, 30%, 20%, 30%

Взвешенные настраиваемые очереди - Weighted Custom Queuing



При взвешенном обслуживании задержки меньше у того класса трафика, у которого отношение λ / μ меньше

Взвешенное справедливое обслуживание - Weighted Fair Queuing



Равные веса для всех потоков

Существует одна приоритетная очередь - для системных сообщений (ICMP, SNMP)