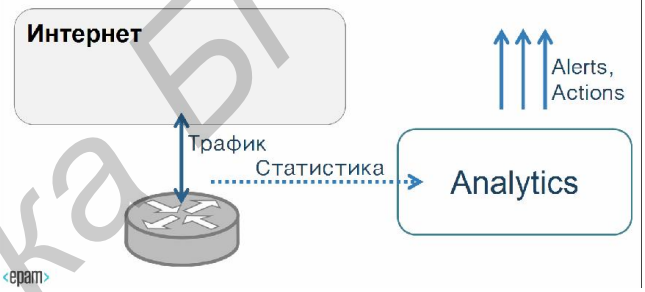
ИСПОЛЬЗОВАНИЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБЕСПЕЧЕНИЯ БЕЗОПАСНОСТИ СЕТИ

Д.В. Орехов
Руководитель группы в EPAM Systems, Республика Беларусь

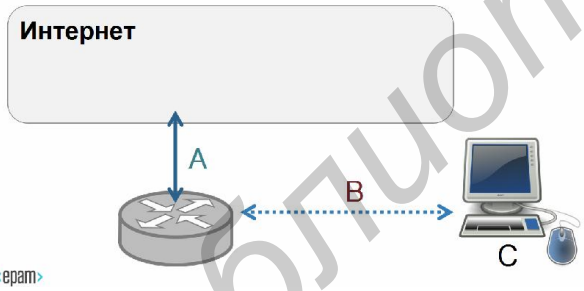
EPAM Systems, Республика Беларусь

Сначала данные надо собрать

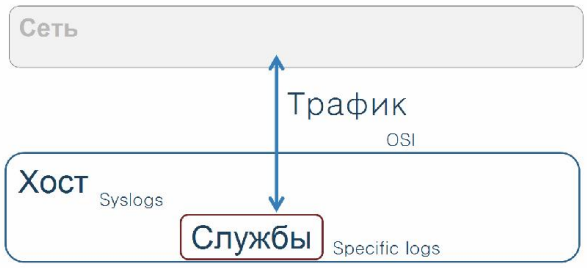
Поток данных: общая топология



Сетевые Сенсоры: Где?

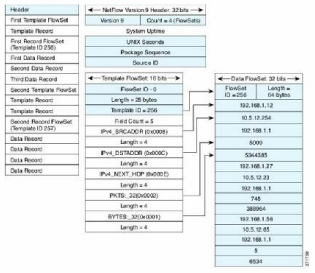


Сетевые сенсоры: Что?



Транспортировка: NetFlow

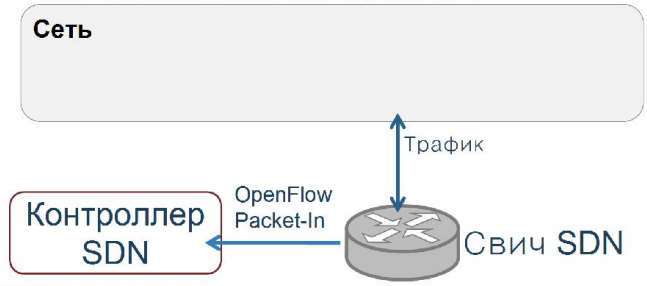
- NetFlow – стандарт для сбора и транспортировки телеметрии и статистической информации о трафике, разработан Cisco Systems, изначально для целей биллинга
- Основа NetFlow – концепция 'flows' структур, каждая из которых хранит информацию об отдельной TCP сессии
- Много открытых решений



Транспортировка: Farsight Nmsg

- Разработана компанией Farsight Security для транспортировки сетевых пакетов, в частности – DNS, до места хранения и анализа
- Скоростная и компактная
- Поддерживает двоичное и текстовое представление, а также конвертацию в pcap
- Исходный код открыт

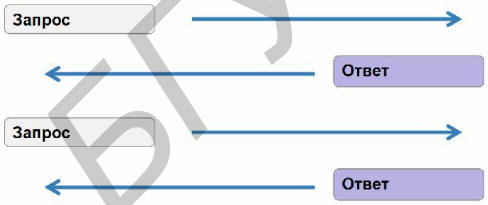
Транспортировка: OpenFlow



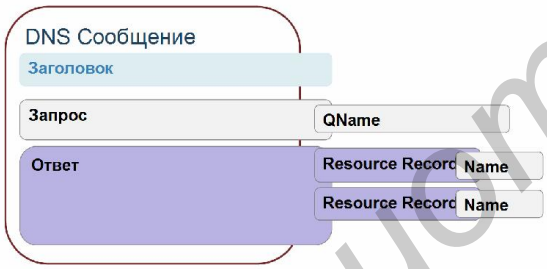
Пример из повседневной практики

DNS Tunneling

DNS



DNS: Формат сообщения



И что же такое мы имеем?

Фактически, текстовый трафик, который не фильтруется брандмауэрами

Кодировка Base32

0	A	8	I	16	Q	24	Y
1	B	9	J	17	R	25	Z
2	C	10	K	18	S	26	2
3	D	11	L	19	T	27	3
4	E	12	M	20	U	28	4
5	F	13	N	21	V	29	5
6	G	14	O	22	W	30	6
7	H	15	P	23	X	31	7

И как это отловить?

Анализ содержимого пакетов

- Размер запросов и ответов
- Энтропия доменных имён
- Статистический анализ
- Необычные типы пакетов

Анализ трафика

- Размер DNS-трафика на IP-адрес
- Размер DNS-трафика на домен
- Количество имён хостов в домене
- Местоположение DNS-сервера
- История запросов/ответов
- «Провисающие» DNS-запросы

То есть, для эффективного анализа…

…нам нужен **deep packet inspection?**

И одновременно

…нам нельзя терять скорость

Решение – смешанная модель обработки данных

Online model

Online Model - Алгоритмы

Главное свойство – инкрементальность

Обнаружение выбросов **(Outlier Detection)**

- Отклонение от медианы (Median Absolute Deviation, MAD)
- Отклонение от среднего значения (Standard Deviation from average)
- Отклонение от скользящего среднего (Standard Deviation from Moving Average)

Потоковая классификация

Инкрементальные деревья решений (Incremental decision tree)

- Hoeffding Tree (VFDT)
- Half-Space Trees

Offline model

Offline Model - Алгоритмы

Анализ данных за фиксированный период

Гипотезы **(Hypothesis tests)**

- Простое обнаружение выбросов за период времени
- Статистический критерий
- Тест Колмогорова-Смирнова

Деревья принятия решений **(Decision Trees)**

Модель авторегрессии — скользящего среднего **(Auto Regressive (AR) Moving Average (MA))**

Вариант архитектуры