



Проектирование высоконагруженных систем

Лекция №2



Быков Александр

Характеристики линии связи

- **Пропускная способность**
Измеряется в битах в секунду (бит/с)
и/или в пакетах в секунду (pps)
- **Задержка передачи**
Измеряется как Round-Trip Time (мс) + отклонение
- **Максимальный размер пакета (MTU)**
Предельный размер пропускаемый оборудованием
- **Интенсивность ошибок**
Вероятность потери пакетов

Пропускная способность

- **Сетевая карта 1Гбит/с (медь)**
Реальная пропускная способность 800-900 Мбит/с
- **Сетевая карта 10 Гбит/с (медь)**
Номинальная пропускная способность 1/2,5/5/10 Гбит/с
- **Сетевая карта 10 Гбит/с (оптика)**
Требуются дорогие оптические SFP+ коннекторы
Альтернатива: китайский коннектор + патчинг ядра

Объединение карт в Bonding

Объединение нескольких физических интерфейсов в один логический интерфейс (агрегация).

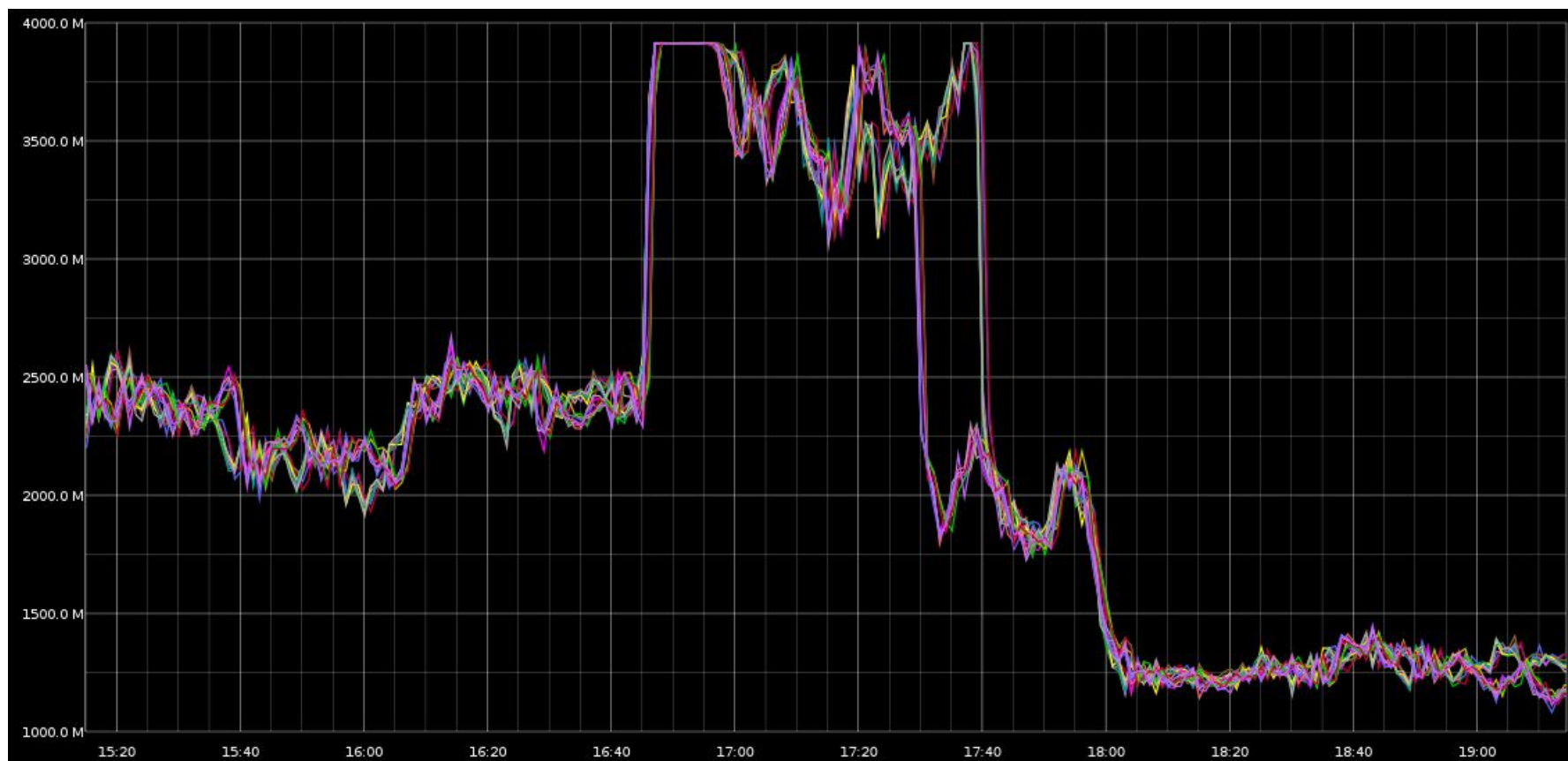
Особенности

- Используется хеширование (src, dst) гарантирующее последовательную доставку данных одного соединения

Примеры

- 4 медных интерфейса – 3.6 Гбит/с
- 2 оптических интерфейса – 20 Гбит/с

r.mradx.net - исходящий трафик («полка»)



Сетевой стек ОС

- **Многопоточные сетевые карты (MSI-X)**
Имеют несколько очередей
Позволяют разнести обработку на несколько CPU
- **Обычная производительность**
150 000 – 250 000 PPS
- **Предельная производительность (с тюнингом)**
до 1 000 000 PPS

HighLoad. Лекция №2

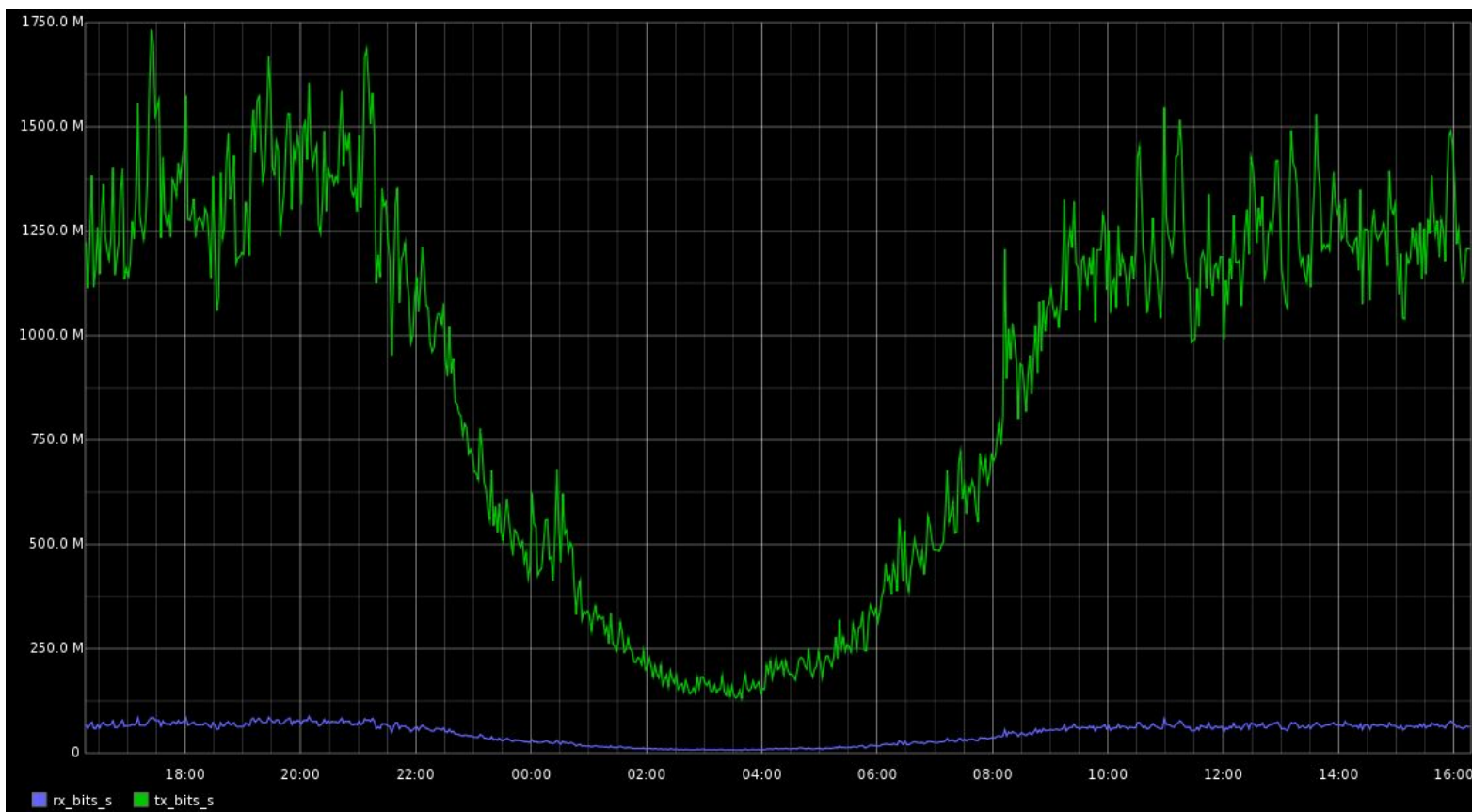


CPU0	CPU1	CPU2	CPU3	CPU4	CPU5	CPU6	CPU7	CPU8	CPU9	CPU10	CPU11	CPU12	CPU13	CPU14	CPU15		
eth0																	
58: 1434978947		0	0	0	0	0	0	0	0	0	0	0	0	0	0	449492	PCI-MSI-edge
eth0-TxRx-0																	
59: 1950517087		0	0	0	0	0	0	0	0	0	0	0	0	0	590577	0	PCI-MSI-edge
eth0-TxRx-1																	
60: 1823482036		0	0	0	0	0	0	0	0	0	0	0	0	497334	0	0	PCI-MSI-edge
eth0-TxRx-2																	
61: 2251074311		0	0	0	0	0	0	0	0	0	0	0	738795	0	0	0	PCI-MSI-edge
eth0-TxRx-3																	
62: 765575858		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth0-TxRx-4																	
63: 329477757		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth0-TxRx-5																	
64: 1517450037		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth0-TxRx-6																	
65: 938808932		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth0-TxRx-7																	
66: 1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth1																	
67: 1645802435		0	0	0	0	0	0	0	0	0	0	482782	0	0	0	0	PCI-MSI-edge
eth1-TxRx-0																	
68: 1688597500		0	0	0	0	0	0	0	0	0	524149	0	0	0	0	0	PCI-MSI-edge
eth1-TxRx-1																	
69: 1805458972		0	0	0	0	0	0	0	0	557581	0	0	0	0	0	0	PCI-MSI-edge
eth1-TxRx-2																	
70: 2068996833		0	0	0	0	0	0	0	621037	0	0	0	0	0	0	0	PCI-MSI-edge
eth1-TxRx-3																	
71: 550403216		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth1-TxRx-4																	
72: 559352604		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth1-TxRx-5																	
73: 1553486445		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth1-TxRx-6																	
74: 1060127510		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth1-TxRx-7																	
75: 1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth2																	
76: 1632663556		0	0	0	0	0	0	587664	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth2-TxRx-0																	
77: 1932803994		0	0	0	0	0	598329	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth2-TxRx-1																	
78: 1890863328		0	0	0	0	559295	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth2-TxRx-2																	
79: 2135731441		0	0	0	630233	0	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth2-TxRx-3																	
80: 564846310		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth2-TxRx-4																	
81: 586755774		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth2-TxRx-5																	
82: 1608909743		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth2-TxRx-6																	
83: 831009992		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge
eth2-TxRx-7																	
84: 1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	PCI-MSI-edge

Полезные утилиты

- **iperf**
Позволяет создавать нагрузку нужной интенсивности TCP и UDP
- **netstat -s**
Статистика по интерфейсам
- **mpstat -P ALL**
Статистика по использованию CPU
- **cat /proc/interrupts**
Распределение обработки по CPU

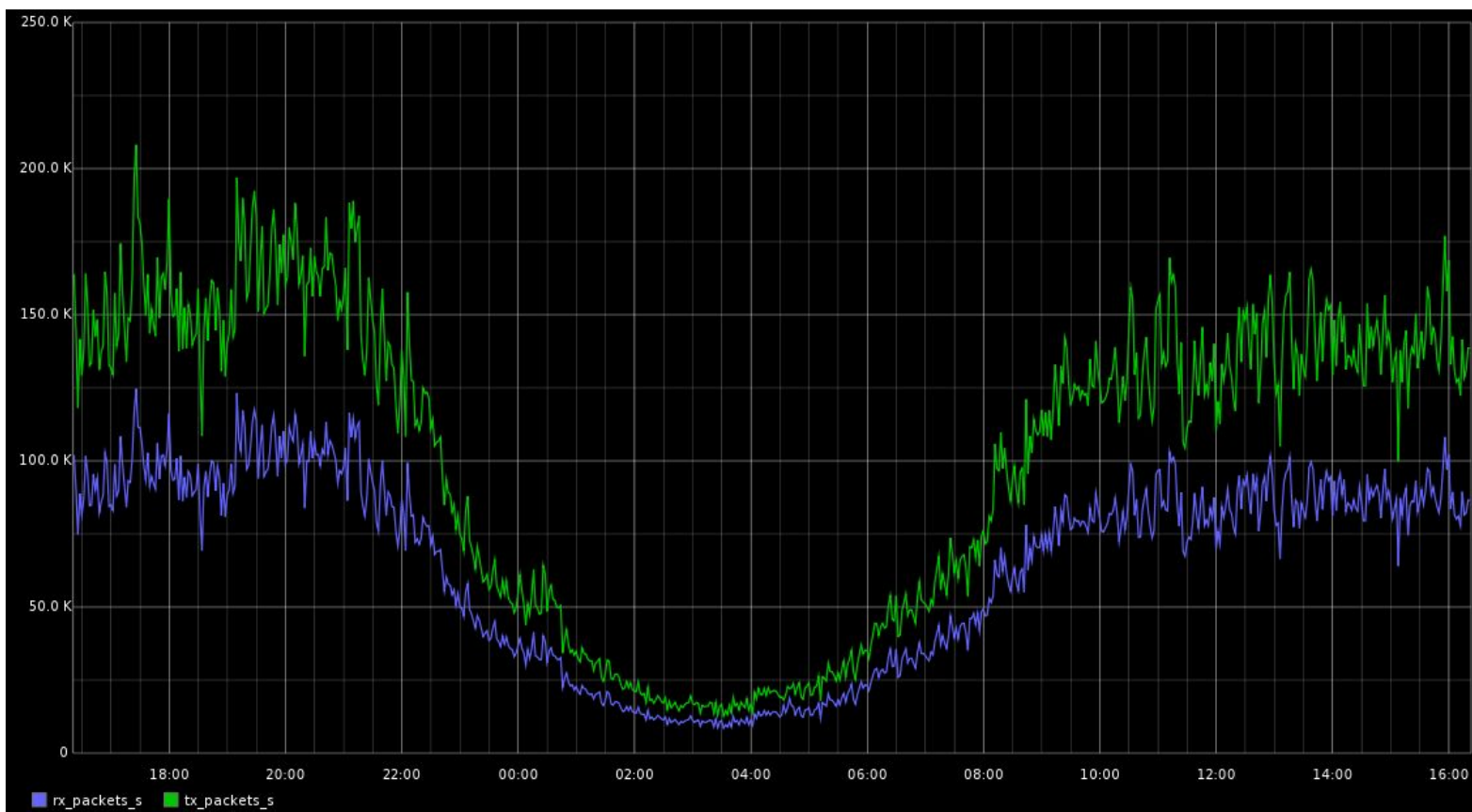
Server #1: bytes per second



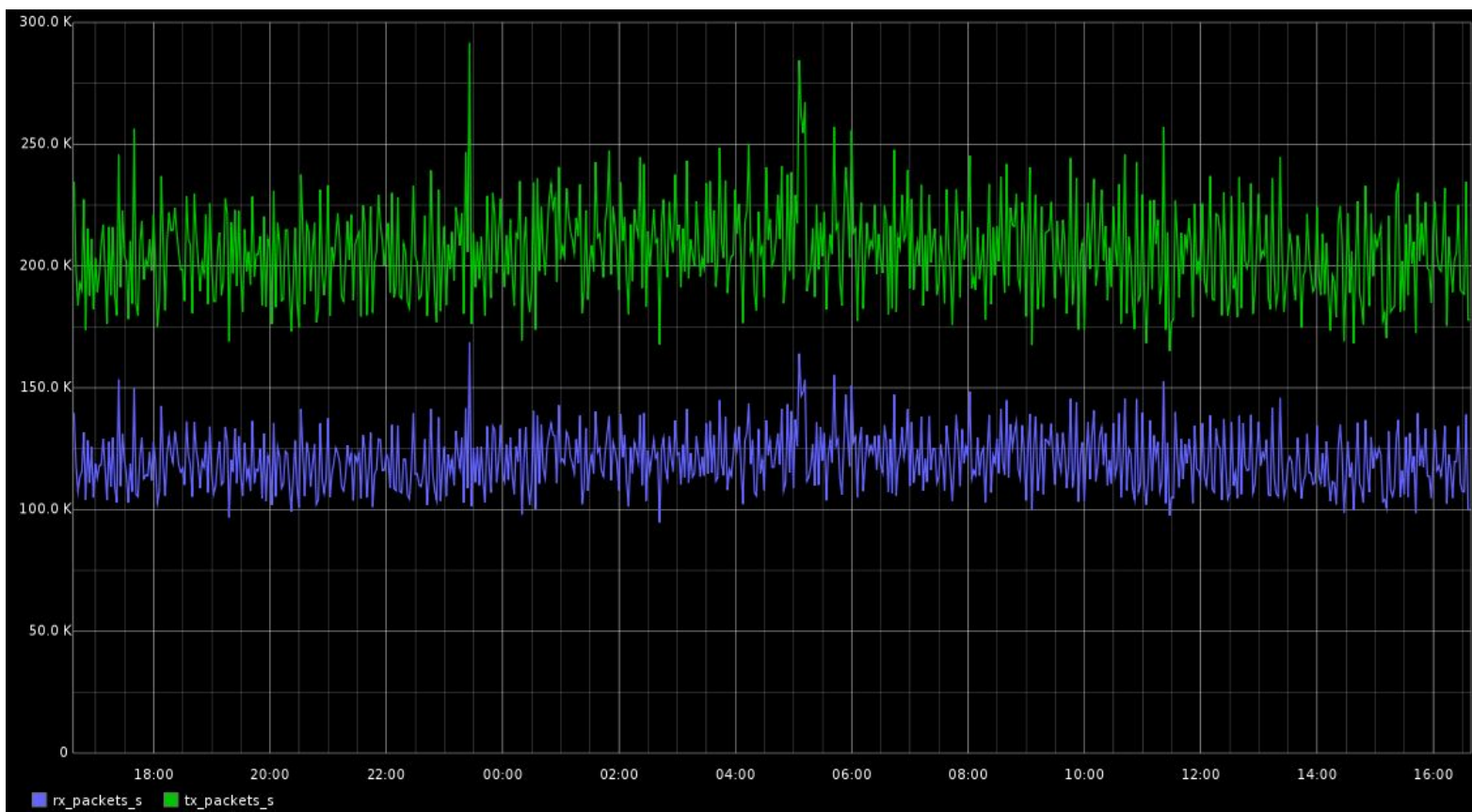
Server #2: bytes per second



Server #1: packets per second



Server #2: packets per second



Сетевой трафик крупных проектов

Проект	Трафик	Данные
Реклама Mail.Ru	45 Гбит/с	2017
Mail.Ru Group	3.4 Тбит/с	2015
Google (с Youtube)	>10% трафика в мире	
CloudFlare	>10% трафика в мире	
Netflix	37% трафика в США	2015

Размер пакета

- **Maximum Transmission Unit (MTU)**
Значение порядка 1400 байт
- **Jumbo Frames**
Технология увеличения размера пакета в контролируемой (собственной) сети

HighLoad. Лекция №2

Survey Peak Attack Size Year Over Year

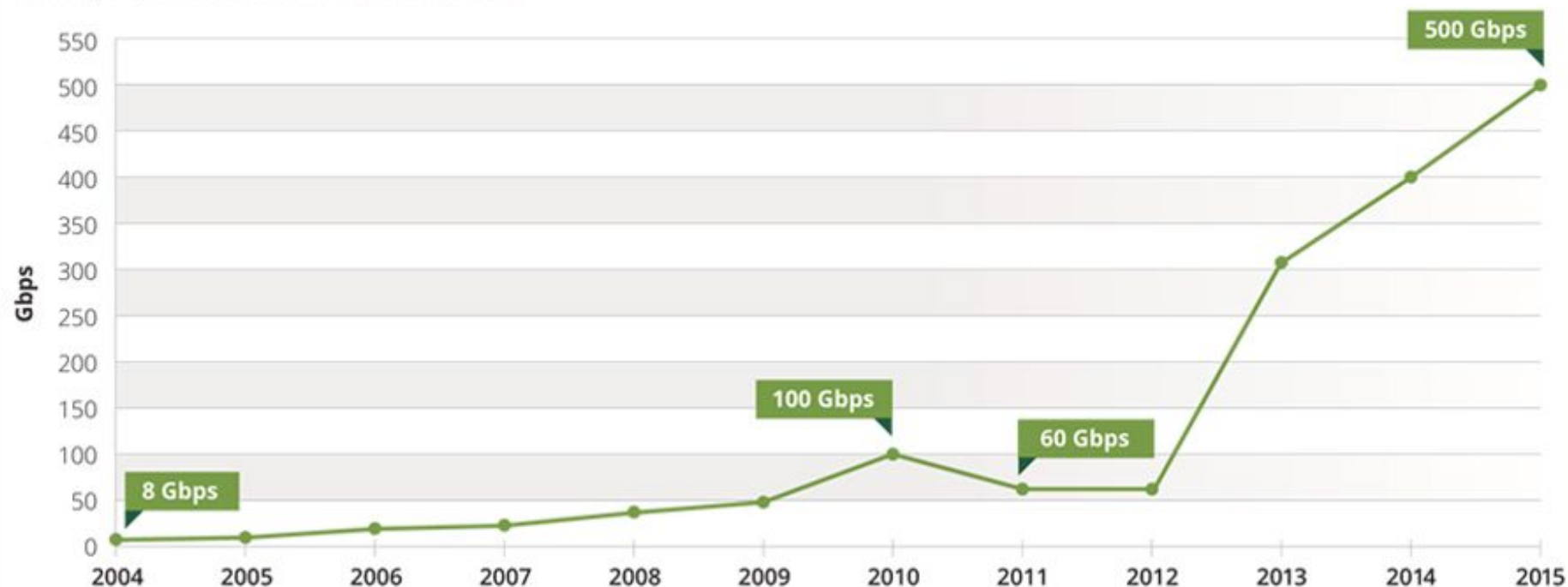
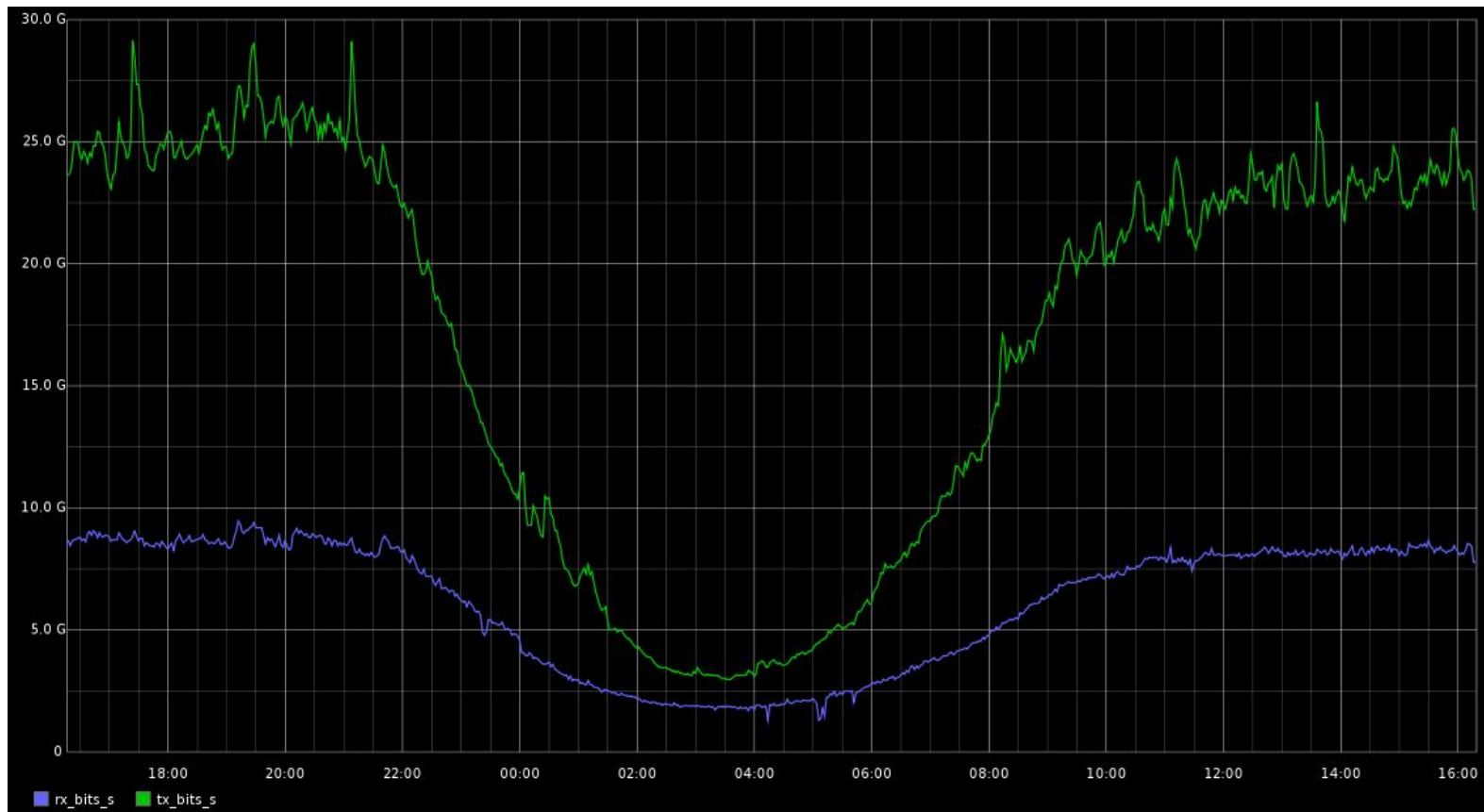


Figure 14 Source: Arbor Networks, Inc.

* 665 Gbps в сентябре 2016 года

Рекламная система Mail.Ru



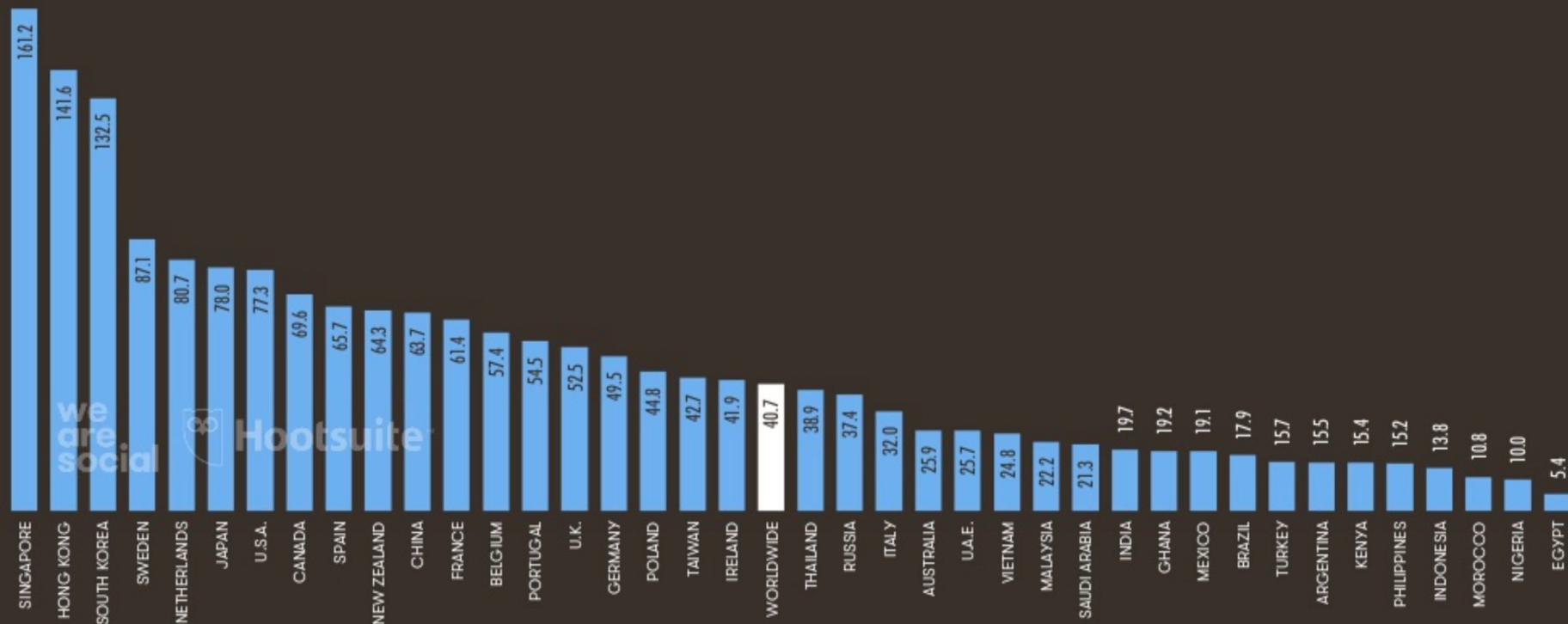
HighLoad. Лекция №2



JAN
2018

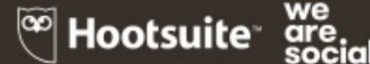
AVERAGE FIXED INTERNET CONNECTION SPEEDS

AVERAGE SPEED OF FIXED INTERNET CONNECTIONS BY COUNTRY, IN MBPS



35

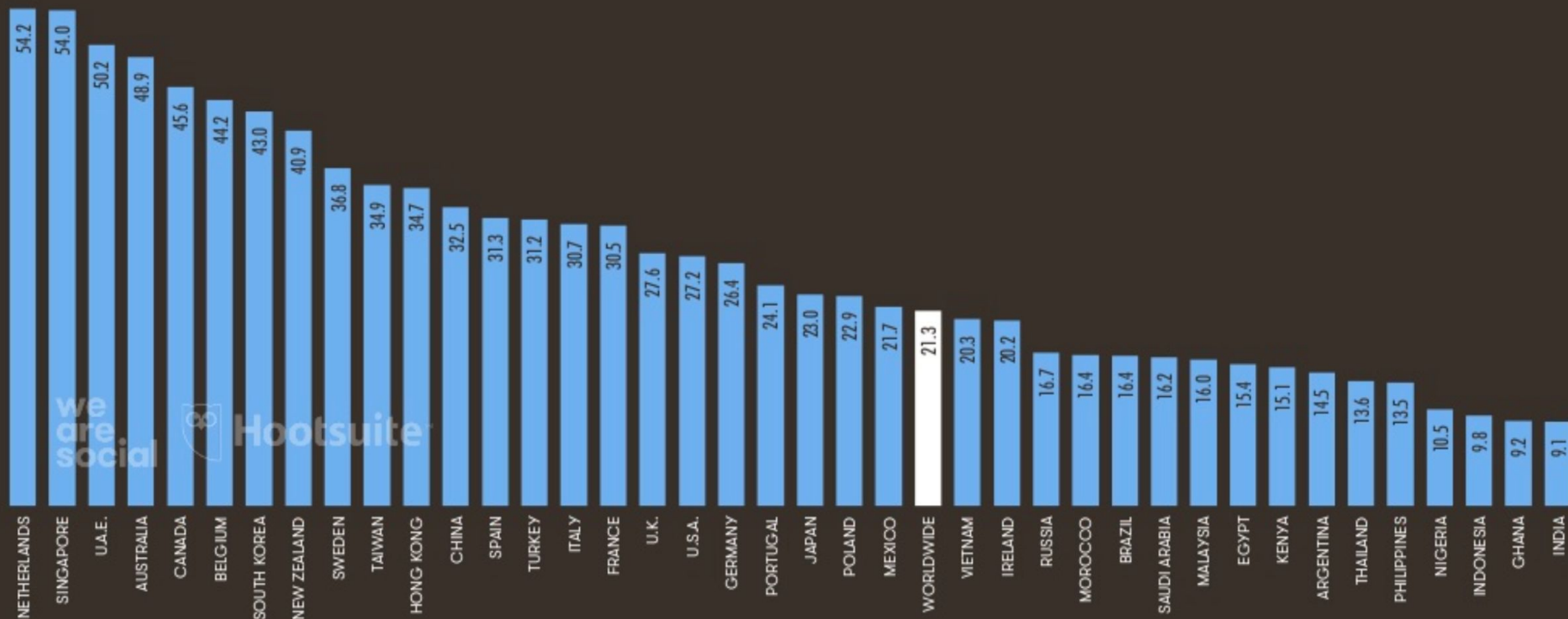
SOURCE: OOKLA SPEEDTEST, DECEMBER 2017. NOTES: FIGURES REPRESENT AVERAGE DOWNLOAD SPEEDS.



JAN
2018

AVERAGE MOBILE INTERNET CONNECTION SPEEDS

AVERAGE SPEED OF MOBILE INTERNET CONNECTIONS BY COUNTRY, IN MBPS



Latency numbers every programmer should know (Jeff Dean)

L1 cache reference	0.5 ns	
Branch mispredict	5 ns	
L2 cache reference	7 ns	
Mutex lock/unlock	25 ns	
Main memory reference	100 ns	
Compress 1K bytes with Zippy	3,000 ns	
Send 2K bytes over 1 Gbps network	20,000 ns	
SSD random read	150,000 ns	
Read 1 MB sequentially from memory	250,000 ns	
Round trip within same datacenter	500,000 ns	0.5 ms
Read 1 MB sequentially from SSD*	1,000,000 ns	1 ms
Disk seek	10,000,000 ns	10 ms
Read 1 MB sequentially from disk	20,000,000 ns	20 ms
Send packet CA->Netherlands->CA	150,000,000 ns	150 ms

HighLoad. Лекция №2

Input Interpretation:

distance	from	Moscow
	to	San Francisco, California, United States

Result:

9472 km (kilometers)

Unit conversions:

5885 miles

9472 km (kilometers)

9.472×10^6 meters

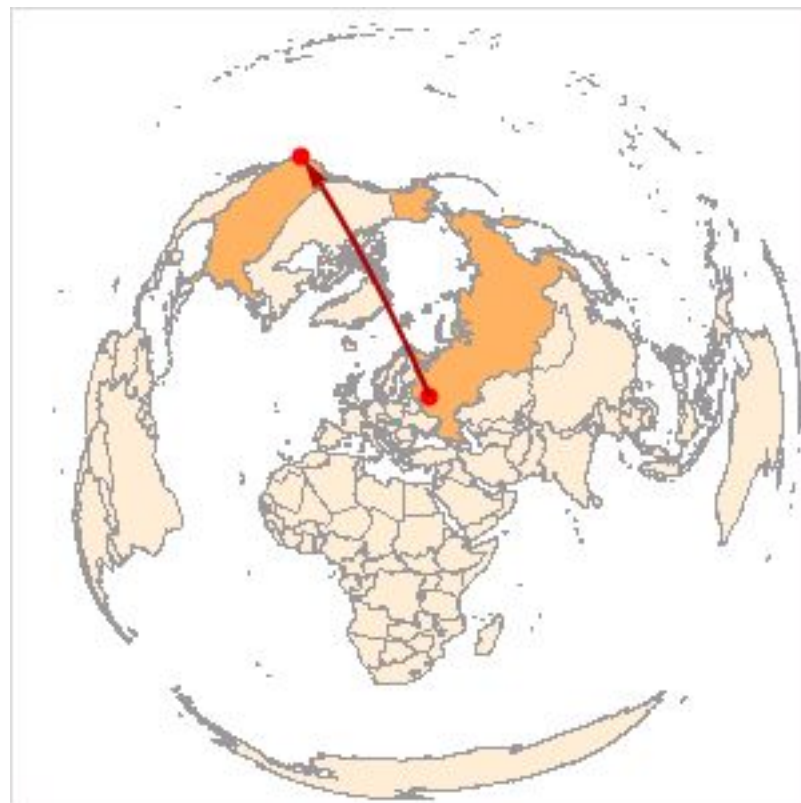
5114 nmi (nautical miles)

Direct travel times:

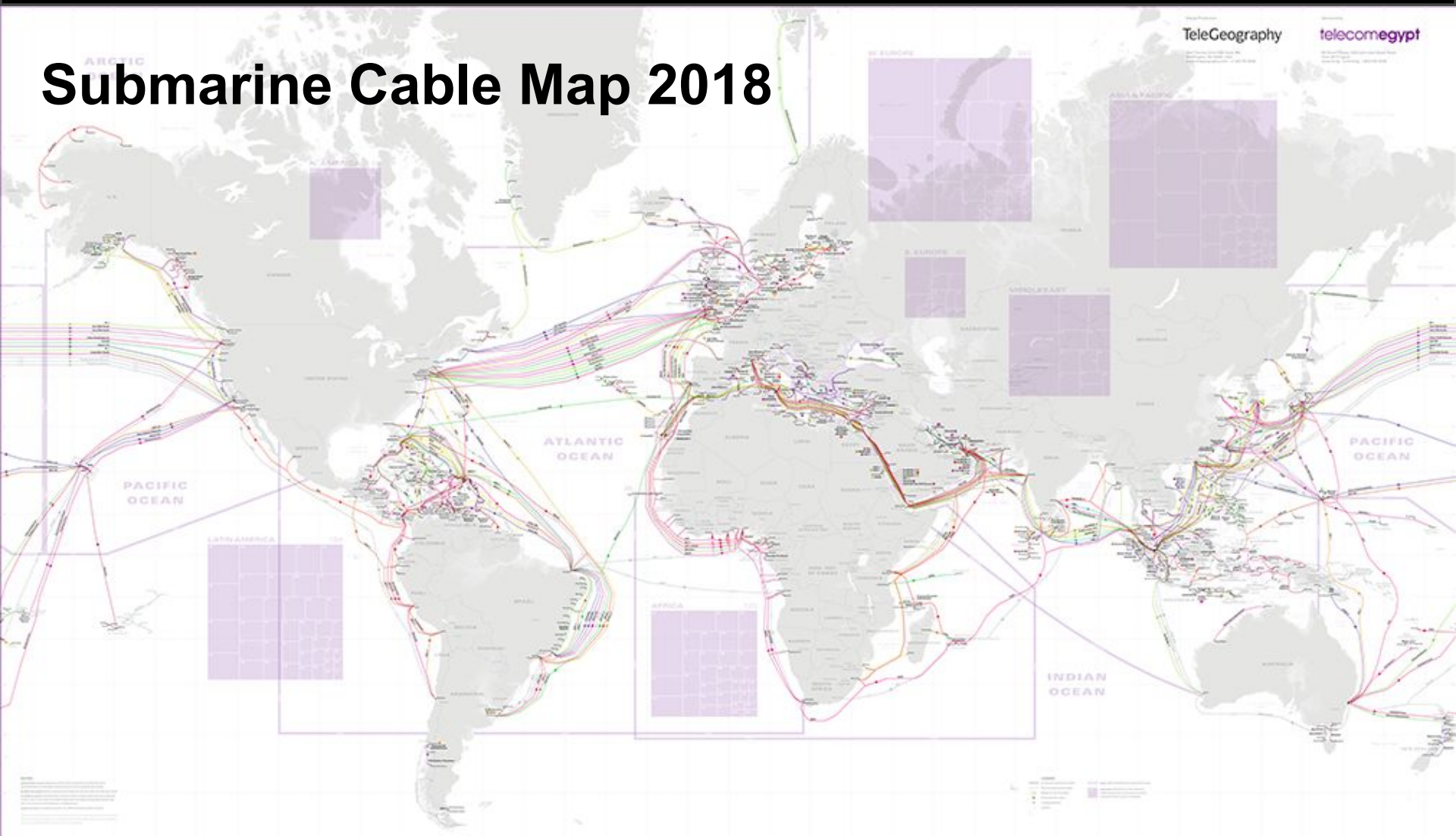
More

aircraft (550 mph)	10 hours 40 minutes
sound	7 hours 40 minutes
light in fiber	44.3 ms (milliseconds)
light in vacuum	31.6 ms (milliseconds)

(assuming constant-speed great-circle path)



Submarine Cable Map 2018



Сетевая задержка

Ping times between cities

	Amsterdam ✕	Copenhagen ✕	Frankfurt ✕	Helsinki ✕	London ✕	Moscow ✕	New York ✕	San Jose ✕	Tokyo ✕
Amsterdam		153.86ms	14.20ms	29.89ms	22.96ms	47.30ms	96.42ms	226.86ms	270.79ms
Copenhagen	82.74ms		112.17ms	52.37ms	40.76ms	60.72ms	115.27ms	162.14ms	275.00ms
Frankfurt	16.50ms	123.25ms		47.21ms	24.63ms	63.82ms	96.70ms	167.43ms	267.34ms
Helsinki	30.06ms	52.32ms	47.16ms		34.95ms	18.31ms	131.37ms	203.62ms	303.50ms
London	22.89ms	90.72ms	25.24ms	34.15ms		62.98ms	77.56ms	142.09ms	270.28ms
Moscow	47.38ms	109.00ms	60.62ms	19.27ms	62.60ms		123.14ms	200.96ms	294.62ms
New York	93.54ms	174.28ms	93.81ms	127.96ms	76.60ms	121.55ms		79.99ms	174.69ms
San Jose	192.61ms	209.72ms	167.60ms	200.84ms	141.93ms	200.95ms	81.45ms		108.03ms
Tokyo	270.55ms	274.68ms	268.34ms	303.05ms	270.80ms	295.01ms	176.68ms	108.09ms	

HighLoad. Лекция №2



```
# host cas.sv.us.criteo.com
```

```
cas.sv.us.criteo.com has address 74.119.117.72
```

```
# whois 74.119.117.72
```

```
NetRange:      74.119.116.0 - 74.119.119.255
```

```
CIDR:          74.119.116.0/22
```

```
OriginAS:
```

```
NetName:       CRITEO-USA
```

```
NetHandle:     NET-74-119-116-0-1
```

```
Parent:        NET-74-0-0-0-0
```

```
NetType:       Direct Assignment
```

```
RegDate:       2009-11-05
```

```
Updated:       2012-03-02
```

```
Ref:           http://whois.arin.net/rest/net/NET-74-119-116-0-1
```

```
OrgName:       Criteo Corp.
```

```
OrgId:         CRITE-6
```

```
Address:       411 High Street
```

```
City:          Palo Alto
```

```
StateProv:     CA
```

```
PostalCode:    94301
```

```
Country:       US
```

```
RegDate:       2009-10-05
```

```
Updated:       2010-07-20
```

```
Ref:           http://whois.arin.net/rest/org/CRITE-6
```

Измеряем реальную задержку

```
# ping cas.sv.us.criteo.com
```

```
PING cas.sv.us.criteo.com (74.119.117.72) 56(84) bytes of data.  
64 bytes from 74.119.117.72: icmp_seq=1 ttl=246 time=196 ms  
64 bytes from 74.119.117.72: icmp_seq=2 ttl=246 time=196 ms  
64 bytes from 74.119.117.72: icmp_seq=3 ttl=246 time=196 ms  
64 bytes from 74.119.117.72: icmp_seq=4 ttl=246 time=196 ms  
64 bytes from 74.119.117.72: icmp_seq=5 ttl=246 time=196 ms  
64 bytes from 74.119.117.72: icmp_seq=6 ttl=246 time=196 ms
```

```
--- cas.sv.us.criteo.com ping statistics ---
```

```
7 packets transmitted, 6 received, 14% packet loss, time 6069ms  
rtt min/avg/max/mdev = 196.745/196.769/196.806/0.256 ms
```

Анализируем маршрут

```
# traceroute cas.sv.us.criteo.com -q 1
```

```
traceroute to cas.sv.us.criteo.com (74.119.117.72), 30 hops max, 60 byte packets
```

```
1  94.100.178.2 (94.100.178.2)  0.408 ms
2  188.254.103.197 (188.254.103.197)  8.875 ms
3  46.61.141.133 (46.61.141.133)  51.420 ms
4  ethernet10-3.ar4.fra4.gblx.net (64.211.193.169)  45.973 ms
5  ae8.scr4.FRA4.gblx.net (67.16.145.241)  51.687 ms
6  po3-20G.ar2.SNV2.gblx.net (67.16.139.98)  195.523 ms
7  CRITEO-CORP.GigabitEthernet4-18.ar2.SNV2.gblx.net (206.41.25.26)  202.311 ms
8  *
9  *
10 *
11 *
12 *
13 *
14 *
15 *
```

Traceroutes from **Moscow, RU** to **Yaroslavl, RU**



Looking Glass

Каталоги:

- <http://www.traceroute.net.ru/>
- <http://www.lookingglass.org/>
- <http://www.bgp4.as/looking-glasses>
- <http://www.bgp4.net/lg>

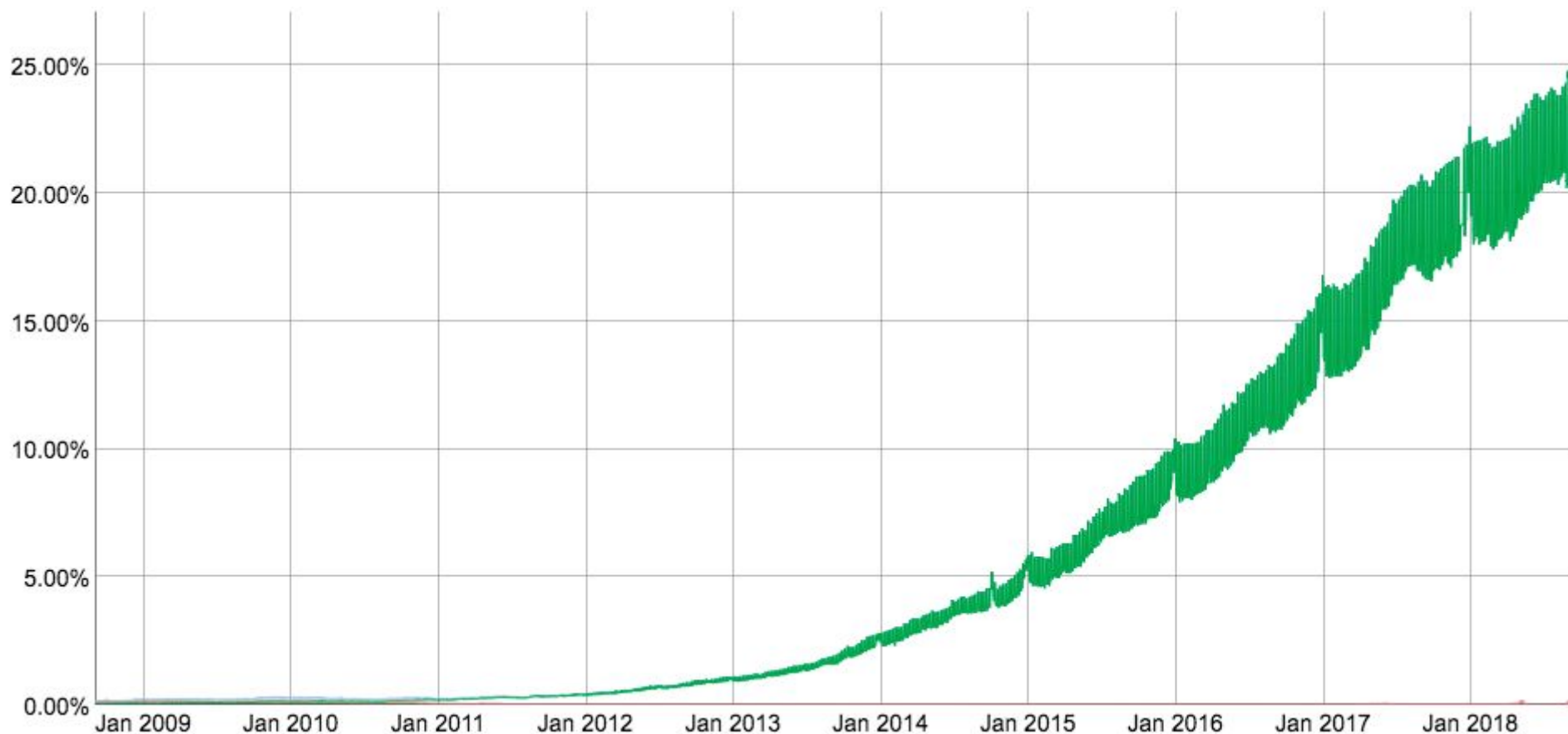
Примеры:

- <http://lg.vk.com/>
- <http://lg.megaфон.ru/>
- <http://lg.transtk.ru/>
- <http://lg.retn.net/>
- <http://lg.he.net/>
- <http://www.msk-ix.ru/network/lookingglass.html>
- <http://www.ris.ripe.net/cgi-bin/lg/index.cgi>

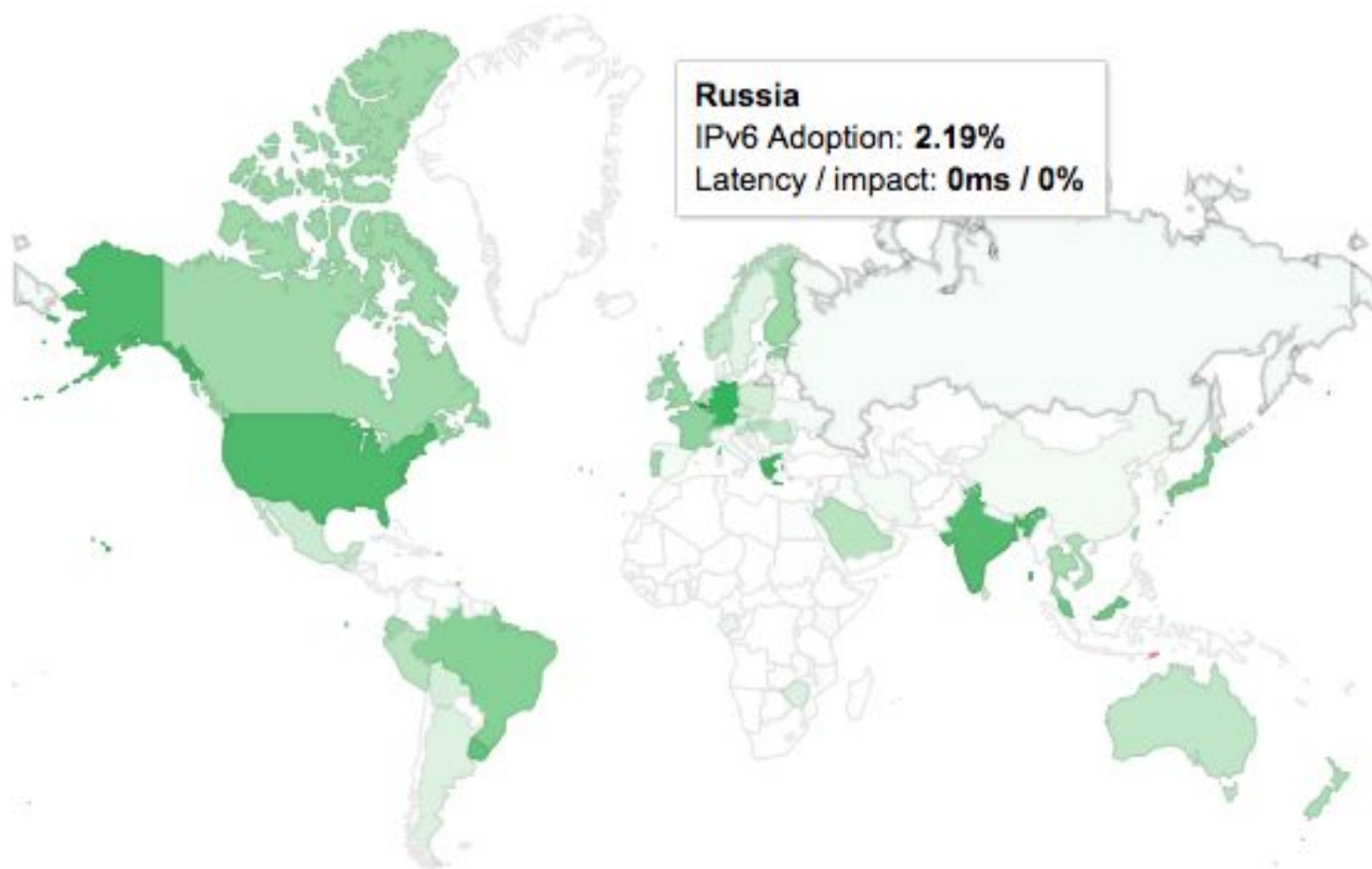
Задержки внутри ДЦ (города)

- Низкие только на первый взгляд
- Для задач с интенсивным синхронным обменом данными по сети весьма ощутимы

IPv6 adoption (Google)



Per Country IPv6 adoption (Google)



Протокол TCP/IP (version 4)

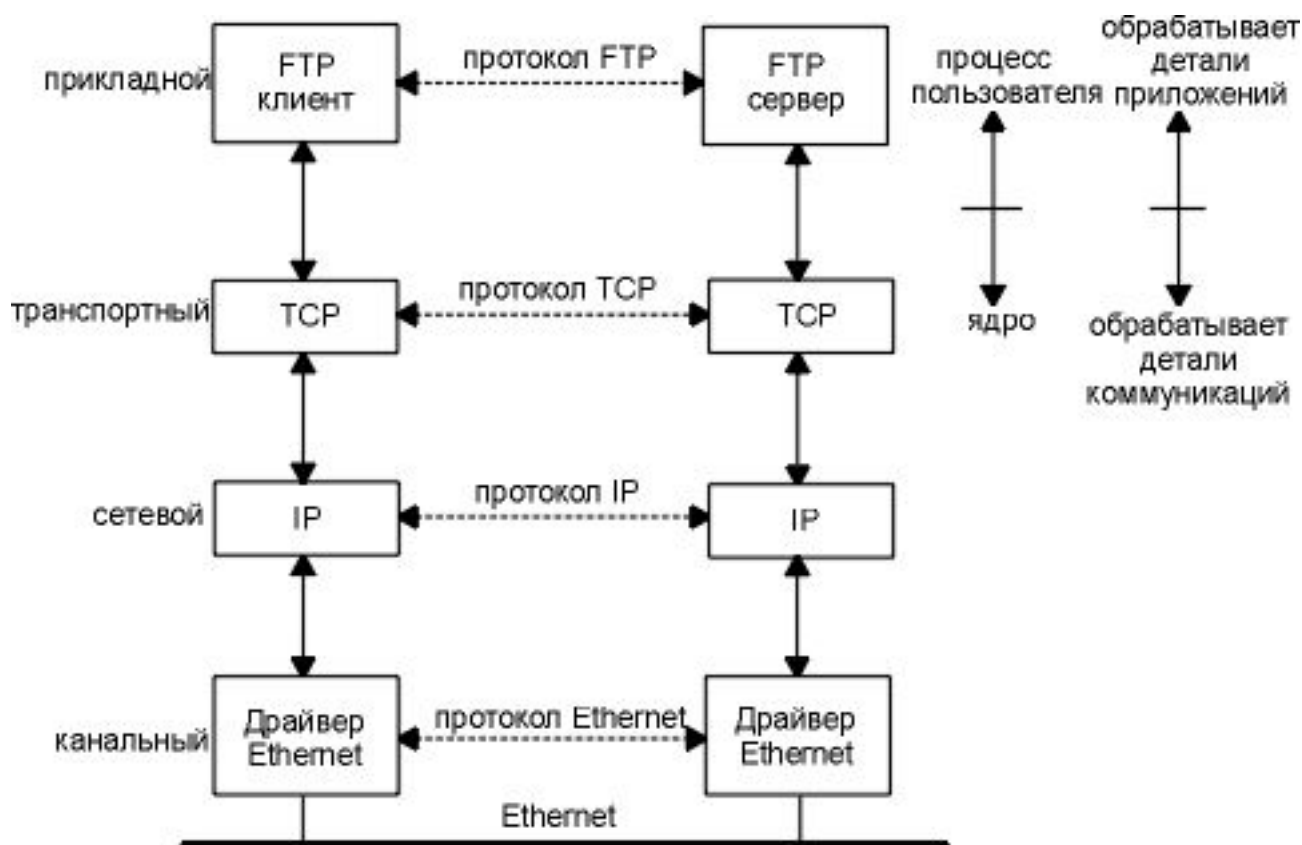
Модель OSI (Open System Interconnection Reference Model)

OSI Model			
	Data unit	Layer	Function
Host layers	Data	7. Application	Network process to application
		6. Presentation	Data representation, encryption and decryption, convert machine dependent data to machine independent data
		5. Session	Interhost communication, managing sessions between applications
	Segments	4. Transport	End-to-end connections, reliability and flow control
Media layers	Packet/Datagram	3. Network	Path determination and logical addressing
	Frame	2. Data link	Physical addressing
	Bit	1. Physical	Media, signal and binary transmission

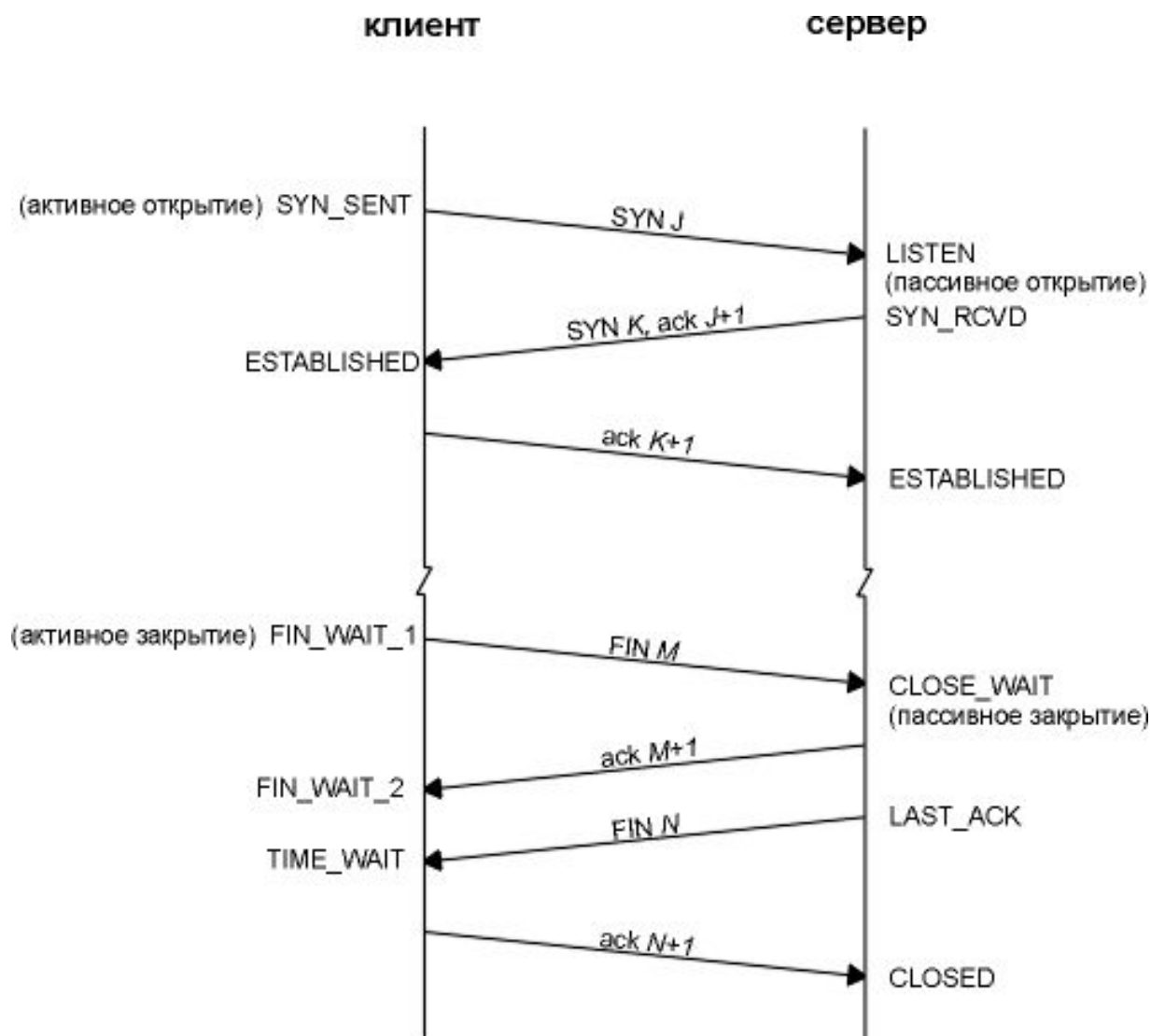
Модель OSI в применении к TCP/IP

7	Application	e.g. HTTP, SMTP, SNMP, FTP, Telnet, SSH and Scp, NFS, RTSP etc.
6	Presentation	e.g. XDR, ASN.1, SMB, AFP etc.
5	Session	e.g. TLS, SSH, ISO 8327 / CCITT X.225, RPC, NetBIOS, ASP etc.
4	Transport	e.g. TCP, UDP, RTP, SCTP, SPX, ATP etc.
3	Network	e.g. IP/IPv6, ICMP, IGMP, X.25, CLNP, ARP, RARP, BGP, OSPF, RIP, IPX, DDP etc.
2	Data Link	e.g. Ethernet, Token ring, PPP, HDLC, Frame relay, ISDN, ATM, 802.11 Wi-Fi, FDDI etc.
1	Physical	e.g. wire, radio, fiber optic etc.

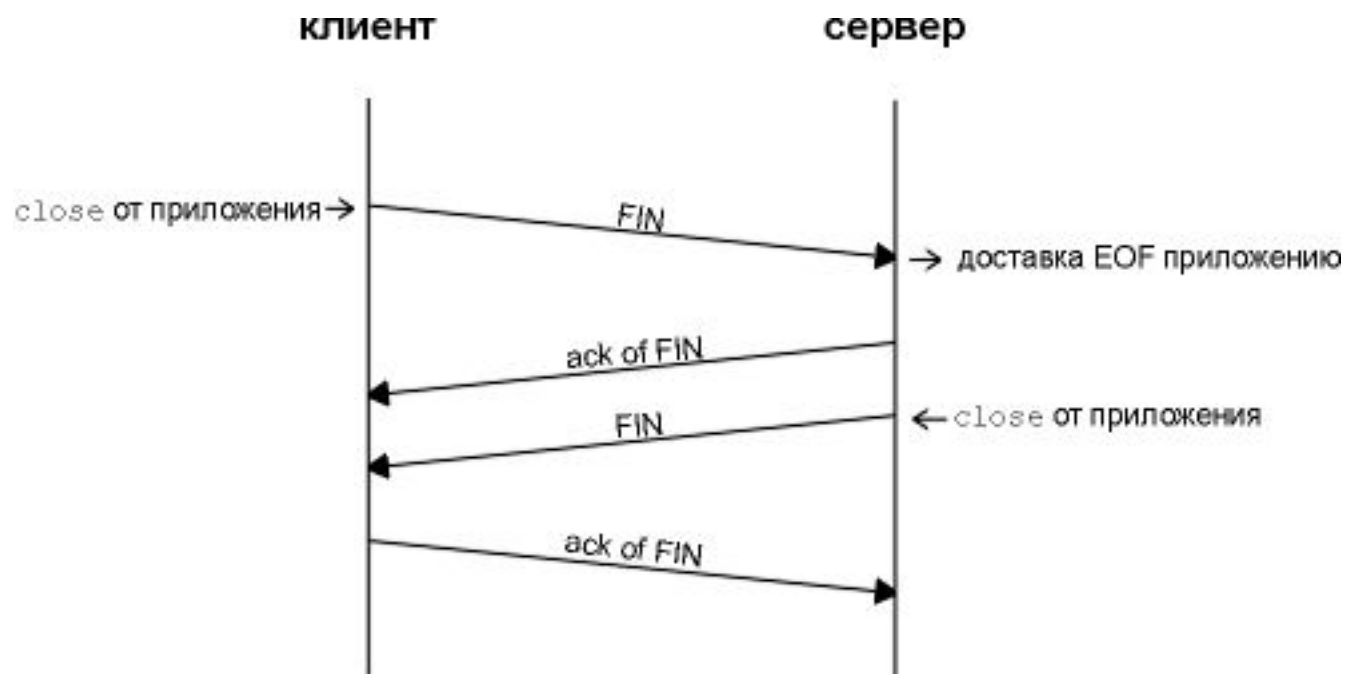
Модель OSI в применении к TCP/IP



HighLoad. Лекция №2



HighLoad. Лекция №2



TCP window scaling

Алгоритм Slow Start:

- Размер окна увеличивается на MSS после каждого ACK
- Размер окна удваивается на RTT
- В случае потери пакета начинаем с начала
- Есть верхнее ограничение размера окна с обеих сторон

TCP window scaling

Альтернативы:

- BBR - алгоритм учитывающий задержки (Google)

Потеря пакетов (packet loss)

Повторная отправка пакета:

- TCP: retransmit timeout = 3 s
- DNS: retransmit timeout = 5 s

Последствия:

- Ощутимая задержка для пользователя
- Повышенная нагрузка на сервис

Потеря пакетов внутри ДЦ

- Рост потребления ресурсов синхронными сервисами
- Каскадное увеличение нагрузки из-за ретрансмиттов

Проблема медленных соединений

- Занимают ресурсы ОС
- Занимают память в веб-сервере
- Блокируют работу синхронного процесса или потока
- Могут использоваться как метод DDoS-атаки

Недостатки TCP/IP

- Большое время установления соединения
- Относительно дорогое создание соединения
- Долгая реакция на потерю пакета
- Медленный разгон window size
- Высокая нагрузка на ОС при обработке большого количества активных соединений

Применение UDP

Достоинства:

- Неблокирующая отправка
- Низкая нагрузка на ОС
- Возможность многоадресной передачи

Недостатки:

- Ненадежный
- Неупорядоченный
- Необходимо следить за нагрузкой на сеть

Применения:

- Сбор статистики
- Уведомления
- Специализированные протоколы

Технология: Point-of-Presence

- Прокси-сервер (NAT) рядом с пользователем
- Прокси-сервер держит постоянное соединение с ДЦ
- Окно передачи разогнанное
- Скорость выше чем при подключении без прокси

Тюнинг TCP

`/etc/sysctl.conf:`

```
net.ipv4.tcp_max_syn_backlog = 1024
```

```
net.ipv4.tcp_max_orphans = 65536
```

```
net.ipv4.tcp_max_tw_buckets = 180000
```

```
net.ipv4.tcp_max_syn_backlog = 32768
```

```
net.ipv4.tcp_max_orphans = 131072
```

```
net.ipv4.tcp_max_tw_buckets = 1800000
```

`/usr/src/linux/Documentation/networking/ip-sysctl.txt`

Тюнинг UDP

/etc/sysctl.conf:

net.core.rmem_default = 129024

net.core.wmem_default = 129024

net.core.rmem_max = 131071

net.core.wmem_max = 131071

net.ipv4.udp_mem = 387840 517120 775680

/usr/src/linux/Documentation/networking/ip-sysctl.t
xt

Список литературы

1. Netflix CDN
<https://events.yandex.ru/lib/talks/2396/>
2. How to receive a million packets per second
<https://blog.cloudflare.com/how-to-receive-a-million-packets/>
3. Тюнинг сетевого стека в Одноклассниках
<https://habrahabr.ru/company/odnoklassniki/blog/266005/>
4. Submarine Cable Map 2018
<https://submarine-cable-map-2018.telegeography.com/>



СПАСИБО ЗА ВНИМАНИЕ

Быков Александр
bykov@corp.mail.ru