西安电子科技大学
XIDIAN UNIVERSITY
计算机科学与技术学院
School of Computer Science and Technology

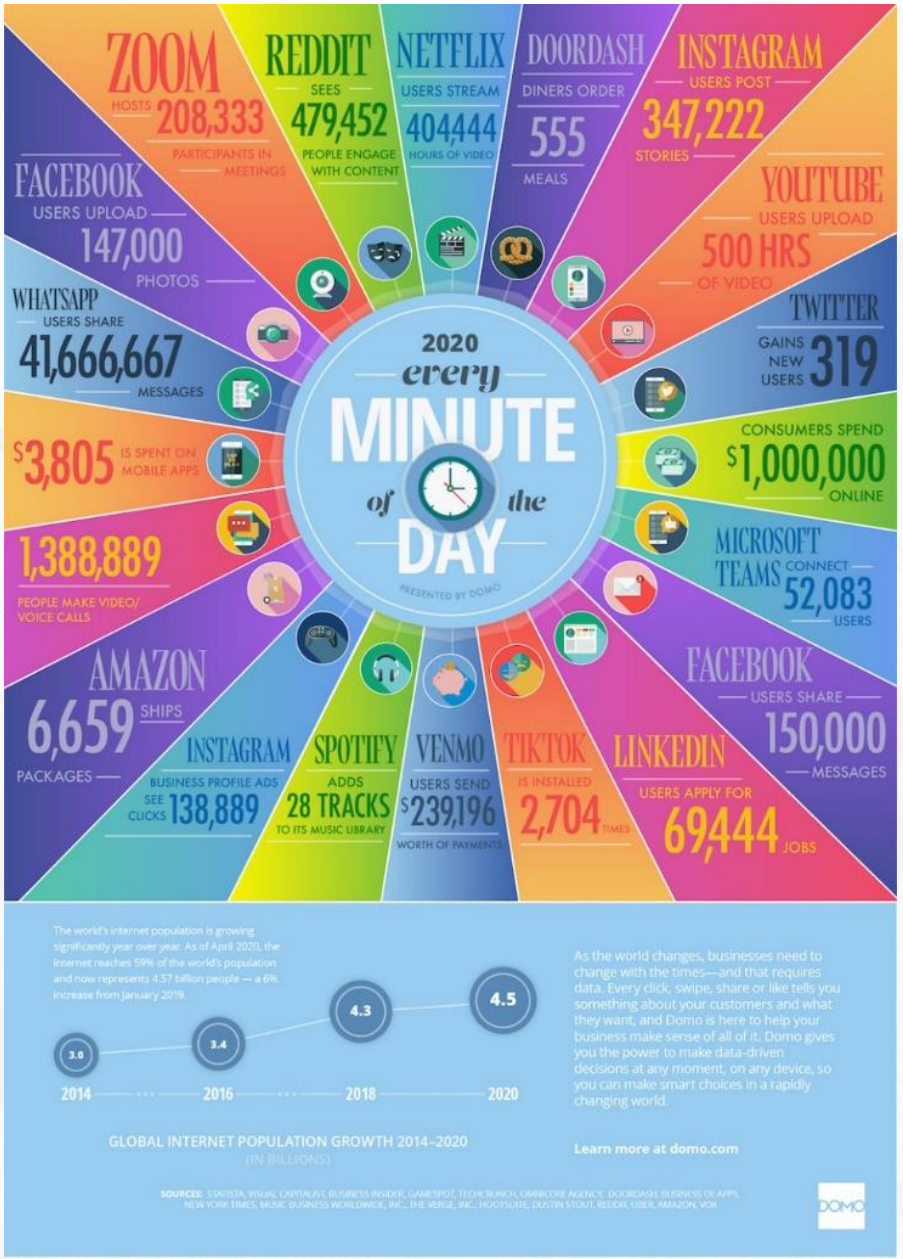
第三章 大数据采集传输的安全与隐私

彭延国

ygpeng@xidian.edu.cn

§3 大数据采集传输的安全与隐私 - 引子

- 大数据的全生命周期
  - “末端”到“神经中枢”；
  - 认证、安全传输。

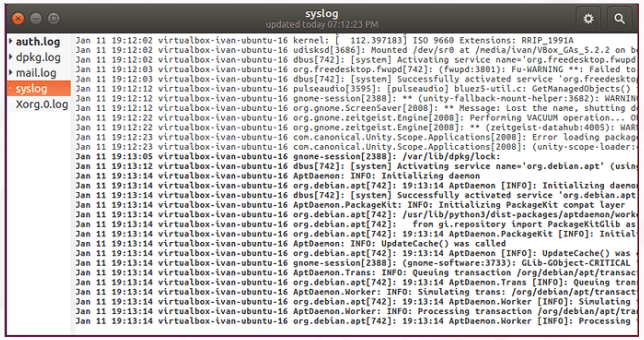


西安电子科技大学
XIDIAN UNIVERSITY

§3.1 大数据采集技术

§3.1 大数据采集技术 - 数据采集方法

- 系统日志采集：
  - 使用日志收集系统，收集业务日志数据供离线和在线的分析系统使用。

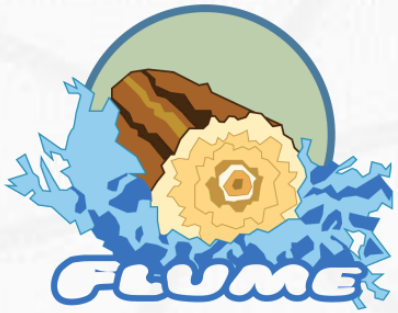


- 网络数据采集：
  - 通过网络爬虫或网站公开API等方式从网站上获取数据信息，可以将非结构化数据以结构化的方式存储。

- 数据库采集：
  - 在采集端部署大量数据库，并对如何在这些数据库之间进行负载均衡和分片进行深入的思考和设计。

§3.1 大数据采集技术 - 数据采集产品

- 数据采集产品有很多，较为常用的是以下六种
  - Apache Flume、Scribe、Fluentd、Apache Chukwa、Logstash、Splunk



splunk>

§3.1 大数据采集技术 - Apache Flume(1)

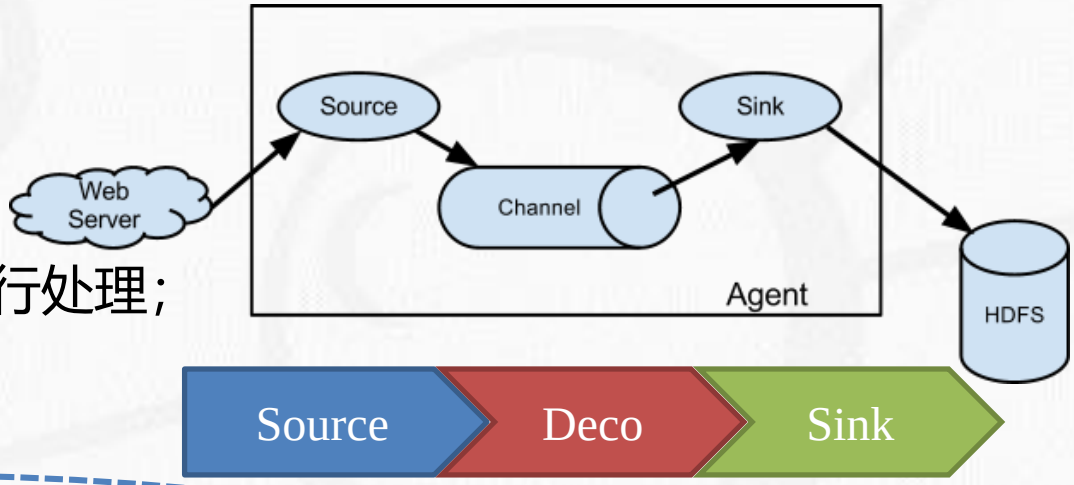
- Apache Flume：
  - 一种分布式、可靠且可用的服务，用于有效地收集、聚合和移动大量日志数据；
- 提出的背景：
  - 大量的运行于末端的服务，会产生大量的日志数据，汇总后需要集中(HDFS)处理。
- 目标：
  - 可靠性(Reliability)、可延展性(Scalability)、可扩展性(Extensibility)、可管理性(Manageability)。

§3.1 大数据采集技术 - Apache Flume(2)

- 流(Flow)模式：
  - 对应不同的数据源：服务日志、机器监视参数等。

§3.1 大数据采集技术 - Apache Flume(3)

- 节点Nodes：
  - 数据来源于Source；
  - 选择性的在decorator中进行处理；
  - 最后通过Sink进行传输。



- Source：
  - Console, Exec, Syslog, IRC, Twitter, other nodes；

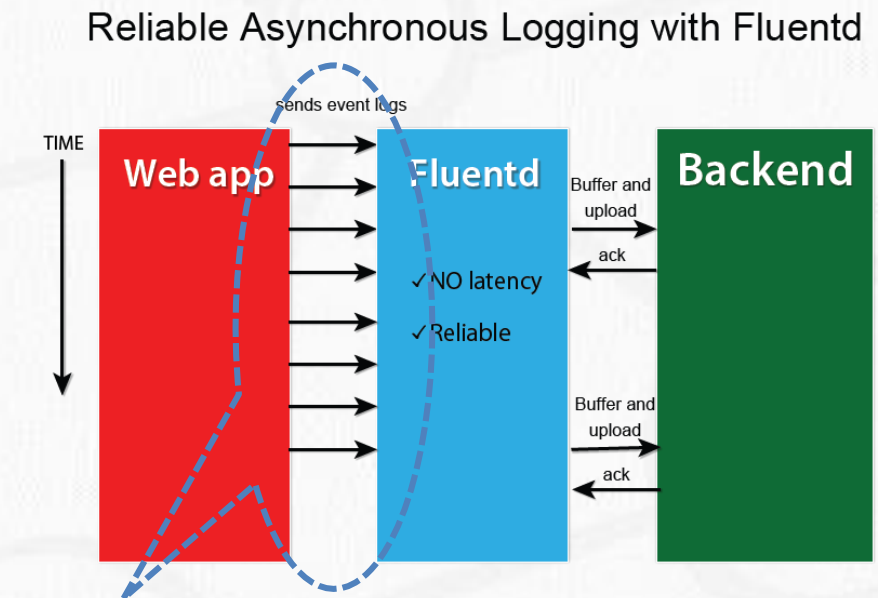
数据采集

- Decorator：
  - wire batching, compression, sampling, projection, extraction...
- Sink：
  - Console, local files, HDFS, S3, other nodes

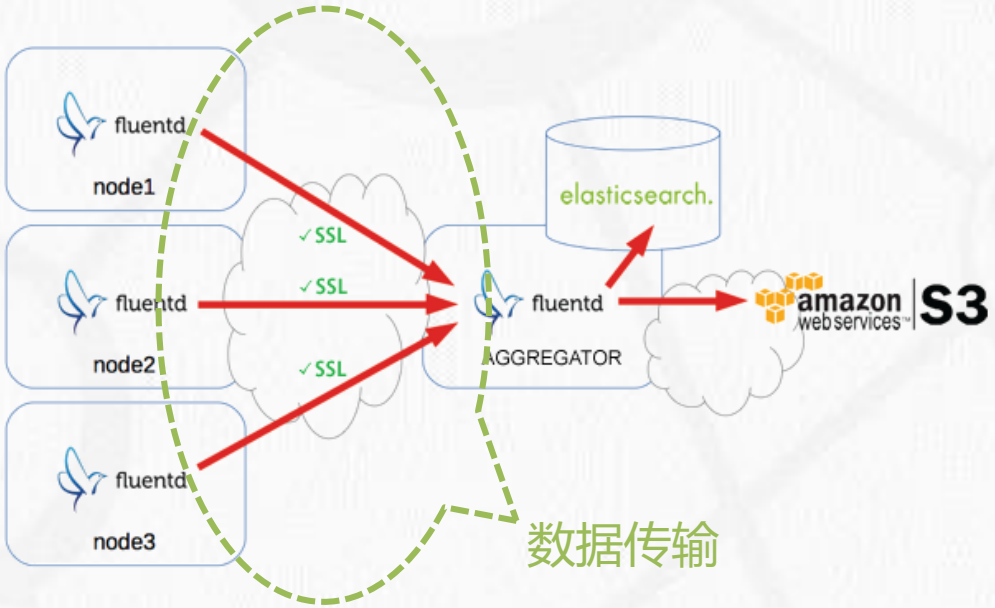
数据传输

§3.1 大数据采集技术 - Fluentd

- Fluentd：
  - 收集来自各种数据源的事件，并把他们写入文件、关系型数据库RDBMS、NoSQL数据库、 IaaS平台等。
  - 一个事件是一个三元组：tag，time，record。



可靠的异步日志记录



Fluentd的系统结构

西安电子科技大学
XIDIAN UNIVERSITY

SSID、WEP、WPA

§3.2 无线网接入安全

§3.2 无线网接入安全 - 无线网初探

- 载体：无线电波
  - 无需物理插入网络；
  - 远程访问。
- 覆盖范围：
  - 个人局域网（PAN）、局域网（LAN）、城域网（MAN）
- 安全问题
  - 无线电信号泄漏到建筑物外面
  - 检测未经授权的设备
  - 拦截无线通信
  - 中间人攻击
  - 验证用户
  - 限制访问

	WiFi 4	WiFi 5		WiFi 6
协议	802.11n	802.11ac		802.11ax
		Wave1	Wave2	
年份	2009	2013	2016	2018
频段	2.4 GHz、5 GHz	5 GHz		2.4 GHz、5 GHz
最大频宽	40 MHz	80 MHz	160 MHz	160 MHz
最高调制	64 QAM	256 QAM		1024 QAM
单流带宽	150 Mbps	433 Mbps	867 Mbps	1200 Mbps
最大带宽	600 Mbps	3466 Mbps	6933 Mbps	9.6 Gbps
最大空间流	4×4	8×8		8×8
MU-MIMO	N/A	N/A	下行	上行、下行
OFDMA	N/A	N/A	N/A	上行、下行

WiFi标准的发展

§3.2 无线网接入安全 - SSID

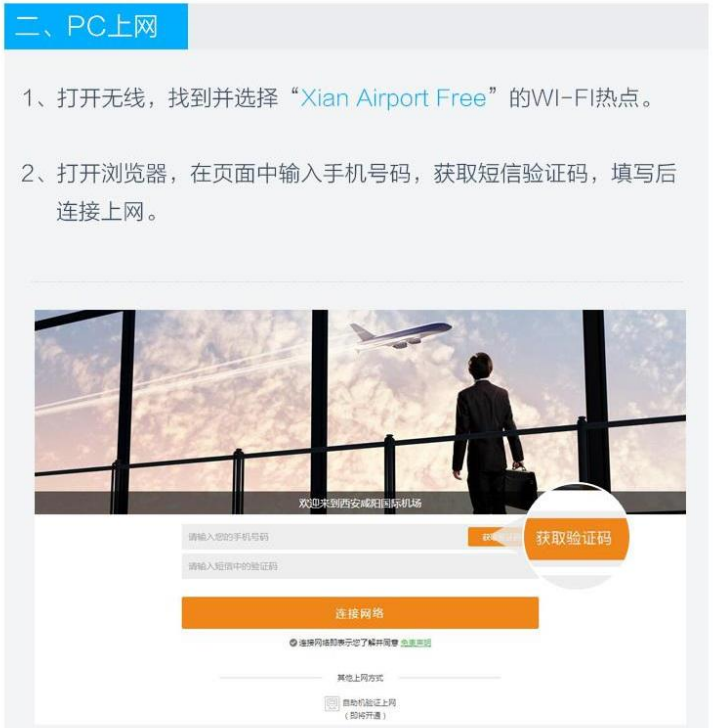
- 多个无线网络可以共存
  - 每个网络由32个字符的服务集ID（SSID）标识；
  - 制造商的名称是接入点的典型默认SSID;
  - 经常广播SSID以使潜在客户能够发现网络。
- SSID未被签名，因此可以进行简单的欺骗攻击
  - 将恶意接入点放置在公共场所（例如，咖啡馆，机场）；
  - 使用ISP的SSID;
  - 设置类似于ISP的登录页面;
  - 等待客户端连接到恶意接入点并进行身份验证;
  - 可能会转发到ISP网络的会话;
  - 通过默认的自动连接进入网络。

§3.2 无线网接入安全 - 窃听和欺骗

- 可以窃听所有无线网络流量：
  - 基于MAC的身份验证通常用于识别公司网络中已批准的计算机；
- MAC欺骗攻击可能：
  - 短暂断开连接后，会话保持活动状态；
  - 如果ISP客户端未明确结束会话，则MAC欺骗允许接管该会话。

§3.2 无线网接入安全 - 强制门户网站

- 协议
  - DHCP提供IP地址；
  - 名称服务器将一切映射到认证服务器；
  - 防火墙阻止所有其他流量
  - 任何URL都会重定向到身份验证页面；
  - 身份验证后，恢复常规网络服务；
  - 通过MAC地址识别的客户端；
  - 由无线ISP使用；



西安咸阳国际机场无线认证

- 安全问题：
  - 如果客户端没有主动断开连接，则可以执行MAC欺骗和会话窃取攻击；
  - 如果在身份验证之前未阻止防火墙之外的DNS流量，则隧道攻击可以绕过强制网络门户；

§3.2 无线网接入安全 - 有线等效保密(WEP)

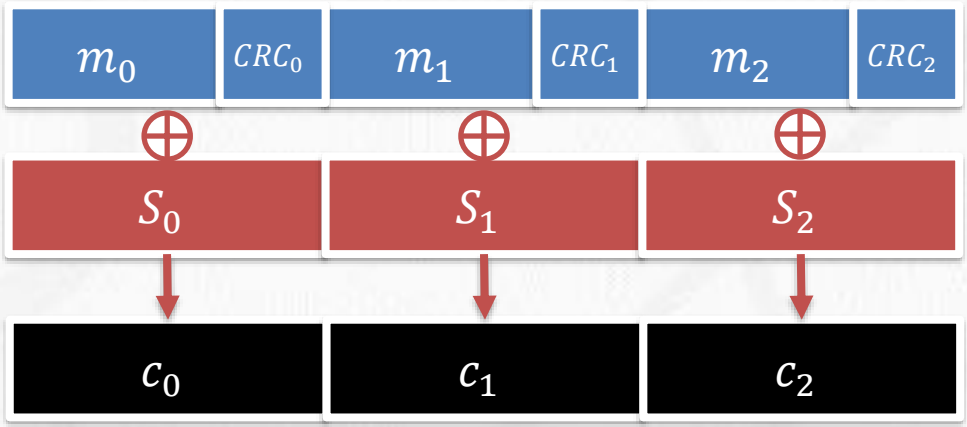
- 有线等效保密(WEP: Wired Equivalent Privacy)
  - 是1999年9月通过的 IEEE 802.11 标准的一部分；
  - 机密性：防止窃听；
  - 数据完整性：数据包不能被篡改；
  - 访问控制：仅路由正确加密的数据包；
- 设计约束
  - 采用90年代的廉价硬件实现
  - 符合美国早期加密设备出口管制法规（40位密钥）；
- 实施和限制
  - 在数据链路级别加密每个帧的主体
  - 要避免传统的IEEE 802.11标准

§3.2 无线网接入安全 - WEP协议的数据传输

- 初始化
  - 接入点和客户端共享40bit密钥K；
  - 密钥在WEP会话期间永远不会更改；

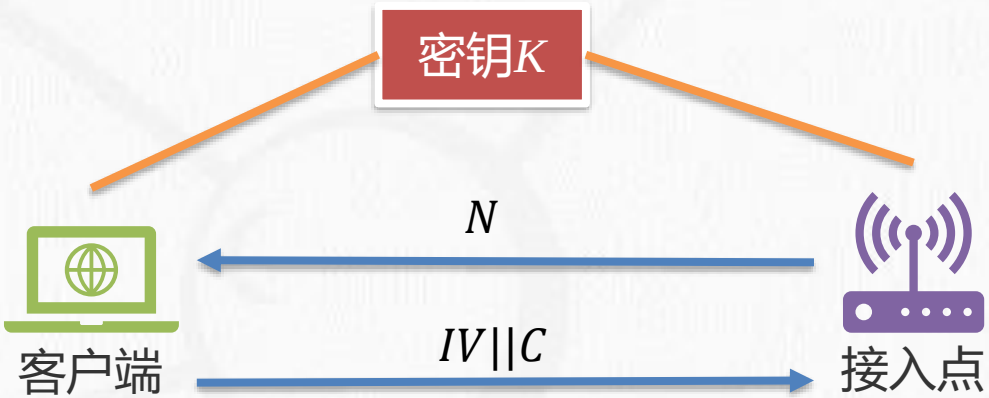
- 加密
  - 计算消息m的CRC-32校验和（帧的有效载荷）；
  - 选择24位初始化向量IV；
  - 使用RC4流密码生成密钥流$S(K, IV)$;
  - 计算密文:
    - $c = (m||CRC(m)) \oplus S(K, IV)$

线性校验和，无法抵抗线性攻击



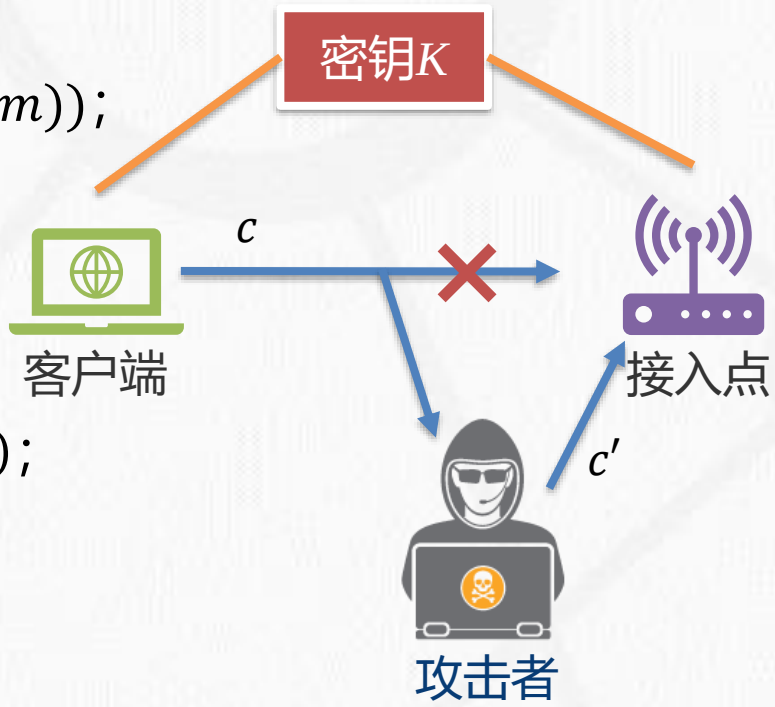
§3.2 无线网接入安全 - WEP协议的认证

- 客户端认证
  - 接入点向客户端发送未加密的随机质询N
  - 客户端响应加密挑战
- 传输
  - 发送$IV||C$，$C=(N||CRC(N))\oplus S(K,IV)$;



§3.2 无线网接入安全 - WEP的消息篡改攻击

- 消息篡改
  - 给定一个任意字符串Δ，我们想用Δ替换消息m
  - 在传输途中截取原始密文$c, c = (m \parallel CRC(m)) \oplus S(K, IV)$;
  - 计算$m' = m \oplus \Delta$;
  - 消息替换$c' = c \oplus (m' \parallel CRC(\Delta) \oplus CRC(m))$;



- 攻击分析：
  - $CRC(\Delta) \oplus CRC(m) \oplus CRC(m) = CRC(\Delta)$;
  - $m' \oplus m = m \oplus \Delta \oplus m = \Delta$。

§3.2 无线网接入安全 - WEP中IV的安全问题

- 初始化向量IV
  - 24位

$$c = (m||CRC(m)) \oplus S(K, IV)$$

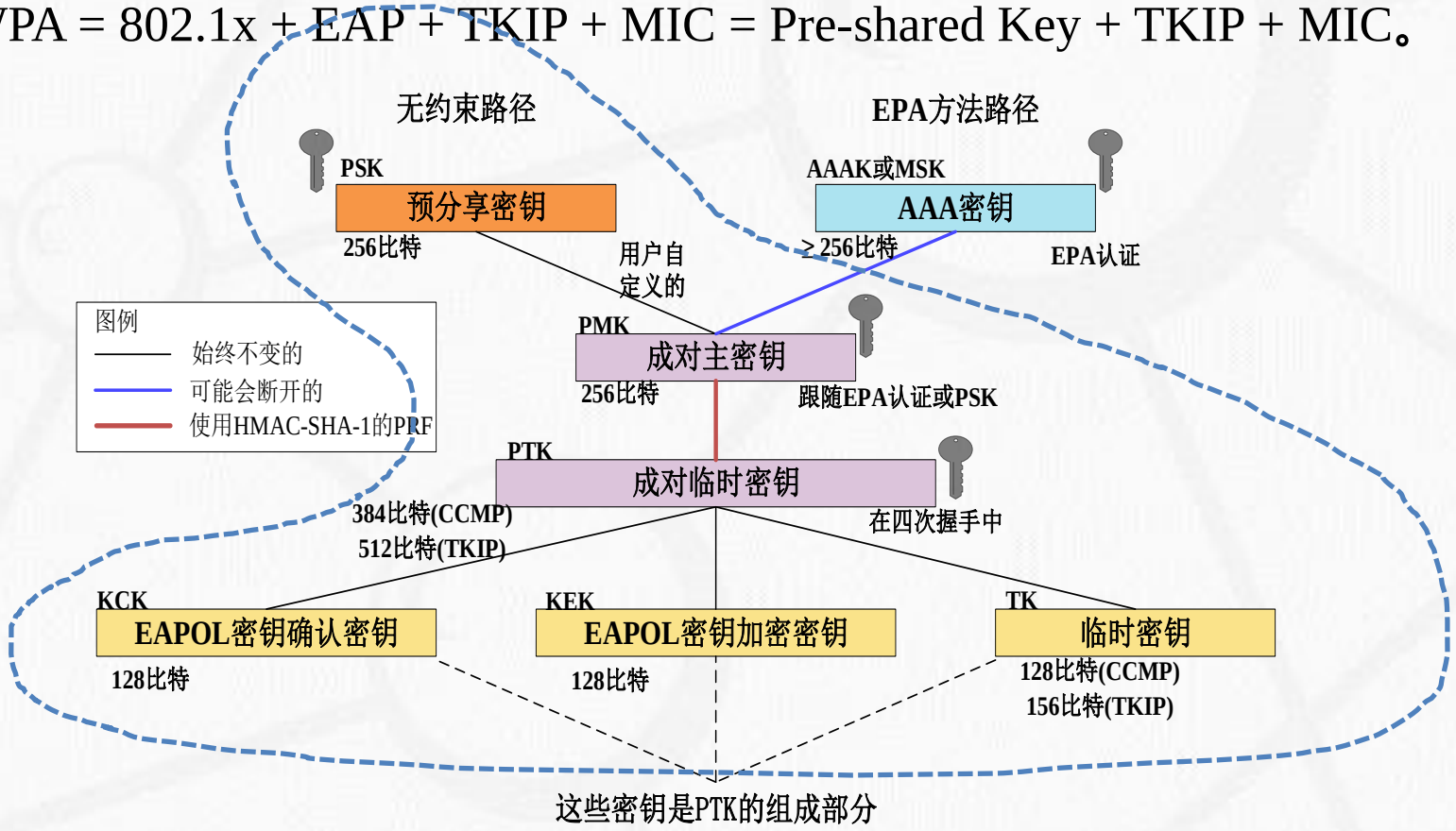
- 存在的风险
  - 重用风险：$c_0 \oplus c_1 = m_0 \oplus m_1$，已知$m_0$可以求得$m_1$。显然不满足IND-CPA安全。
  - 部署风险：实际部署中，IV被初始化为0，造成碰撞概率增大。
  - 空间风险：IV的实际取值空间仅有2^{24}
    - 即使IV是完全随机生成，在短时间内碰撞的概率也很高。
    - 例：粗略计算，AP的发包速度为1000 p/s，也会每4秒碰撞一次。

§3.2 无线网接入安全 - WPA

- WiFi网络安全接入(WPA: WiFi Protected Access)
  - 2003年WPA作为802.11i标准(候选)的子集进行发布
    - 在客户端硬件兼容WEP协议，只需要通过固件更新就能支持WPA；
    - 在接入端需要进行硬件更新；
    - 基于RC4、TKIP、EAPoL实现。
  - 2004年WPA2作为802.11i标准(正式)的子集进行发布
    - 加入AES加密的支持(CBC模式)；
    - 2006年之后生产的设备都支持WPA和WPA2。
  - 2018年1月WiFi联盟宣布WPA3将会取代WPA2
    - 加入AES-256和SHA-384的支持；

§3.2 无线网接入安全 - WPA的密钥管理

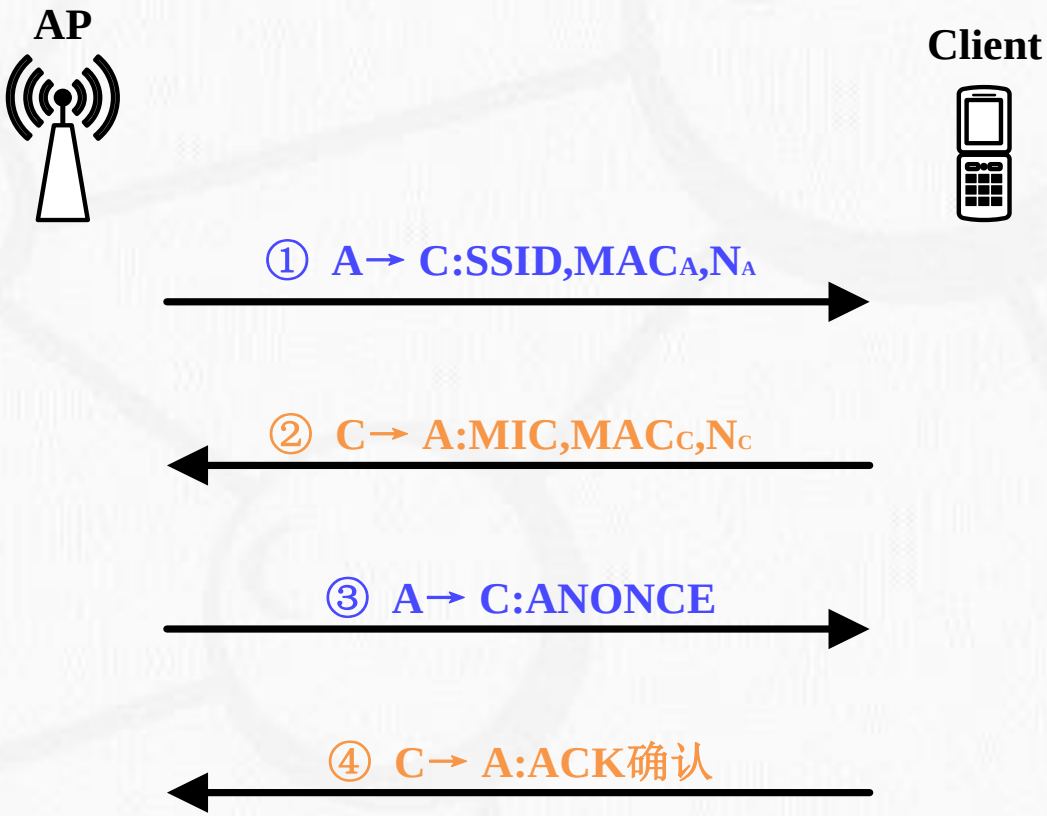
- 针对WPA-PSK模式：
  - PSK作为PMK使用。
  - WPA = 802.1x + EAP + TKIP + MIC = Pre-shared Key + TKIP + MIC。



成对密钥的层次

§3.2 无线网接入安全 - WPA的四次握手过程(1)

- WPA的四次握手过程：
  - 发现认证阶段中确保双方有共同的主密钥PMK；
  - PMK一经生成，双方留存，不用于会话加密。



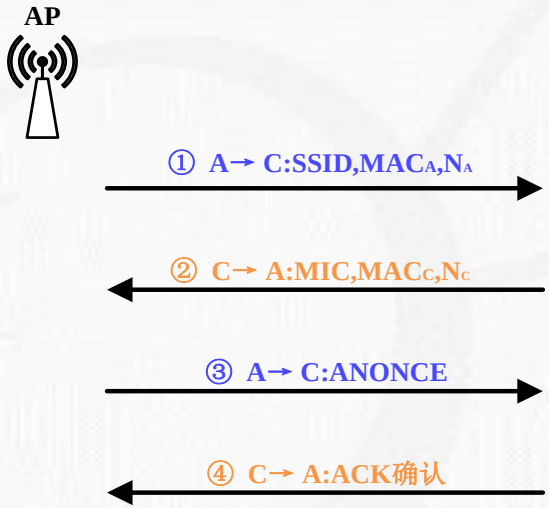
§3.2 无线网接入安全 - WPA的四次握手过程(2)

- 具体过程(概要):

$PSK = PMK = SHA(passphrase, SSID, MAC_A);$

1. A→C:
   - $SSID, MAC_A, \boldsymbol{N_A}$
   - $PSK = SHA(passphrase, SSID, MAC_A);$
2. C→A:
   - $E_{PSK}(MAC_C, N_C)||MIC$
   - $PTK = SHA1\_PRF_{PMK}\ (MAC_A, \boldsymbol{N_A}, MAC_C, \boldsymbol{N_C});$
   - MIC_{KEY}为PTK的前16个字节;
   - $MIC = HMAC(MIC_{KEY});$
3. A→C:
   - $E_{PSK}(\boldsymbol{N_C})$
4. C→A:
   - ACK

预共享密钥

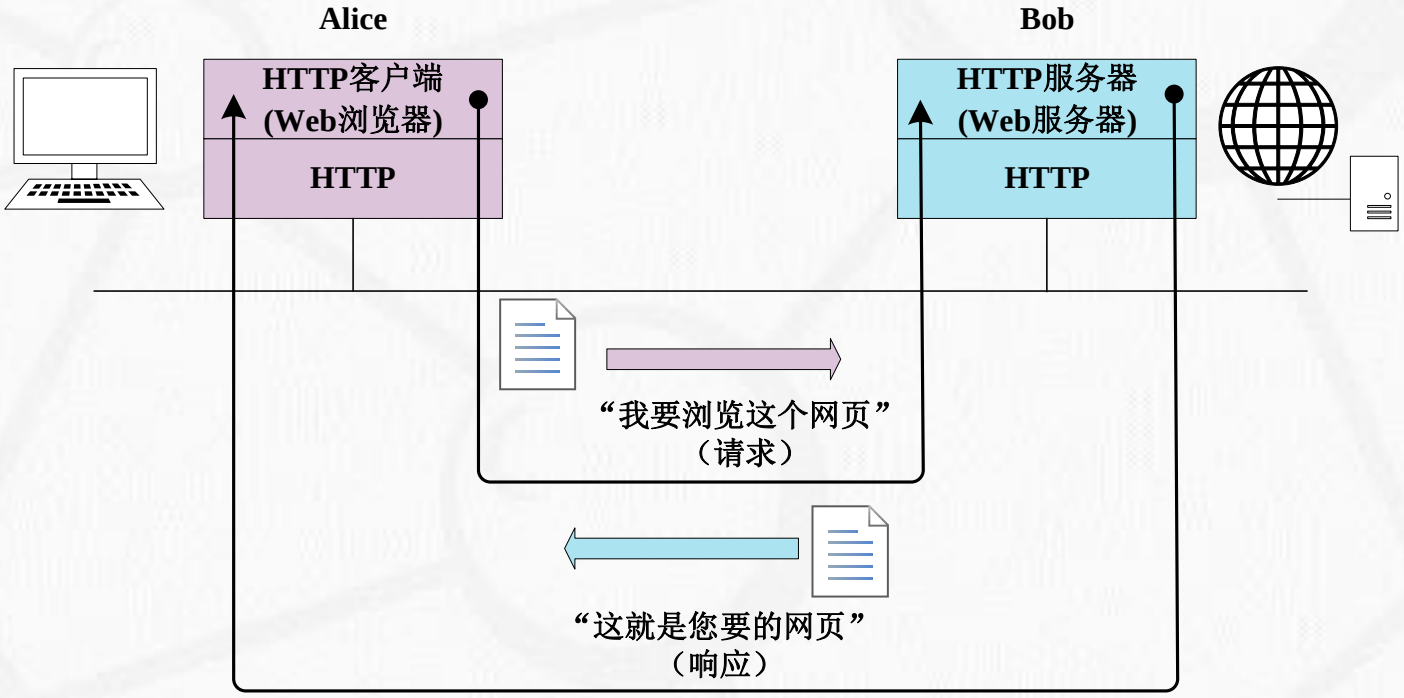


SSL、HTTP

§3.3 HTTPS

§3.3 HTTPS - 超文本传输协议

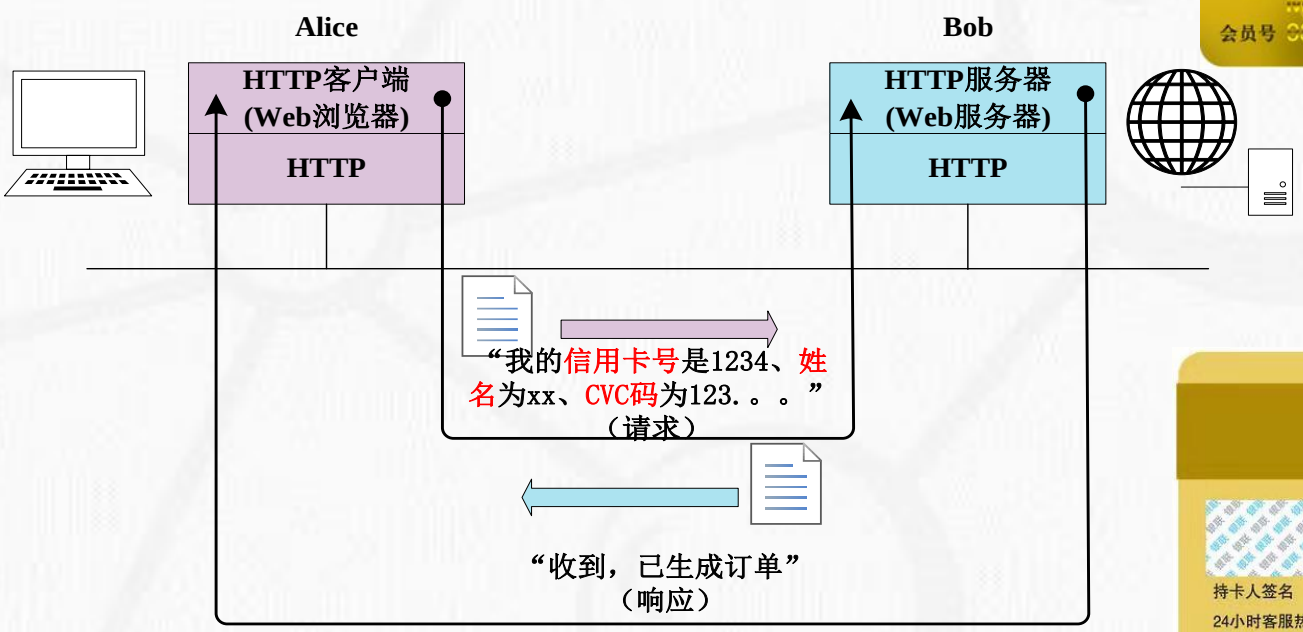
- 超文本传输协议(HTTP: HyperText Transfer Protocol)
  - 基于“请求-响应” 实现；
  - 因特网上应用最为广泛的一种网络传输协议；
  - WWW文件必须遵守这个标准。



Alice的Web浏览器（客户端）和Bob书店的网站（服务器）进行HTTP通信

§3.3 HTTPS - HTTP的缺陷

- HTTP的缺陷：
  - 一切有效载荷全部是明文传输。

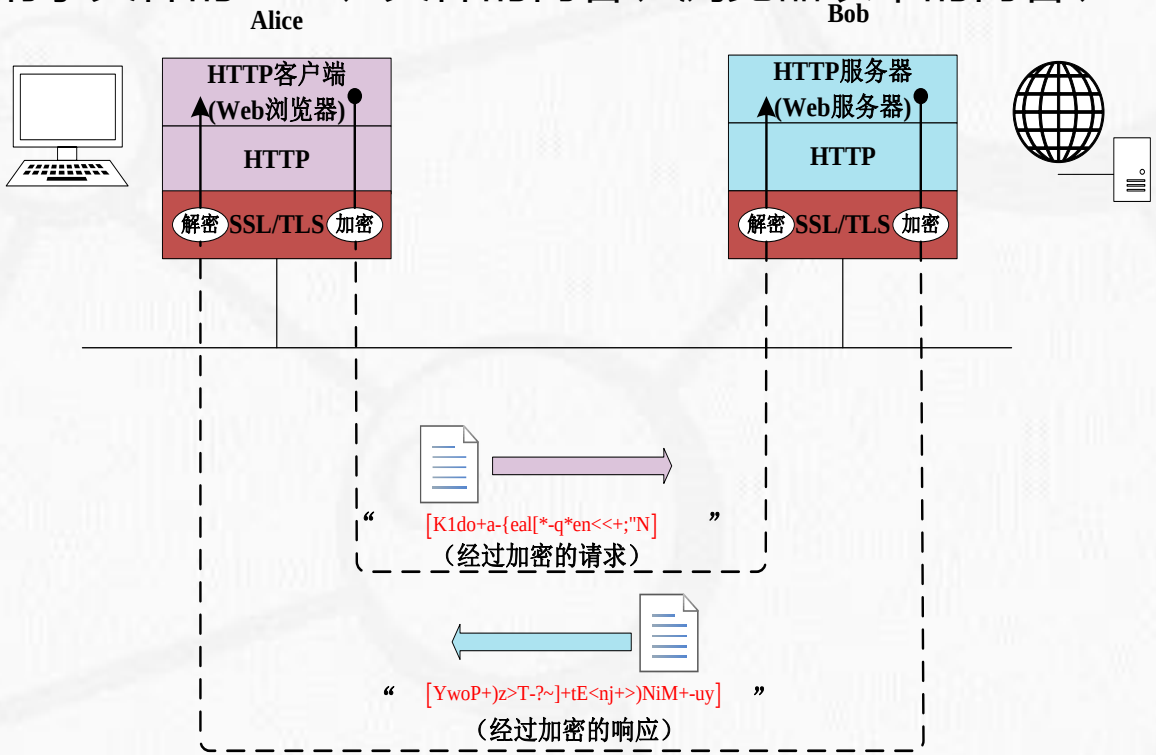


姓名+卡号+CVC码=
可免密支付

不使用TSL/SSL发送信用信息的情形

§3.3 HTTPS - HTTPS的由来

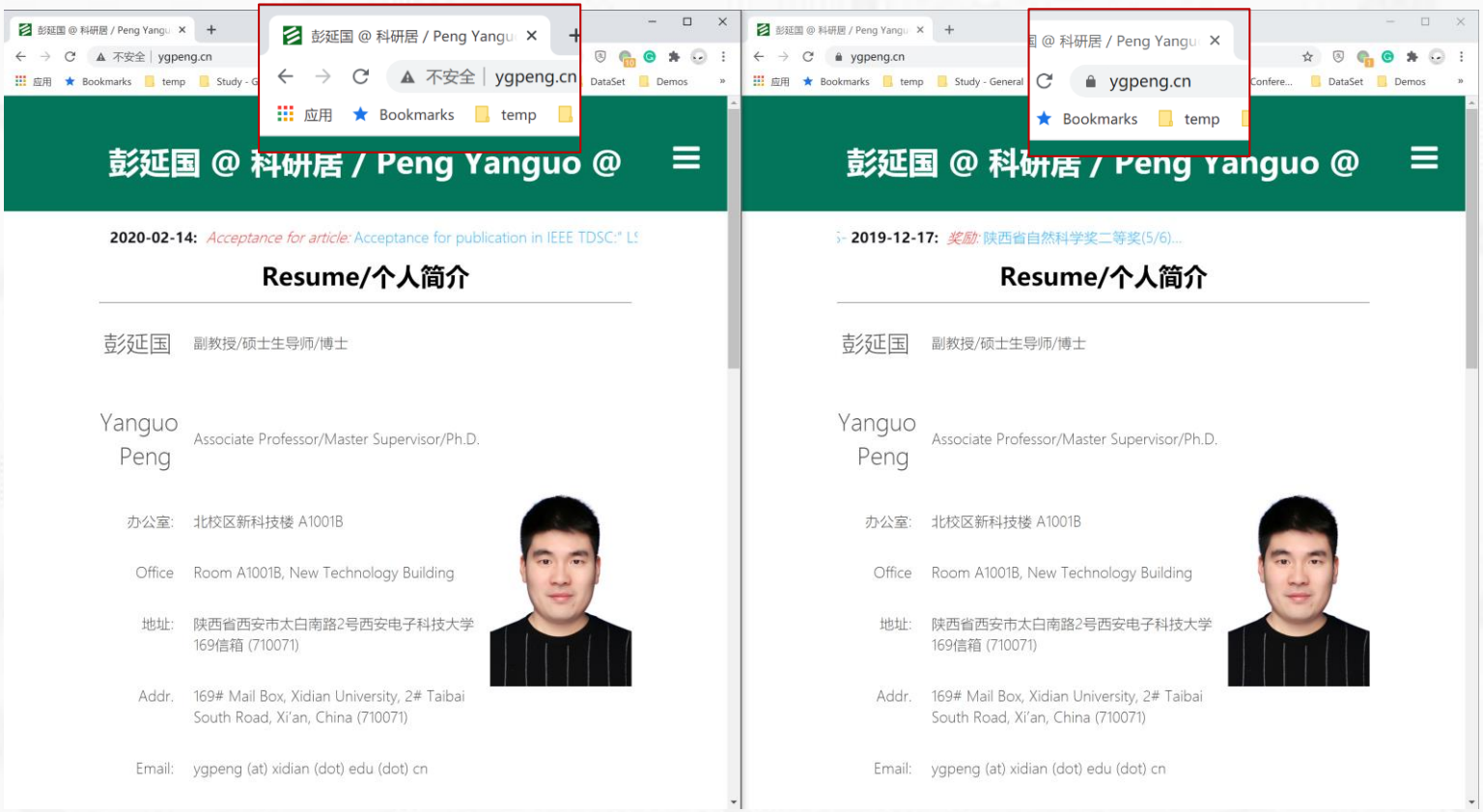
- HTTPS(Hyper Text Transfer Protocol over Secure Socket Layer)
  - 底层基于TLS/SSL协议实现(RFC 2818)。
  - 下列元素被加密：
    - 请求文件的URL、文件的内容、浏览器表单的内容、Cookie、HTTP报头。



使用TSL/SSL承载HTTP，对请求和响应进行加密

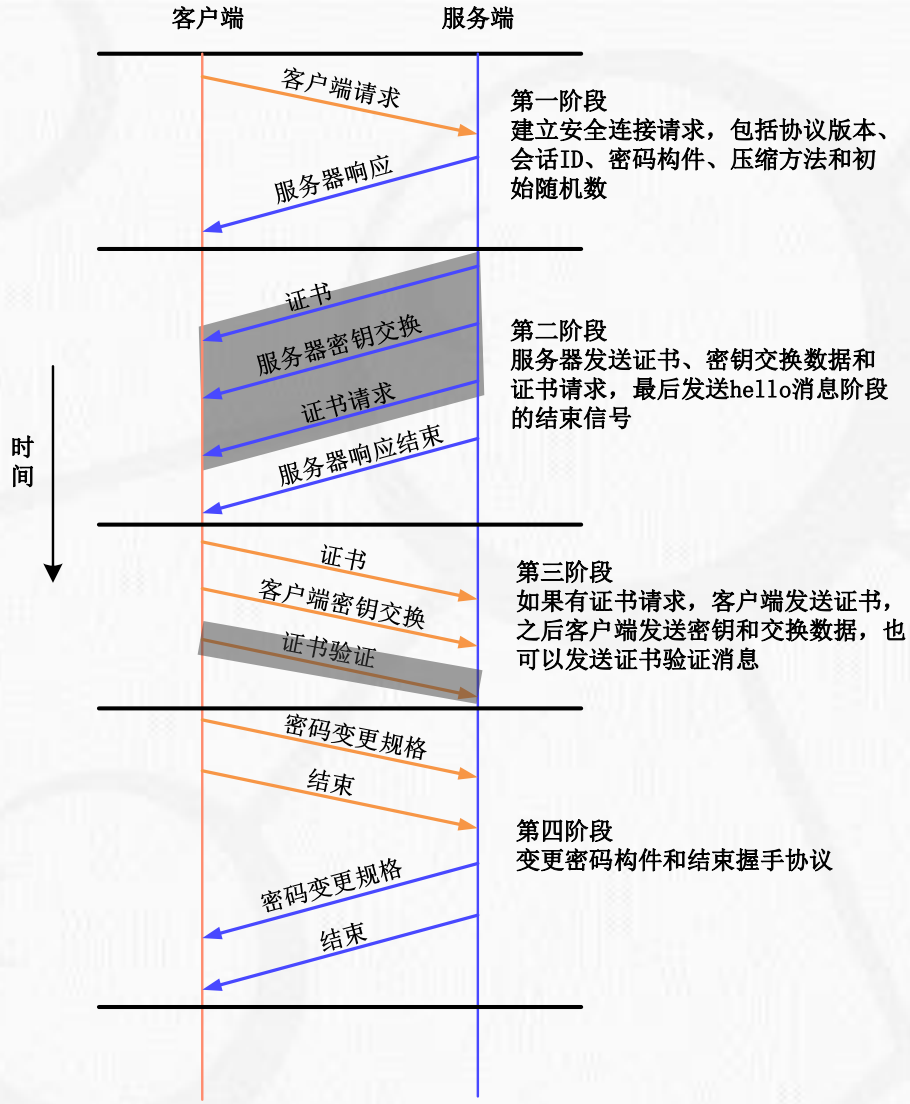
§3.3 HTTPS - HTTPS的效果

- HTTPS的效果：
  - 与HTTP呈现的内容保持一致；
  - 与HTTP内容的格式无关。



§3.3 HTTPS - SSL/TLS的工作原理

- 第一阶段：
  - 建立安全会话
- 第二阶段：
  - 获取有效证书
- 第三阶段：
  - 密钥交换
- 第四阶段：
  - 密钥变更



注：加阴影的传输是可选的，或者是与情况相关的消息，它们并非总会被发送

握手协议过程

隐私保护没有终点

§3.4 浏览器的DNT标准

§3.4 浏览器的DNT标准 – HTTP/HTTPS的特性

- HTTP/HTTPS协议是无状态的
  - 来自客户的每个请求，即使是同一会话中访问相同服务器的请求，也被看作是一个全新的请求。
  - 这给Web应用程序的设计带来了一些麻烦。
- Cookie的引入：
  - 解决HTTP协议无状态的问题。
  - 是存储在客户机上的文本文件。
    - 记录用户的身份信息、配置信息等。
  - 可以是瞬态的，即与一个会话的生存期相同；
  - 也可以是持久的，其生存期可以超过一个会话的生存期。
  - 经常被用于第三方网站追踪。

§3.4 浏览器的DNT标准 - 原理

- DNT标准(Do Not Track)
  - 是一个能避免用户被来自从未访问过的第三方网站跟踪的浏览器功能；
  - 并不采用手段过滤或阻止追踪Cookies；
  - 当用户提出DNT请求时，具有DNT功能的浏览器会在HTTP数据传输中添加一个“头字段”（headers），这个头字段会告诉商业网站的服务器用户不希望被追踪。
    - 1：表示用户不允许被追踪（即选择退出）
    - 0：表示用户允许被追踪（即选择加入）
    - 空值：（即不发送头字段的默认设定）表明用户并未表达出特定的喜好。

§3.4 浏览器的DNT标准 - 历史(概览)

2007年

2007年，民间隐私保护团体要求美国联邦贸易委员会(FTC)为网络广告商列出一份“不要跟踪我”的名单

2009年

2009年，克里斯托弗·索菲安和希德·斯塔姆为火狐浏览器开发了一个插件的雏形，其中第一次用到了Do Not Track头字段。

2010年

2010年7月，美国联邦贸易委员会主席乔恩·莱博维茨在一次有关隐私保护的听证会中告知美国参议院商务委员会他们正在探索这一想法的可行性。

2011年

2011年1月，Mozilla宣布火狐浏览器将提供一个拒绝跟踪的选项；
2011年2月，Opera浏览器发表声明称支持这一功能；
2011年4月，苹果的Safari浏览器增加了支持；
随后，Opera和Chrome正式添加了对DNT的支持。

2012年

2012年6月，微软宣布其DNT选项将在Windows 8系统搭载的IE10中默认开启；
2012年9月，DNT标准的创始者之一Roy Fielding向Apache的HTTP服务器的源代码中加入了补丁，主动忽略了由IE10用户发送的DNT头字段数据；
2012年10月，雅虎声明将忽略来自IE10的DNT请求。

§3.4 浏览器的DNT标准 – DNT对OBA的影响

- DNT会严重影响在线行为广告的运营和收入。

单位：亿美元

序号	公司名	2018年第四季度广告收入	2019年第四季度广告收入	年变化（增长/下降）
1	Alphabet	325.2	379.3	17%
2	Facebook	166.4	207.36	25%
3	Microsoft (Search advertising + LinkedIn)	36.69	42.65	16%
4	Amazon	33.88	47.82	41%
5	Twitter	7.91	8.85	12%
6	Snap*	3.89	5.61	44%
7	eBay (Classifieds)	2.63	2.69	2%
8	Pinterest*	2.73	3.99	46%

备注*：Snap和Pinterest约90%的收入均来源于广告，故此数据为总收入。

MORKETING Global

国内互联网公司

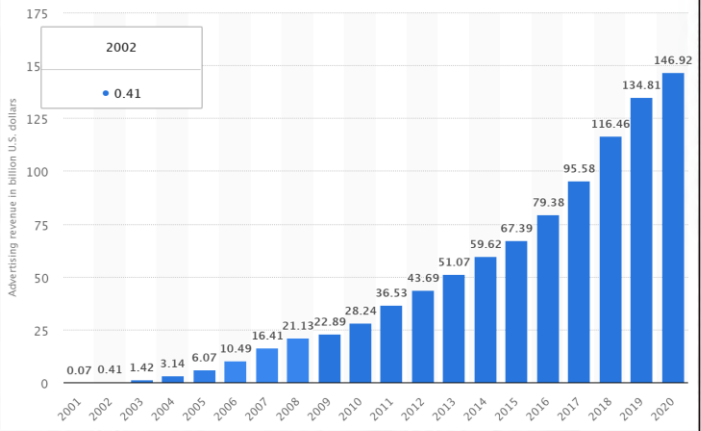
单位 亿元人民币

2020年Q3广告营收情况

序号	公司名	Q3广告营收	同比增速	Q3总营收	Q3广告营收总占比	上半年广告营收	备注
1	阿里巴巴	693.38	20%	1550.59	44.72%	823.4	该数值为客户管理收入，其中包含营销服务和展示广告，淘宝客计划等。
2	腾讯	213.51	16%	1254.47	17.02%	362.65	
3	百度	202.01	14.41%	282.32	71.55%	319.31	百度广告营收中包含爱奇艺广告营收。
4	拼多多	128.78	47.89%	142.1	90.63%	165.47	第三季度为在线营销服务+其他收入
5	京东	124.12	24.29%	1742.14	7.12%	235.8	该数值包含第三方佣金。
6	美团	56.6	28.37%	354.01	16%	71.87	
7	小米	33	13.7%	722	4.57%	58	
8	爱奇艺	18.4	-11%	71.88	25.6%	31.23	
9	搜狗	13.63	-33.2%	15.35	88.84%	33.87	
10	唯品会	10.19	-10.3%	231.8	4.39%	17.27	此项为“其他收入”，主要包含第三方物流、产品推广和在线广告的营收，及第三方商家收取的费用等。
11	汽车之家	9.27	0.03%	23.16	40.05%	14.98	此项为“媒体服务”营收，其中大部分为广告营收。
12	哔哩哔哩	5.58	126%	32.26	17.28%	5.63	
13	搜狐	2.91	-11%	11.18	26.03%	38.37	搜狗成为腾讯全资子公司，第三季财报仅披露了品牌广告收入
14	欢聚时代	2.37	-20.8%	62.86	3.77%	6.25	此项为“其他收入”，包含广告收入。2020年第二季度欢聚时代与虎牙财报拆分。
15	斗鱼	1.98	1.02%	25.47	7.77%	3.53	此项为广告及其他收入。
16	同程艺龙	1.75	-23.6%	19.15	9.13%	1.79	此项为“其他收入”，主要为广告服务收入，配套增值用户服务所得收入及景点门票收入。
17	虎牙	1.58	44.6%	28.15	5.6%	2.69	此项为广告及其他收入。
18	陌陌	0.5	-38%	37.67	1.33%	0.95	
19	蘑菇街	0.18	-71.5%	1.13	15.98%	0.42	
	字节跳动	媒体报道2020年全年字节跳动媒体广告将达1500-1550亿流水收入					

注1：该表格中部分企业未公布人民币单位的营收，故采用北京时间2020年3月20日13：29的美元汇率进行换算（1:7.0826）。
注2：部分企业尚未公布财报数据，因此目前暂不加入榜单。
注3：该表格中的各公司营收、广告占比数值，部分公司同比增速为四舍五入（仅保留小数点后两位），以及汇率波动，将会存在少量误差，仅供参考

MORKETING



Google的广告营收

§3.4 浏览器的DNT标准 - DNT的困境

- 理论上：
  - 那些支持DNT的浏览器应该遵从协议。
- 实际上：
  - DNT并不是一个技术协议，而是一个政策协议，实际约束力有限。
  - 这个协议只是说Don't track，但是对于具体的track的方法和范围，目前并没有清晰的界定。
  - 到现在为止，都没有一个统一的标准来规定什么样的HTTP response可以体现一个server/tracker尊重DNT。

§3.4 浏览器的DNT标准 - 隐私獾

- 隐私獾(Privacy Badger)：
  - 是由电子前哨基金会开发的适用于Google Chrome与Firefox浏览器的隐私保护扩展，其用于阻止那些不遵守DNT协议的广告商跟踪行为。
  - 2017年4月，隐私獾的用户已经达到100万。

- 自动检测潜在的跟踪器并阻止跟踪行为：
  - 绿色：有第三方资源，但未检测出跟踪行为，不做任何拦截。
  - 黄色：有第三方资源，并且检测出跟踪行为，但为避免网页显示异常，没有完全拦截跟踪器，但跟踪性Cookie将会被拦截。
  - 红色：有第三方资源，并且检测出跟踪行为，彻底拦截跟踪器与跟踪Cookie。

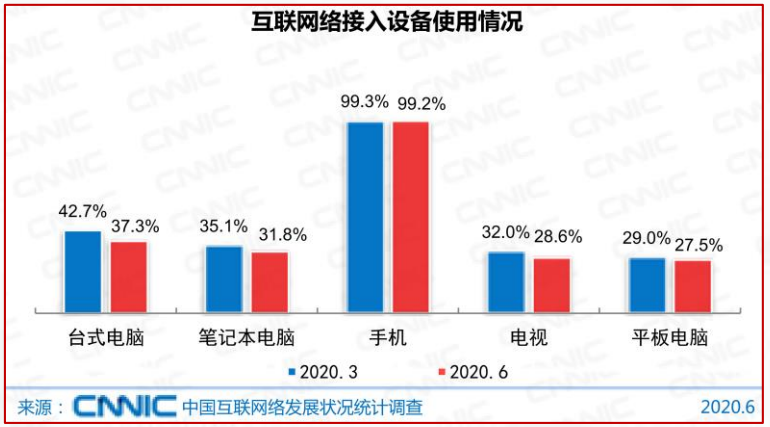
西安電子科技大學
XIDIAN UNIVERSITY

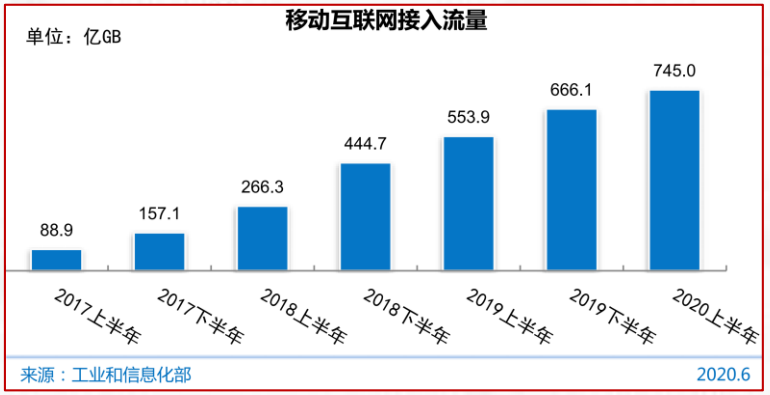
隐私保护没有终点

§3.5 移动终端的隐私保护

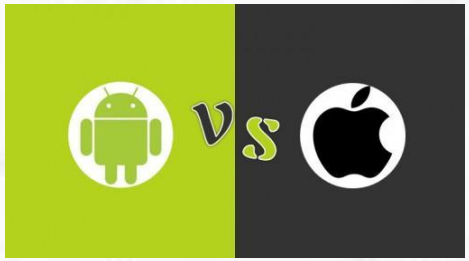
§3.5 移动终端的隐私保护 - 背景

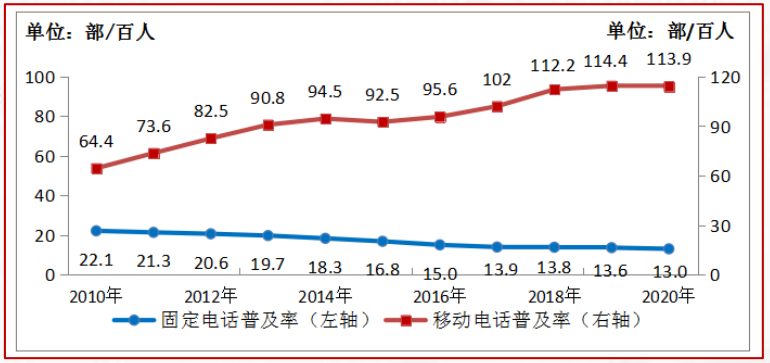
- 移动终端的快速发展
  - 大于99%的流量来自移动端；
  - 2020年移动终端数量相较于2010年翻一番；
  - 2020年移动流量总量相较于2018年翻一番；





- 移动终端的两大阵营
  - 苹果：IOS
  - 谷歌：Android

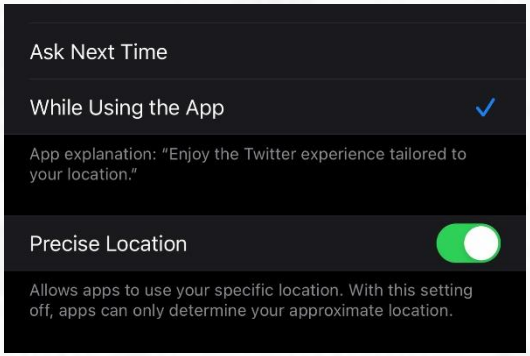




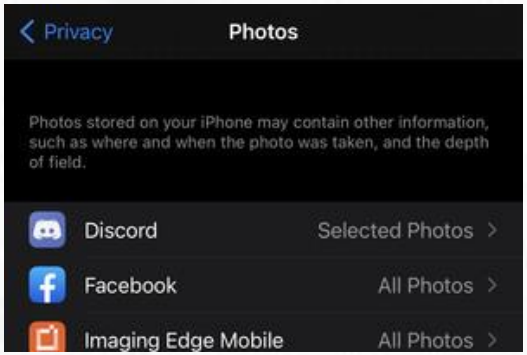
§3.5 移动终端的隐私保护 - IOS端的隐私保护

- 2020年9月17日，苹果公司发布IOS 14正式版。
  - 集成“应用追踪透明度(App Tracking Transparency)”措施：必须通过应用追踪透明化 API 获得用户的明确许可，才能追踪他们的活动。了解更多关于追踪的信息。
  - 隐私保护新功能：
    - 摄录提示灯
    - 近似位置
    - 严格的图像访问
    - 不允许APP追踪
    - 访问剪贴板提示
    - ……

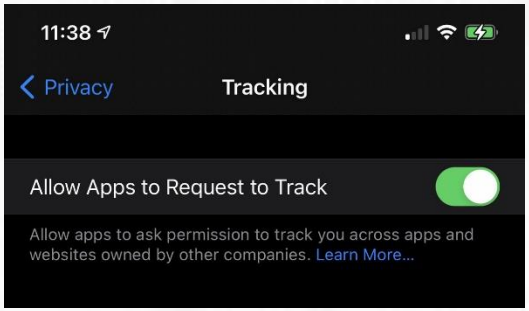
摄录提示灯



近似位置



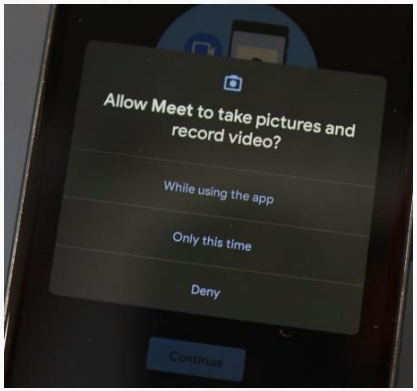
严格的图像访问



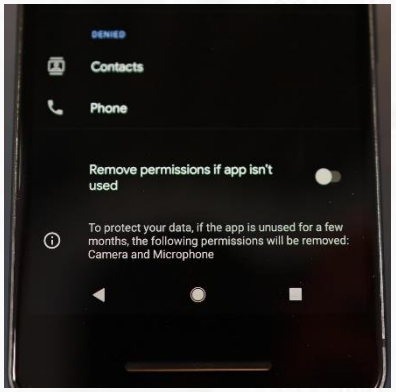
不允许APP追踪

§3.5 移动终端的隐私保护 - Android的隐私保护

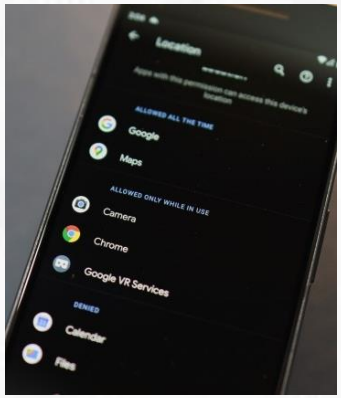
- 2020年9月10日，谷歌公司发布Android 11正式版。
  - 隐私保护新功能：
    - 强制执行分区存储机制
    - 单次授权
    - 自动重置权限
    - 后台位置信息访问权限
    - 前台服务
    - ……



单次授权

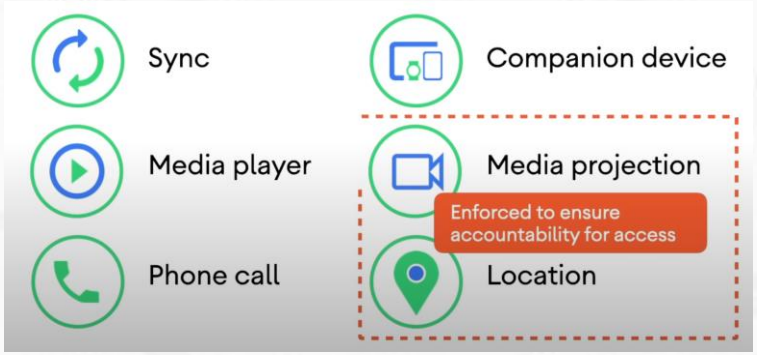


自动权限重置



后台位置信息访问权限

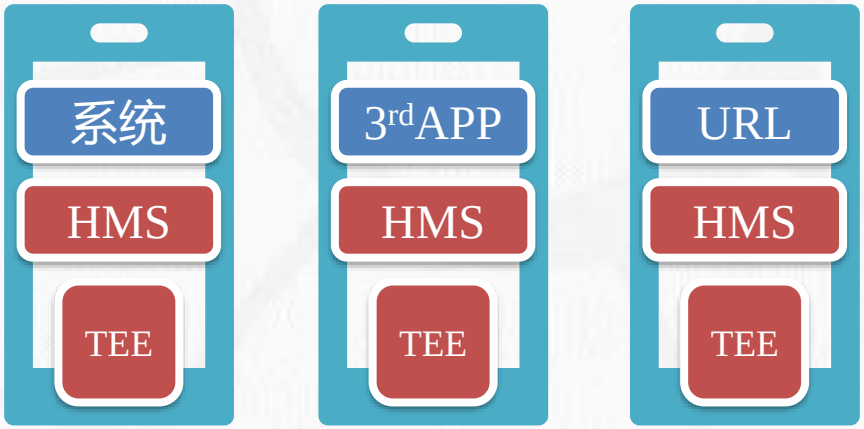
强制分区存储



前台服务

§3.5 移动终端的隐私保护 - EMUI的隐私保护

- 2020年9月15日，华为发布了EMUI 11的正式版。
  - 华为Safety Detect服务：包括系统完整性检测、应用安全检测、恶意URL检测、虚假用户检测能力，发挥华为手机独特优势，助力开发者快速构建应用安全。
  - 隐私保护新功能：
    - 分享隐私保护
    - 单次授权
    - 隐私备忘录
    - ……



§3.5 移动终端的隐私保护 - MIUI的隐私保护

- 2020年12月28日，小米发布MIUI 12.5正式版本。
  - 隐私保护新功能：
    - 剪贴板隐私保护
    - 差分位置隐私
    - 沙盒机制
    - URL隐私保护
    - 后台相机禁止使用
    - ……

剪贴板隐私保护

差分位置隐私

URL隐私保护

后台禁止相机

§3 小结

- 内容回顾
  - 数据采集过程中存在的安全和隐私风险
  - WEP、WPA的基本工作流程
  - HTTPS的工作原理
  - 移动端隐私保护的进展

- 掌握
  - WEP的工作流程和可能存在的安全威胁
  - HTTPS的工作原理

Thanks!
Questions & Advices!