CSE 486/586分布式系统
案例研究。Facebook照片商店

史蒂夫-高
计算机科学与工程
布法罗大学

设计一个系统

- 一般来说，当你设计一个系统时，你需要了解你的工作负荷。
  - 并根据工作量来设计你的系统
  - (也许在开始时没有，因为没有工作量)
- 工程原理
  - 让常见的情况快速，让罕见的情况正确
  - (摘自帕特森和轩尼诗的书)
  - 这一原则贯穿了几代人的系统。
- 例子？
  - 缓存
- 了解常见案例==了解你的工作量
  - 例如，阅读占主导地位？写为主？混合型？
- 我们来看看Facebook的例子。

Facebook的工作量

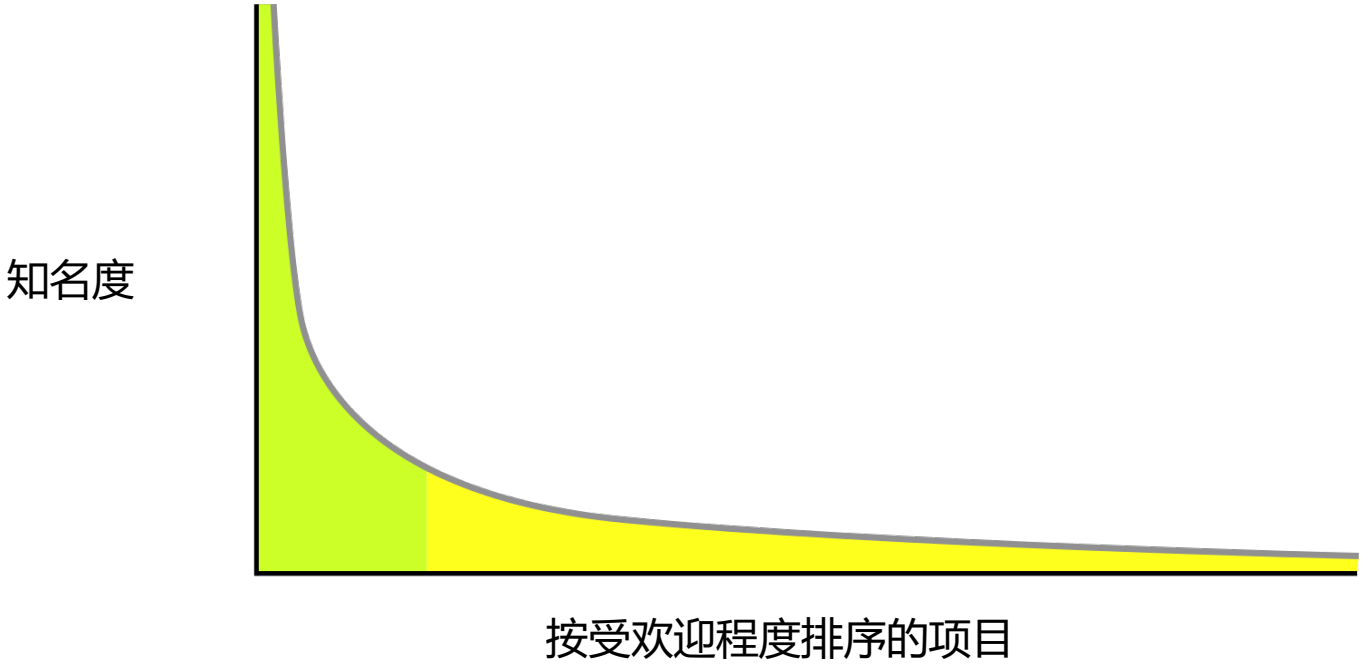
- 你在Facebook上最常做的事情是什么？
  - 读/写墙上的帖子/评论/喜欢
  - 查看/上传照片
  - 其特点非常不同
- 读/写墙上的帖子/评论/喜欢
  - 读取和写入的混合，因此在一致性方面需要更加小心。
  - 但体积小，所以可能对性能不太敏感
- 照片
  - 一次写入，多次读取，因此在一致性方面不需要太多关注
  - 但体积大，所以对性能更敏感

脸书照片的工作量

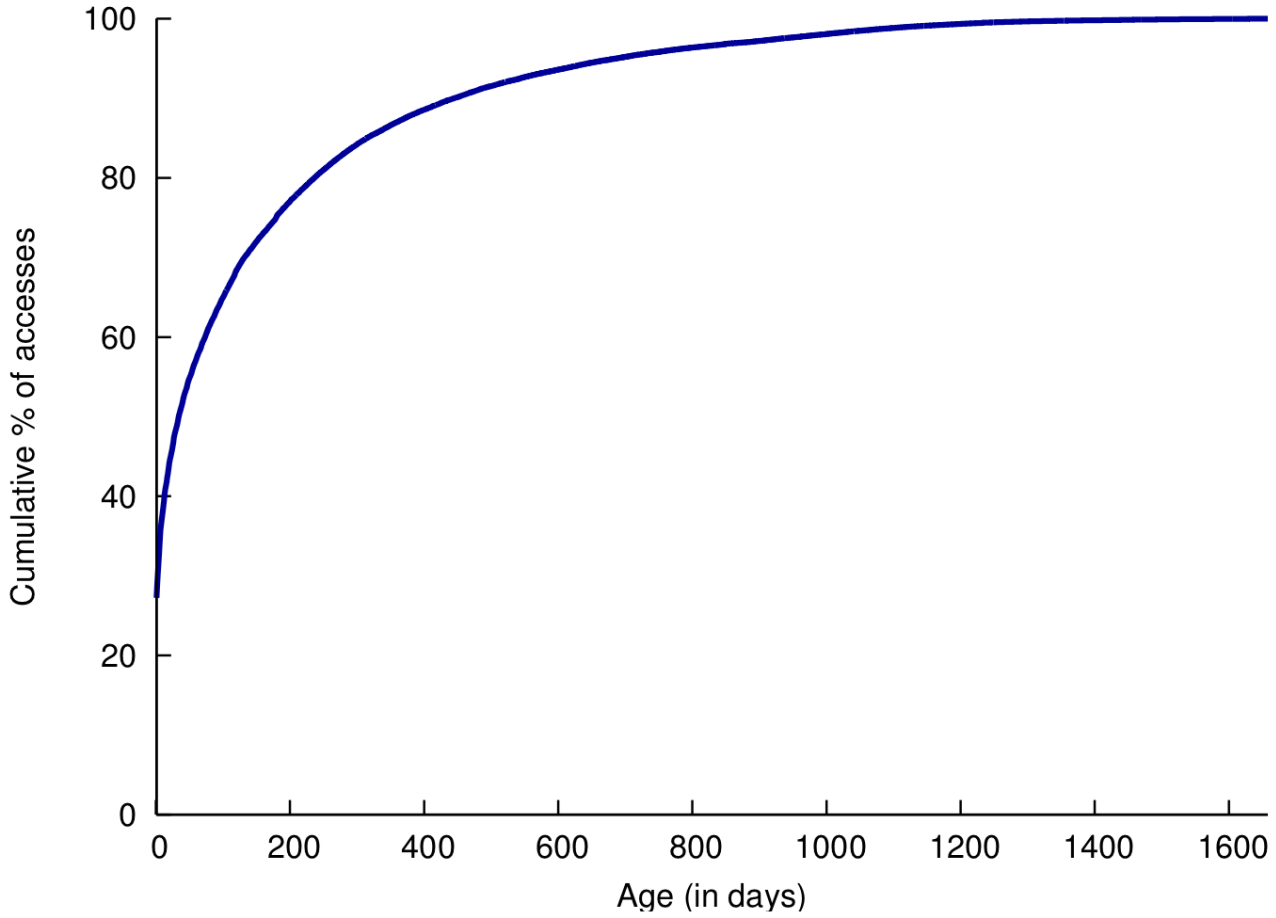
- (这是从2010年开始的)。
- 2600亿张图片(~20PB)。
- 每周10亿张新照片(约60TB)。
- 峰值时每秒一百万次的图像浏览
- 两个特点。Facebook分析了他们的照片工作量，发现了两个特点。
  - 流行分布遵循Zipf。
  - 随着时间的推移，照片的受欢迎程度会发生变化，因为照片会"老化"。

Zipf分布

- 基于幂律
- 对很多自然现象进行建模
- 社会图谱、媒体知名度、财富分布等。
- 也有很多网络内容。

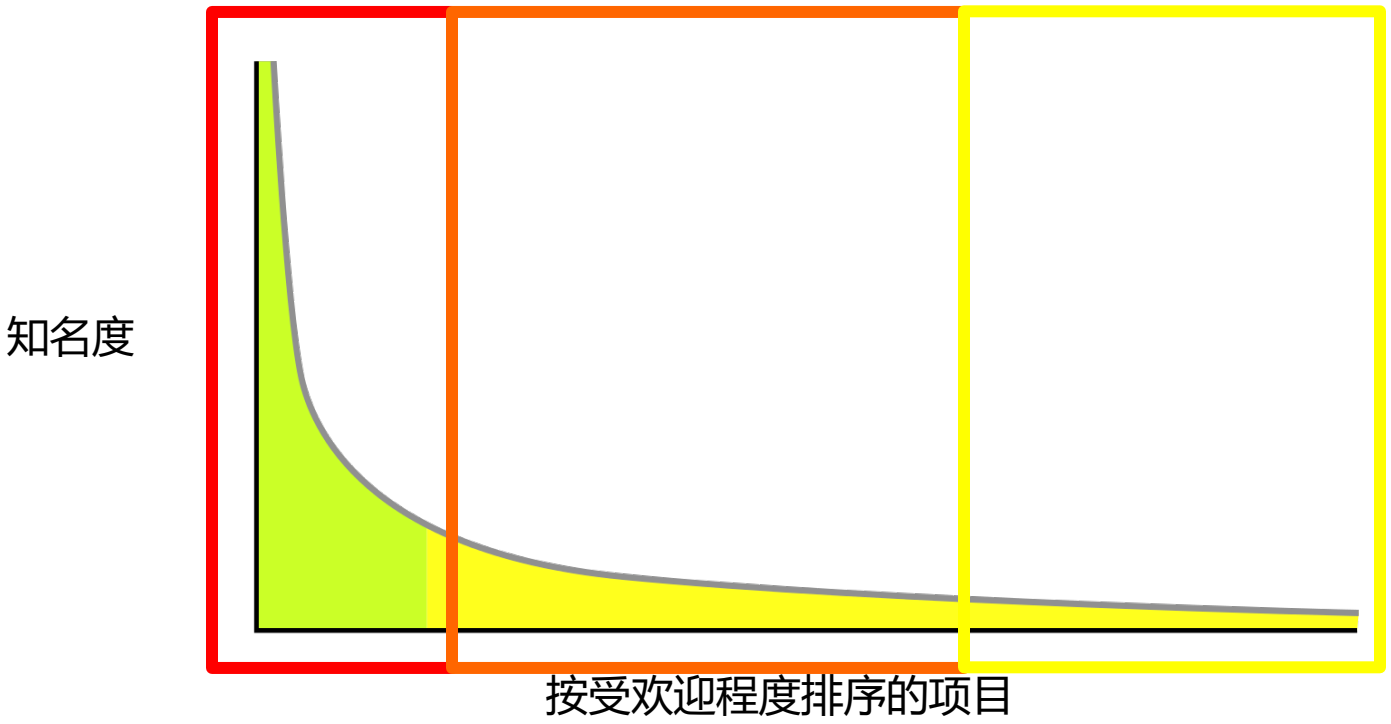


人气随年龄增长而增加


100
80
60
40
20
0
Cumulative % of accesses
0
200
400
600
800
1000
1200
1400
1600
Age (in days)

脸书照片分发

- "热 "与 "暖 "与 "冷 "的照片
  - 热门：受欢迎，浏览量大（约占浏览量的90%）。
  - 温暖。有点流行，但总的来说还是有很多意见
  - 寒冷。不受欢迎的，偶尔的观点



处理不同类型的照片

- 热门照片
  - Facebook使用CDN（内容分发网络）来处理这些问题。
  - 性能非常好，但没有可靠性保证
  - CDN是一个缓存，而不是一个永久存储。
- 温馨的照片
  - Facebook已经设计了自己的存储，名为Haystack。
  - 兼顾性能和可靠性
- 寒冷的照片
  - Facebook已经设计了一个名为f4的 "档案 "存储。
  - 在存储复制的照片时，以存储效率为目标（但不是高性能）。

CSE 486/586 行政管理学

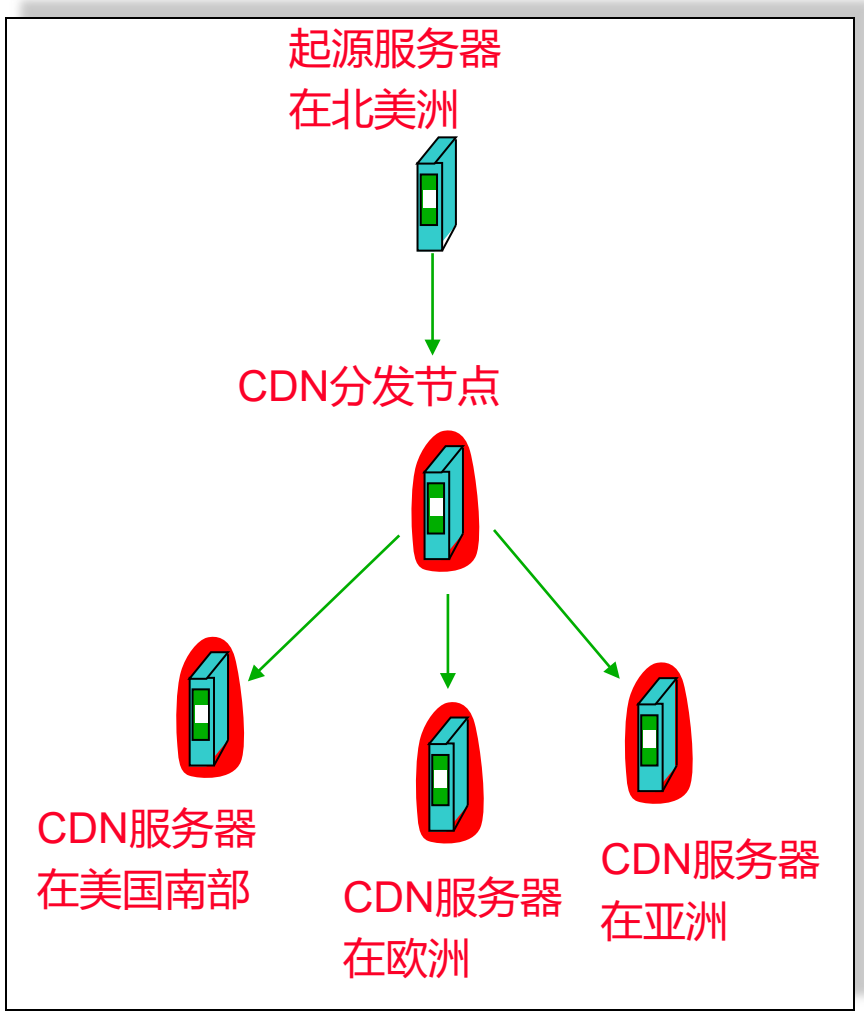
- PA4截止日期：5/10
- 调查和课程评估
  - 调查：https://forms.gle/eg1wHN2G8S6GVz3e9
  - 课程评价：https://www.smartevals.com/login.aspx?s=buffalo
- 如果两者都有80%或更多的参与。
  - 对于你们每个人来说，我会在期中考试和期末考试之间选取较好的一个，并给较好的一个30%的权重，给另一个20%的权重。
  - (目前，期中考试为20%，期末考试为30%)。
- 本周没有背诵；用办公时间代替

热门照片的CDN

- 内容提供商是CDN客户

<u>内容复制</u>

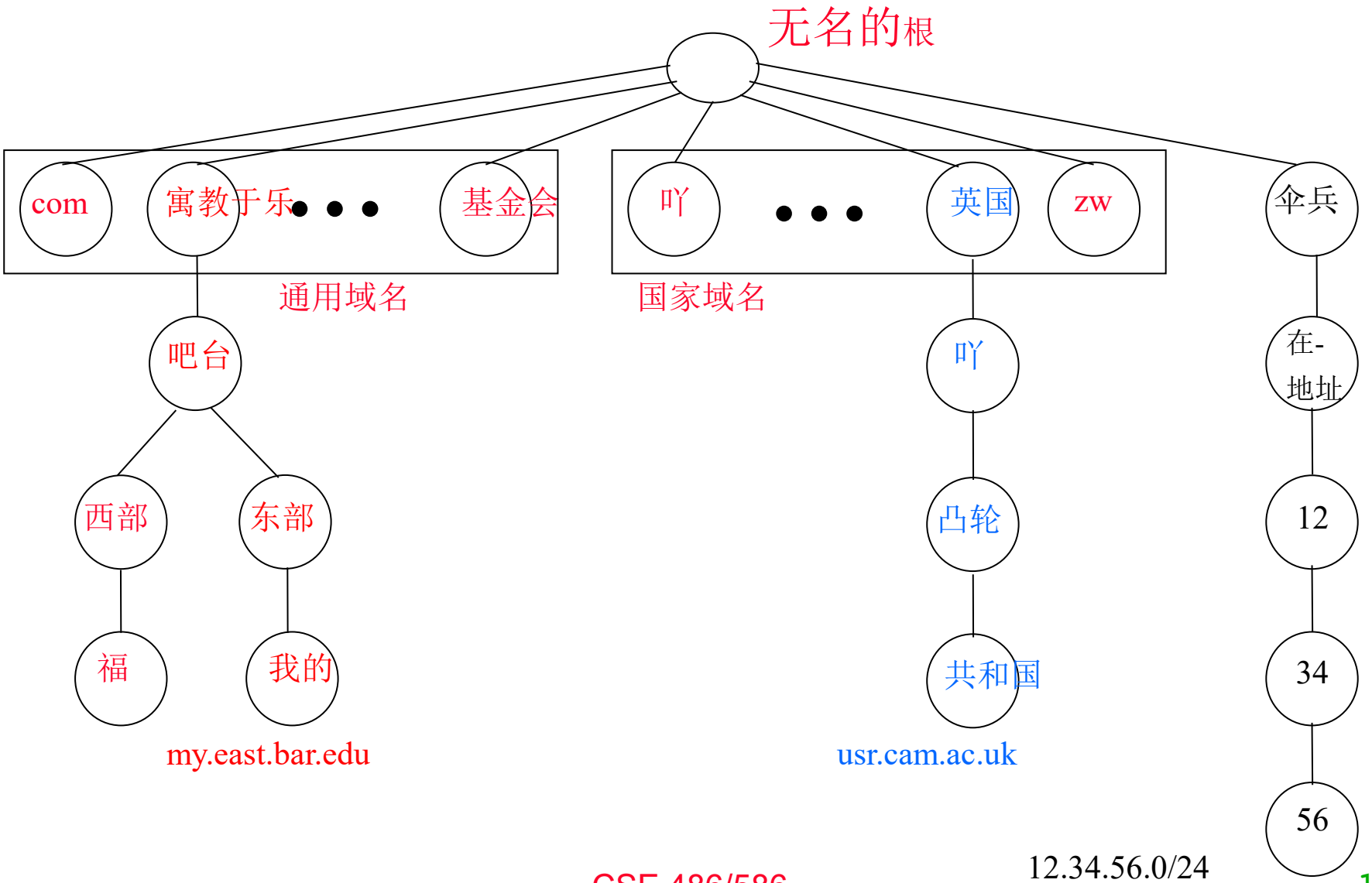
- CDN公司（如Akamai）在整个互联网上安装了成千上万的服务器
  - 在靠近用户的大型数据中心内
- CDN复制了客户的内容
- 当供应商更新内容时，CDN会更新服务器



域名系统

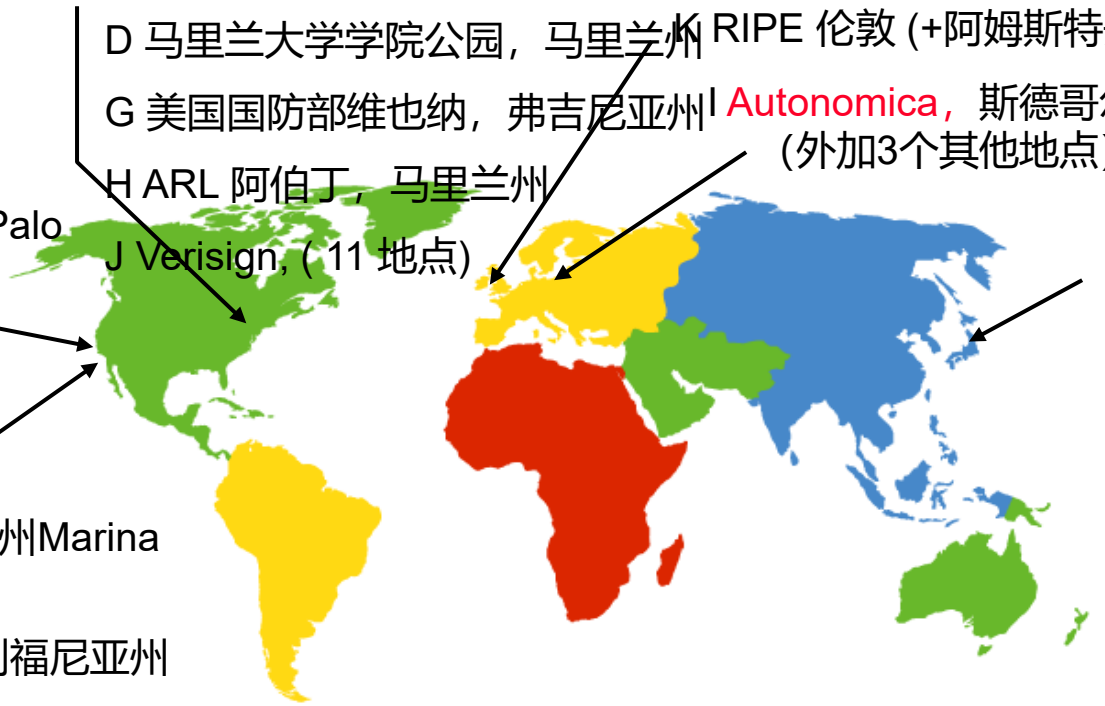
- 对于一个给定的用户，如何定位一个接近的服务器？
- 许多CDN依赖于域名系统(DNS)。
  - DNS将一个DNS名称映射到一个IP地址或另一个DNS名称(别名)。
  - 例如，www.cse. buffalo.edu
    - » 域名：每个顶级域名的注册商
    - » 主机名称：本地管理员分配给每个主机的名称
- DNS的属性
  - 层次化的名称空间
  - 分布在一系列的DNS服务器上
- DNS服务器的层次结构
  - 根部服务器
  - 顶级域名(TLD)服务器
  - 权威的DNS服务器

分布式层次结构数据库


无名的根
com
寓教于乐
基金会
叮
英国
zw
伞兵
通用域名
国家域名
吧台
叮
在-
地址
西部
东部
凸轮
12
福
我的
共和国
34
my.east.bar.edu
usr.cam.ac.uk
56
12.34.56.0/24

DNS根服务器

- 由12个独立的根服务器运营商运营的1088个实例（见http://www.root-servers.org/）。
- 标记为A至M

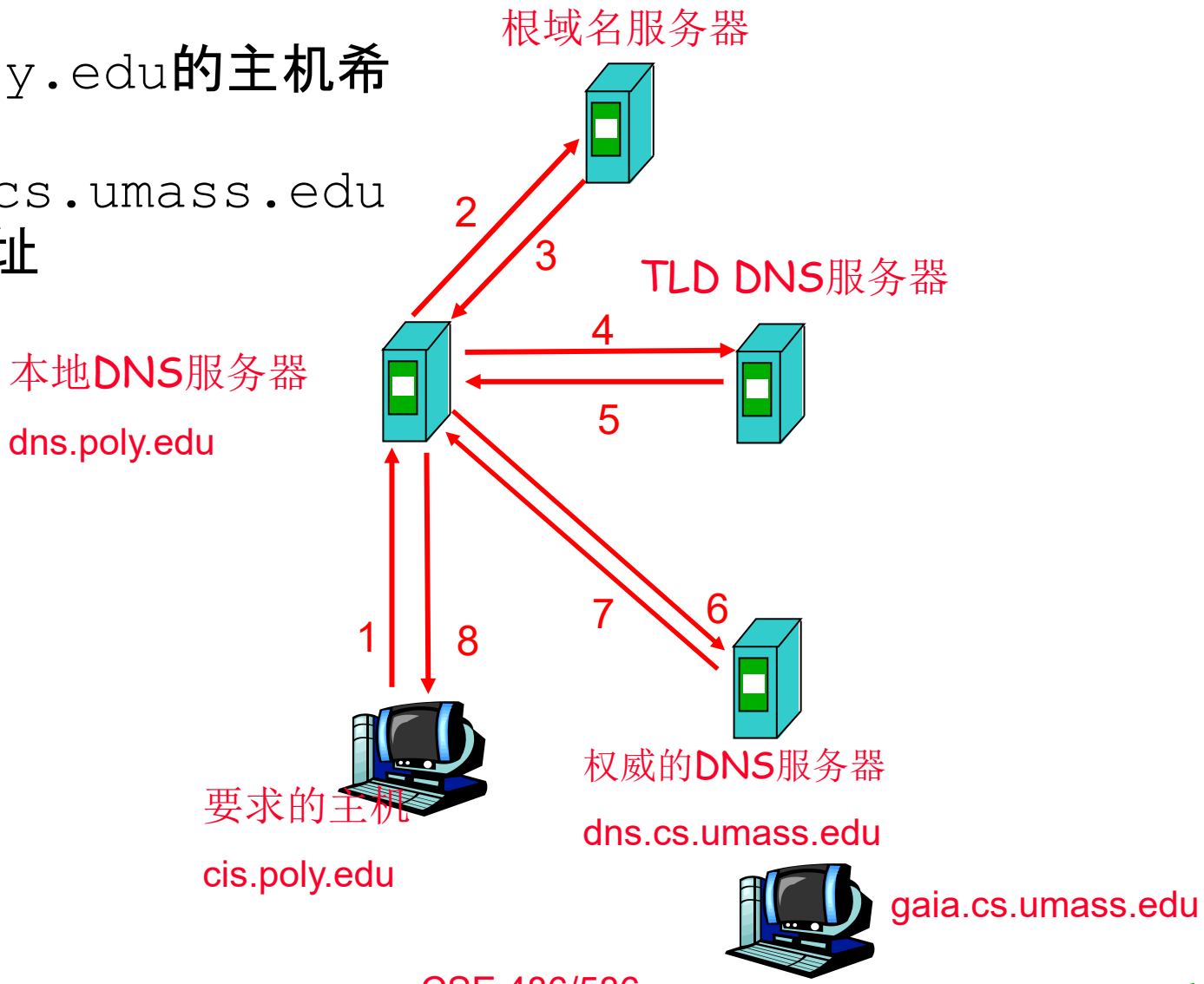


顶级域名和权威性DNS服务器

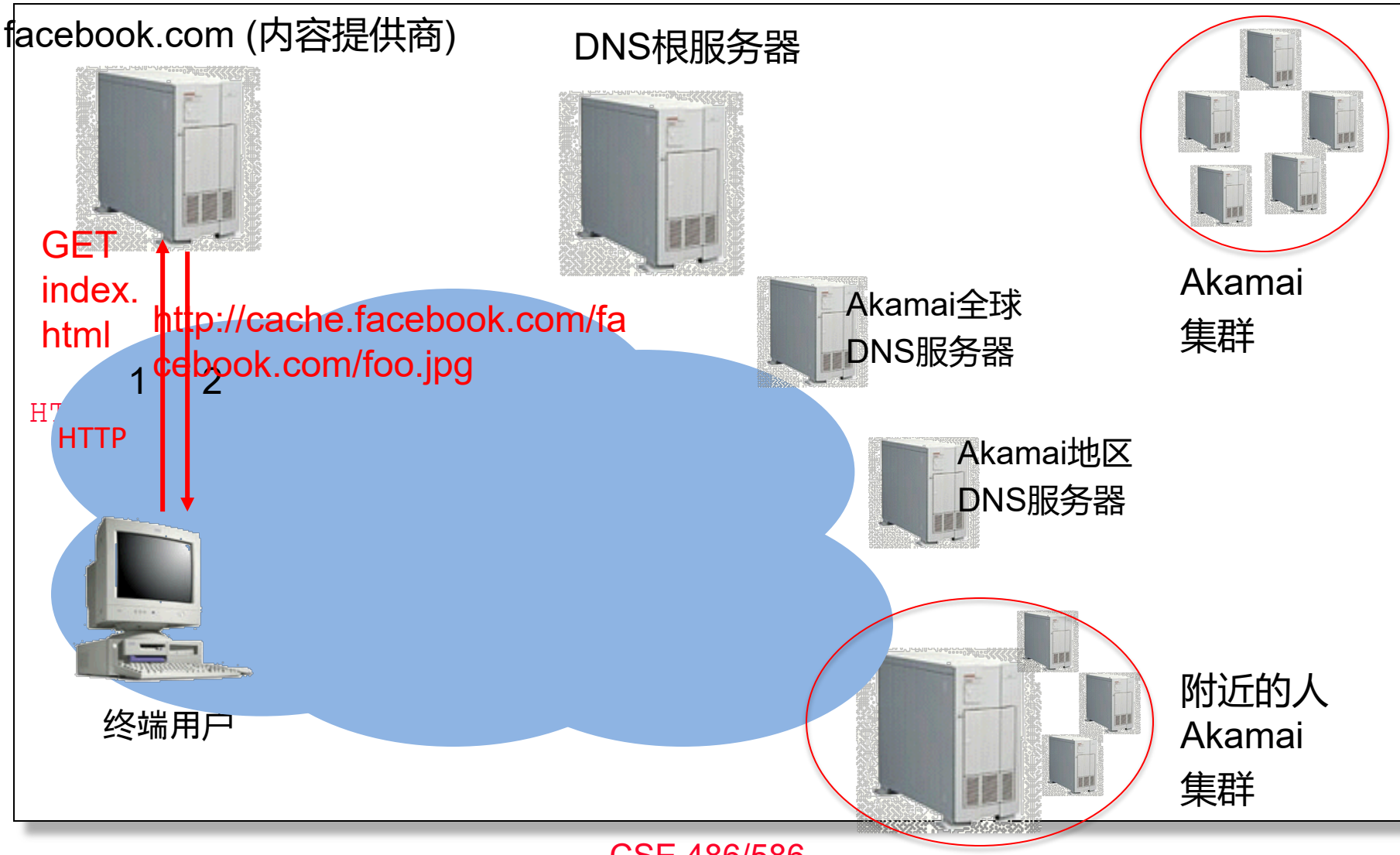
- 顶级域名（TLD）服务器
  - 通用域名（如com、org、edu）。
  - 国家域名（例如，英国、法国、加拿大、日本）。
  - 通常以专业方式管理
    - » 网络解决方案为 "com "维护服务器
    - » 教育机构为 "edu "维护服务器
- 权威的DNS服务器
  - 为一个组织的主人提供公共记录
  - 对于组织的服务器（例如，网络和邮件）来说
  - 可以在本地或由服务提供商维护

例子

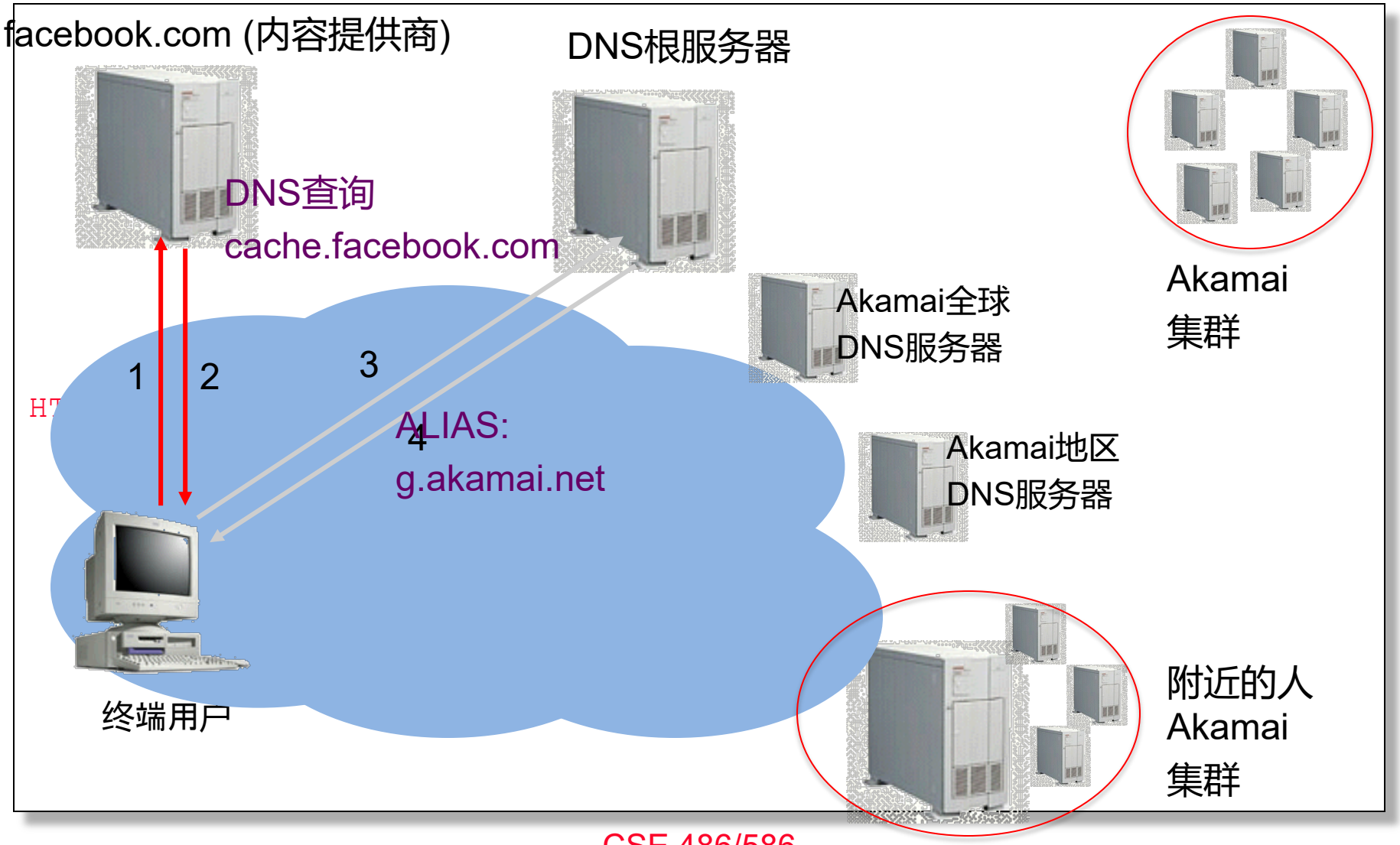
cis.poly.edu的主机希望得到gaia.cs.umass.edu的IP地址



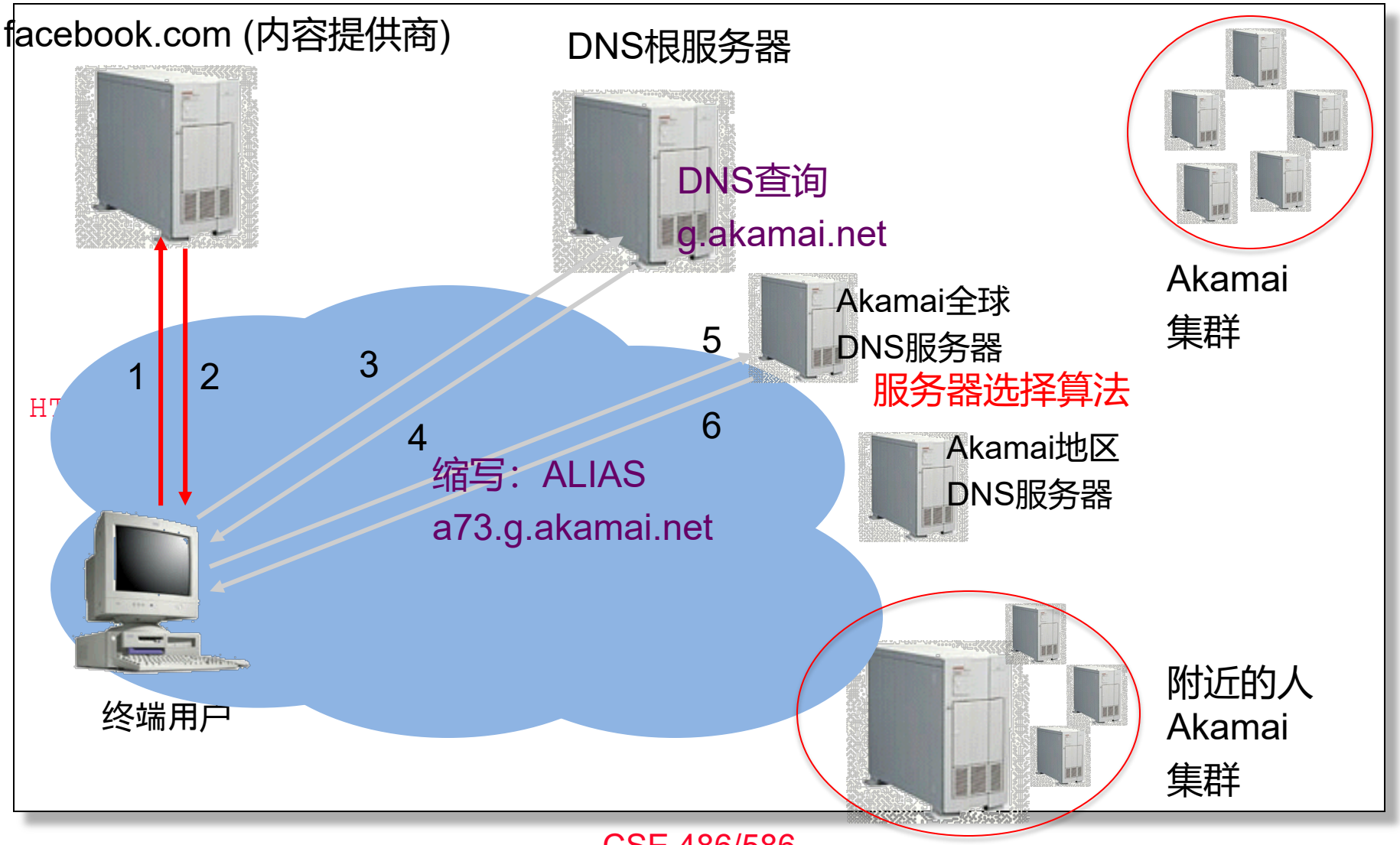
CDN如何工作


facebook.com (内容提供商)
DNS根服务器
Akamai
集群
GET
index.
html
http://cache.facebook.com/fa
cebook.com/foo.jpg
1
2
HTTP
Akamai全球
DNS服务器
Akamai地区
DNS服务器
终端用户
附近的人
Akamai
集群

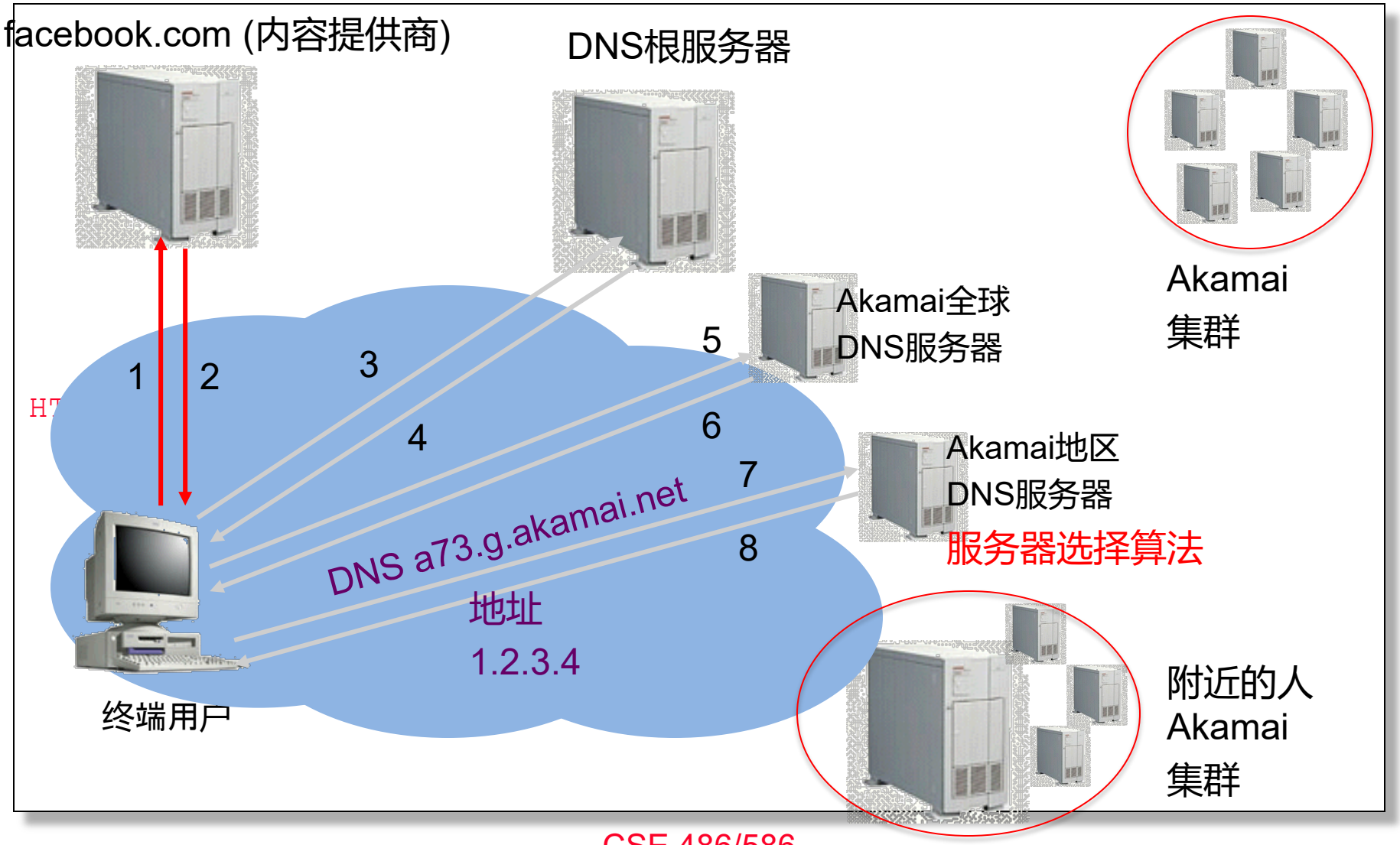
CDN如何工作


facebook.com (内容提供商)
DNS根服务器
DNS查询
cache.facebook.com
1
2
3
4
ALIAS:
g.akamai.net
Akamai全球
DNS服务器
Akamai
集群
Akamai地区
DNS服务器
终端用户
附近的人
Akamai
集群

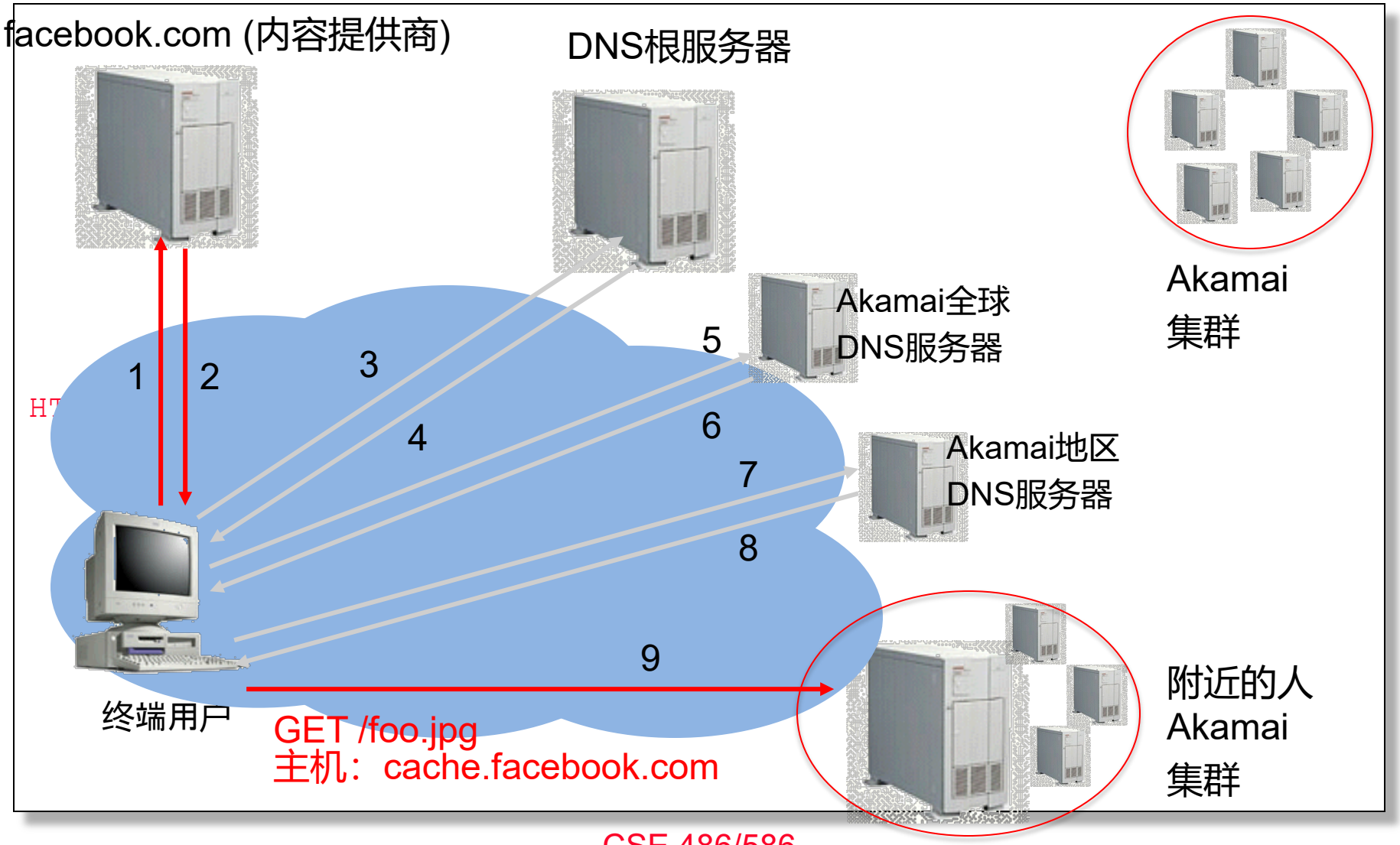
CDN如何工作


facebook.com (内容提供商)
DNS根服务器
DNS查询
g.akamai.net
Akamai
集群
Akamai全球
DNS服务器
服务器选择算法
Akamai地区
DNS服务器
1
2
3
4
5
6
缩写：ALIAS
a73.g.akamai.net
终端用户
附近的人
Akamai
集群

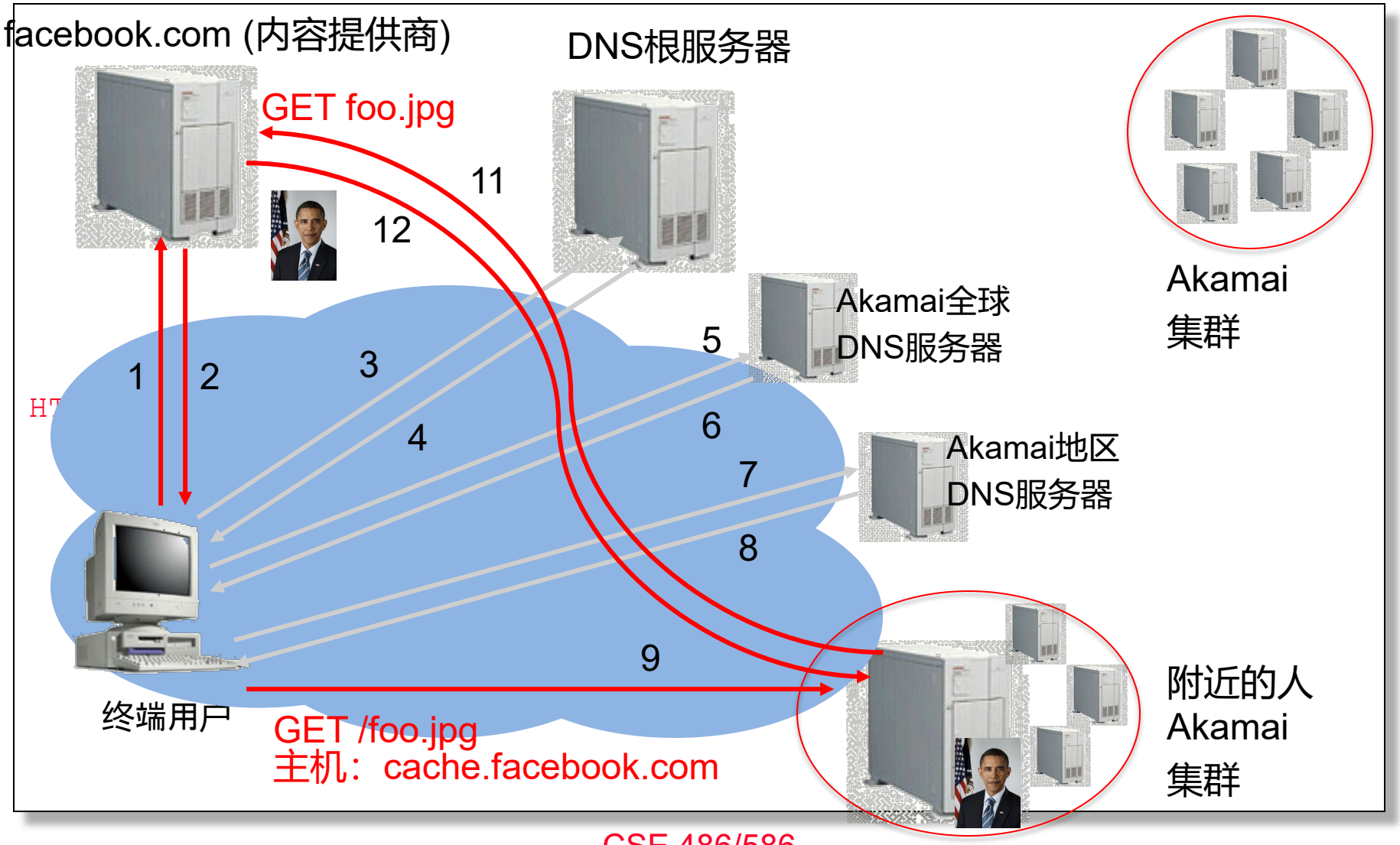
CDN如何工作


facebook.com (内容提供商)
DNS根服务器
Akamai
集群
Akamai全球
DNS服务器
5
1
2
3
4
6
7
8
Akamai地区
DNS服务器
服务器选择算法
DNS a73.g.akamai.net
地址
1.2.3.4
终端用户
附近的人
Akamai
集群

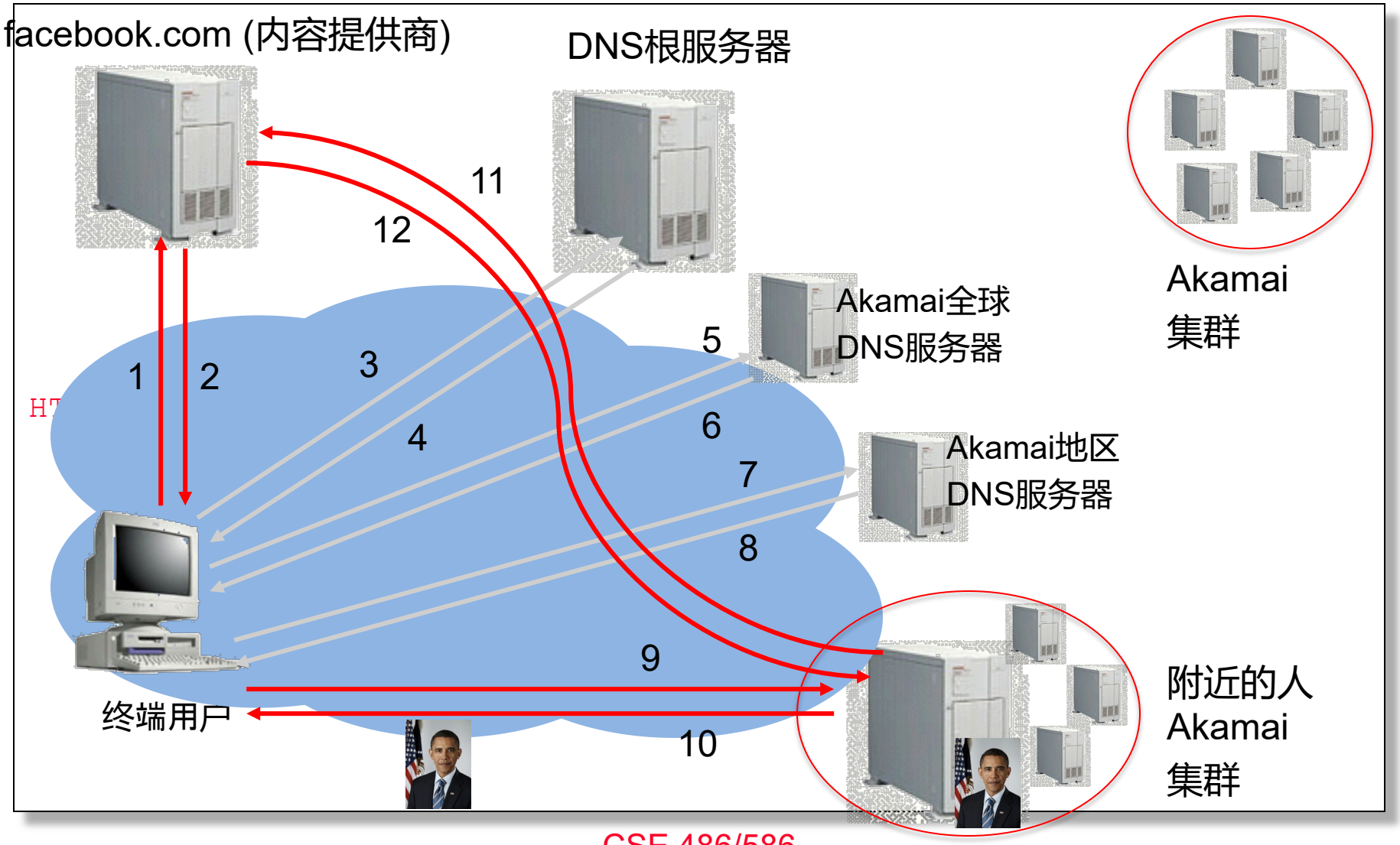
CDN如何工作


facebook.com (内容提供商)
DNS根服务器
Akamai
集群
Akamai全球
DNS服务器
Akamai地区
DNS服务器
1
2
3
4
5
6
7
8
9
终端用户
GET /foo.jpg
主机：cache.facebook.com
附近的人
Akamai
集群

CDN如何工作

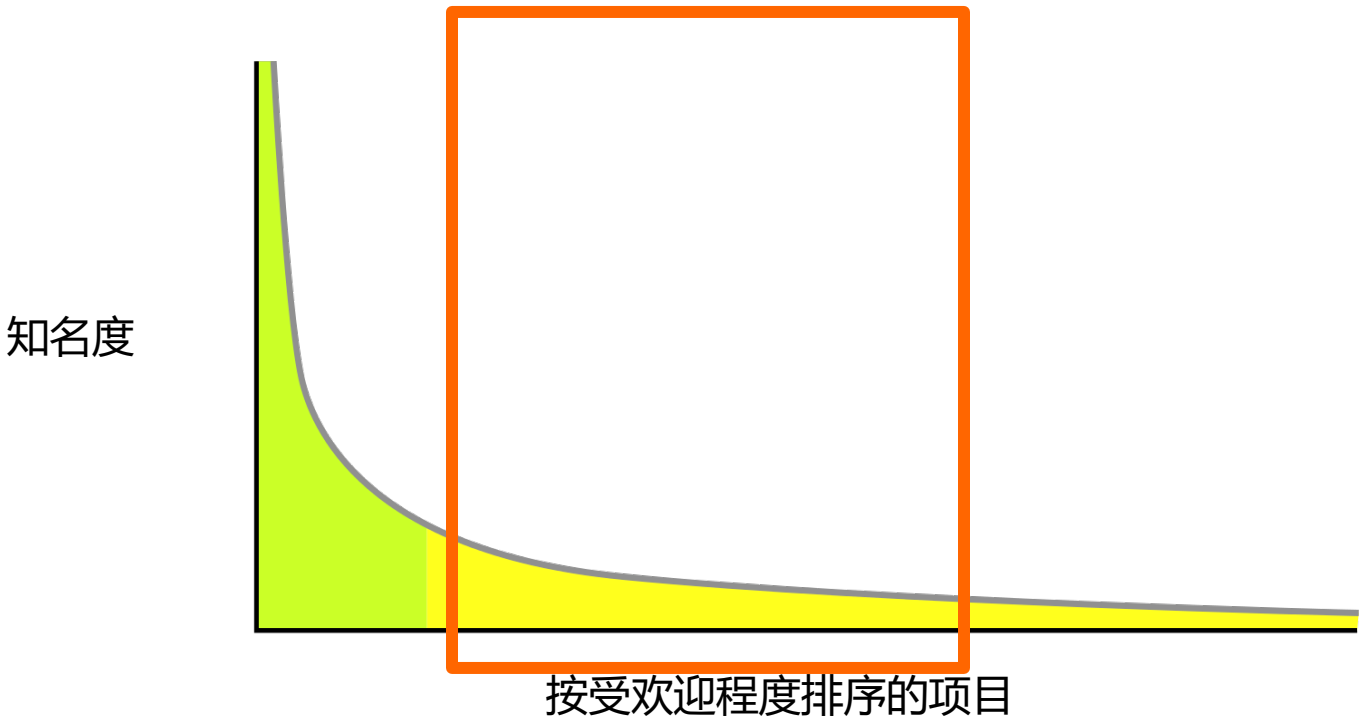

facebook.com (内容提供商)
DNS根服务器
GET foo.jpg
11
12
Akamai全球
DNS服务器
5
Akamai
集群
1
2
3
4
6
7
8
Akamai地区
DNS服务器
9
终端用户
GET /foo.jpg
主机：cache.facebook.com
附近的人
Akamai
集群

CDN如何工作


facebook.com (内容提供商)
DNS根服务器
Akamai
集群
Akamai全球
DNS服务器
Akamai地区
DNS服务器
终端用户
附近的人
Akamai
集群
1
2
3
4
5
6
7
8
9
10
11
12

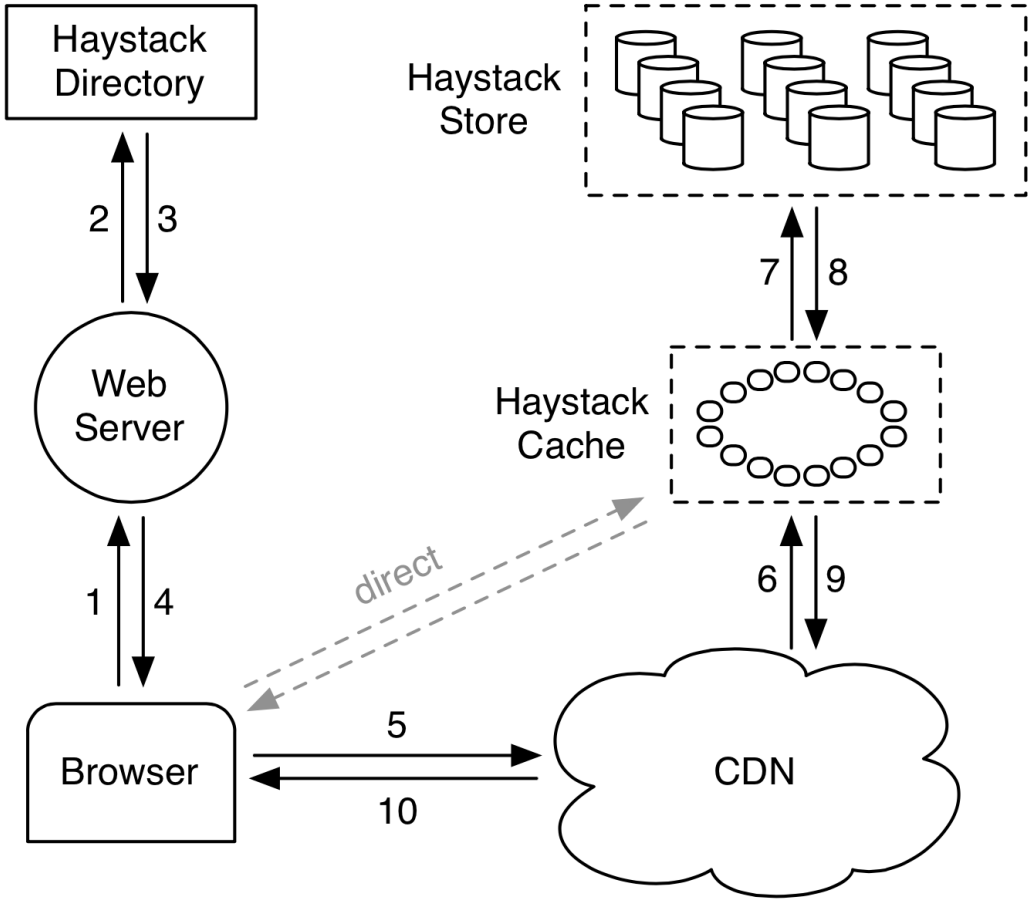
脸书照片分发

- "热 "与 "暖 "与 "冷 "的照片
  - 热门：受欢迎，浏览量大（约占浏览量的90%）。
  - 温暖。有点流行，但总的来说还是有很多意见
  - 寒冷。不受欢迎的，偶尔的观点



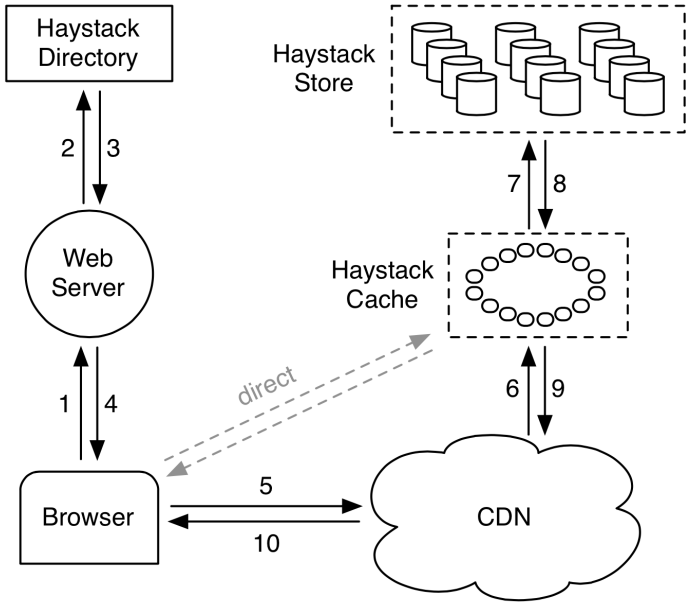
处理温暖的照片。干草堆

- 专为性能和可靠性设计
- "默认 "的照片存储



干草堆目录

- 有助于图像的URL构建
  - http://⟨CDN⟩/⟨Cache⟩/⟨Machine id⟩/⟨logical volume, Photo⟩
  - 阶段性查询
  - CDN将其部分剥离出来。
  - 缓存将其部分剥离出来。
  - 机器将其部分剥离出来



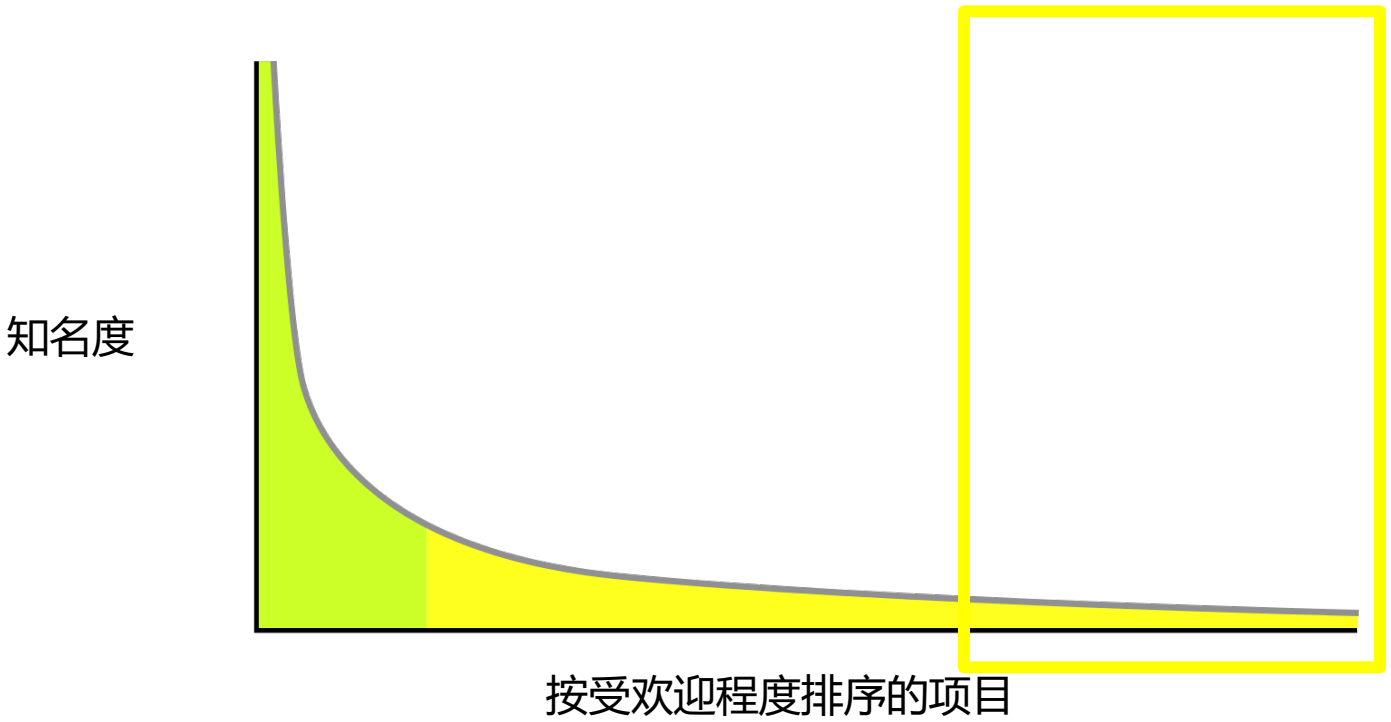
- 逻辑和物理卷
  - 一个逻辑卷被复制成多个物理卷
  - 物理卷的存储。
  - 每卷都包含多张照片。

干草堆储藏室和商店

- 干草堆缓存
  - 使用DHT的Facebook操作的第二级高速缓存
  - 带照片的身份证作为钥匙
  - 进一步消除了商店的流量
- 干草堆商店
  - 保持物理量
  - 一个卷是一个大文件(100GB)，有许多照片(针)。
  - 性能优化:检索图像时只需读取一次磁盘即可

脸书照片分发

- "热 "与 "暖 "与 "冷 "的照片
  - 热门：受欢迎，浏览量大（约占浏览量的90%）。
  - 温暖。有点流行，但总的来说还是有很多意见
  - 寒冷。不受欢迎的，偶尔的观点



CDN / Haystack / f4

- CDN为热门照片吸收了很多流量。
- Haystack的权衡：良好的吞吐量和可靠性，但存储空间的使用效率有点低（主要是由于复制）。
- f4的权衡：吞吐量较小，但存储效率更高。
  - ~ 在上传1个月后，照片/视频被移至f4。
  - f4使用纠错*编码方案*来有效地复制数据。

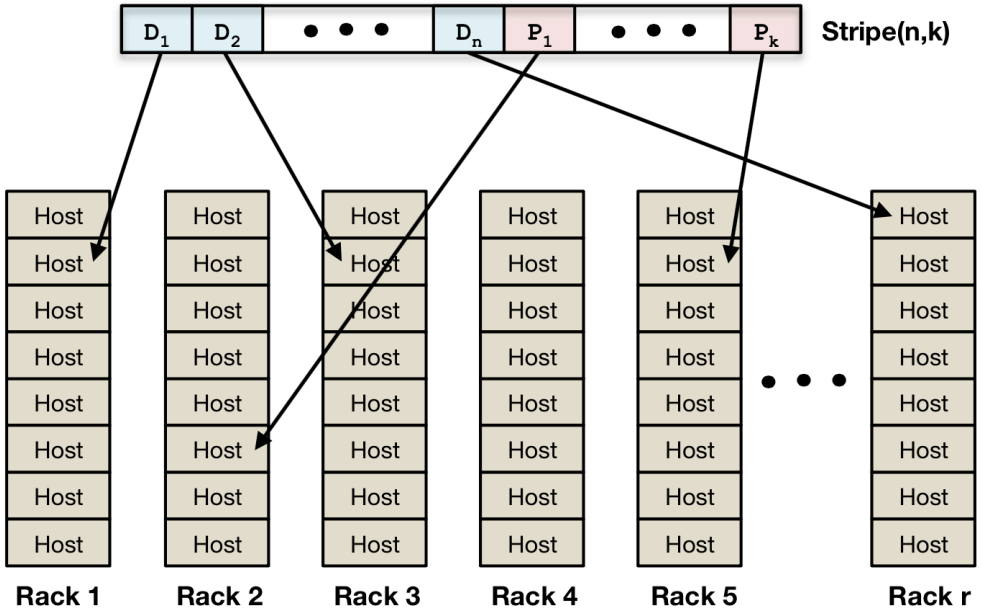
f4的复制

- (n, k) 里德-索洛蒙码
  - k个数据块，f==(n-k)个奇偶数块，n个总块
  - 一旦发生故障，任何k个区块都可以重构丢失的区块。
  - 可以容忍多达f个区块的故障
  - 需要通过编码器/解码器进行读/写，这影响了吞吐量

- 奇偶校验的例子。XOR
  - (Reed-Solomon使用了比这更复杂的东西)。
  - XOR位，例如，（0，1，1，0）P：0
  - 失败后的重建。(0, 1, 1, 0) P: 0

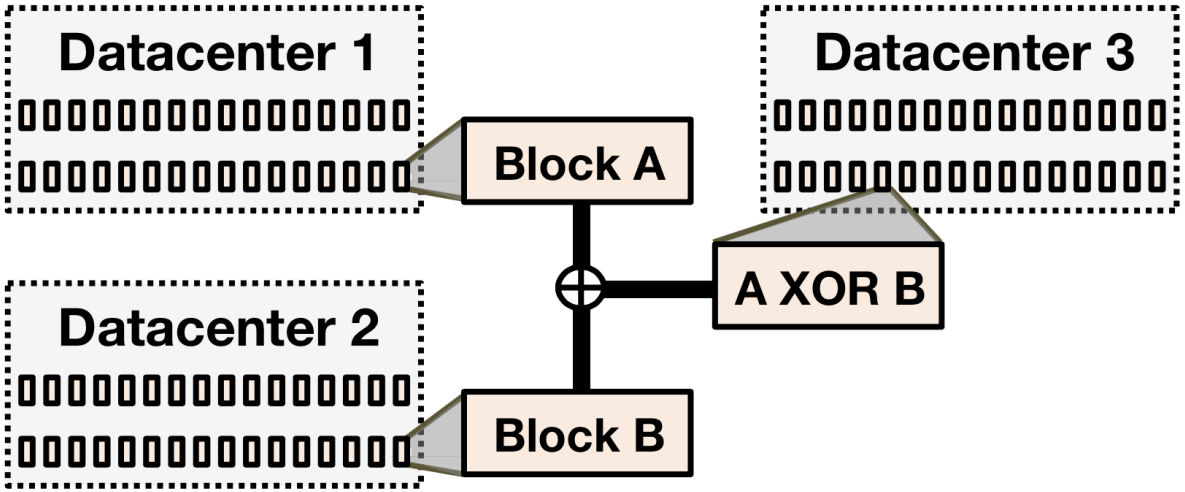
f4:单一数据中心

- 在单个数据中心内，(14，10)里德-所罗门码
  - 这最多可以容忍4个区块的故障
  - 每个区块1.4倍的存储用量
- 将区块分布在不同的机架上
  - 这可以容忍四个主机/机架的故障



f4:跨数据中心

- 额外的奇偶校验块
  - 可以容忍单个数据中心的故障



- 每个区块的总体平均空间使用率：2.1倍
  - 例如，A区和B区的平均值：(1.4*2+1.4)/2=2.1
- 具有2.1倍的空间使用率。
  - 可容忍4个主机/机架故障
  - 可容忍1个数据中心故障

摘要

- 设计一个系统需要对工作量的理解。
- 脸书照片的工作量
  - 热、暖、冷。
- 热门照片的CDN
  - 业绩
- 温馨照片的干草堆
  - 性能和可靠性
- f4用于拍摄冷门照片
  - 可靠性和存储效率

鸣谢

- 这些幻灯片包含由Indranil Gupta（UIUC）、Michael Freedman（Princeton）和Jennifer Rexford（Princeton）开发并拥有版权的材料。