7.6.3 代理缓存的层次结构

在实际中，实现*层次化*（hierarchy）的缓存是很有意义的，在这种结构中，在较小缓存中未命中的请求会被导向较大的*父缓存*（parent cache），由它来为剩下的那些“提炼过的”流量提供服务。图 7-9 显示了一个两级的缓存层次结构。[7]其基本思想是在靠近客户端的地方使用小型廉价缓存，而更高层次中，则逐步采用更大、功能更强的缓存来装载多用户共享的文档。[8]

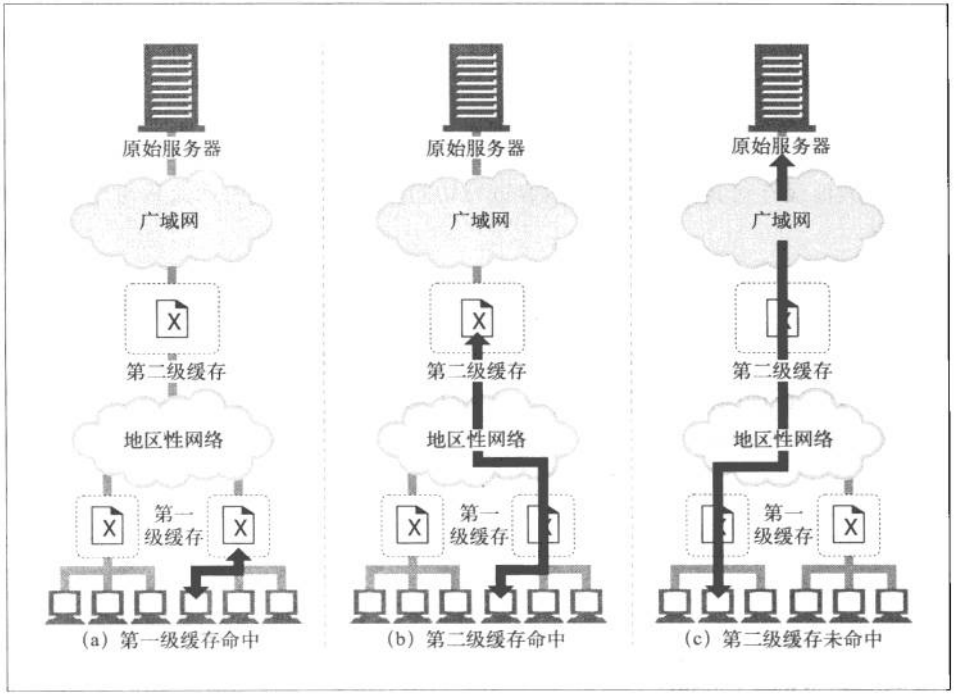

图 7-9 在两级的缓存层次结构中访问文档

我们希望大部分用户都能在附近的第一级缓存中命中（参见图 7-9a）。如果没有命中，较大的父缓存可能能够处理它们的请求（参见图 7-9b）。在缓存层次结构很深 169
的情况下，请求可能要穿过很长一溜缓存，但每个拦截代理都会添加一些性能损耗，当代理链路变得很长的时候，这种性能损耗会变得非常明显。[9]

注 7：如果客户端浏览器自带缓存，那么从技术上来讲，图 7-9 显示的就是一个三级的缓存层次结构。

注 8：父缓存可能要更大一些，以便装载在多用户间流行的文档，它们还要接收来自很多子缓存的聚合流量，这些子缓存的兴趣点可能很分散，所以还需要更高的性能。

注 9：在实际中，网络结构会尝试着将其深度限制在连续的两到三个代理以内。但是，新一代的高性能代理服务器会使代理链的长度变得不那么重要。