题目六 Cache 预取

假设某计算机由 1 个 CPU core,L1 和 L2 两级 cache，以及 DRAM 构成。本题将对不同的预取策略进行分析，包括 prefetch accuracy, coverage 和 bandwidth overhead 等。为了获得预取器稳定状态时的性能指标，每次实验前 6 个请求作为 warm-up，不计入统计范围。如果对一个 cache block 已经有一个未完成的内存访问请求，那么对同一个 cache block 的新请求将不会被生成，新请求将会与已经存在的请求合并成一条记录存在历史信息中。

已知某个程序的内存访问模式如下：（注意是 cache block 地址）：

A A+1 A+2 A+7 A+8 A+9 A+14 A+15 A+16 A+21 A+22 A+23 A+28 A+29 A+30...

（a）假如设计了一个 stride prefetcher，观察最近的三个 cache block 请求。如果最近三个请求之间存在一个恒定的 stride，它将使用该 stride 预取下一个 cache block。假设上面的程序运行了很长时间。那么该 stride prefetcher 的 accuracy 和 coverage 分别是多少？（Accuracy 的计算方法：有用的 prefetched blocks/所有的 prefetched blocks；Coverage 的计算方法：因为预取命中的内存访问/所有的内存访问）

解答：Accuracy 和 coverage 都是 0%。在每组三个请求之后，由于检测到步长，会触发预取操作，但预取的块始终是无用的；这个预取操作没有覆盖到任何需求请求。

（b）假如设计了一个 next-N-block prefetcher，每当访问一个 cach block，预取接下来的 N 个连续的 cache blocks。对于上面的内存访问模式，已知这个 prefetcher 的 coverage 和 accuracy 分别是 66.67%和 50%。那么 N 是多少？这个 prefetcher 的 bandwidth overhead 是多少？

Prefetcher 的 bandwidth overhead 定义为：
有 prefetcher 时从内存读取的总的 cache block 数/没有 prefetcher 时从内存读取的总的 cache block 数

解答：N=2, bandwidth overhead 是 5/3

(c)如果想提升 next-N-block prefetcher 的 coverage,但是必须保证 bandwidth overhead 小于 2，请问这可能么？

解答：为了获得更好的 coverage，预取器必须预取到前一组中下一组 3 个步长请求，因为第一组已经预取了完整的 3 个请求。例如，在访问 A+14 时，A+15 和 A+16 已经被预取。为了提高覆盖范围，应该预取 A+21（即下一组 3 个步长请求中的第一个）。然而，这将需要预取 A+16 和 A+21 之间的四个缓存块(A+17、A+18、

A+19、A+20）。这将使带宽开销超过 2 倍。

(d)对于 next-N-block prefetcher，如果想要 coverage 达到 100%，最小的 N 是多少？此时的 bandwidth overhead 是多少？

解答：N=5, 此时的 bandwidth overhead 是 7/3