In-class questions for Presentation:

*Summary Cache: A Scalable Wide-Area Web Cache Sharing Protocol*
• What is the caching in the Web proxy? What are the possible benefits of Web cache sharing? And what are the technical difficulties that hindered the wide deployment of web cache sharing?

• The right figure illustrates the Bloom Filter. In the summary cache, several bits of the vector will be set when a document is added to the cache. When a document is removed from the cache, how would the filter track the change?

![Bloom Filter Diagram]

Figure 2: A Bloom Filter with 4 hash functions.

The idea (illustrated in Figure 2) is to allocate a vector $v$ of $m$ bits, initially all set to 0, and then choose $k$ independent hash functions, $h_1, h_2, \ldots, h_k$, each with range $\{1, \ldots, m\}$. For each element $a \in A$, the bits at positions $h_1(a), h_2(a), \ldots, h_k(a)$ in $v$ are set to 1. (A particular bit might be set to 1 multiple times.) Given a query for $b$ we check the bits at positions $h_1(b), h_2(b), \ldots, h_k(b)$. If any of them is 0, then certainly $b$ is not in the set $A$. Otherwise we conjecture that $b$ is in the set although there is a certain probability that we are wrong. This is called a “false positive.” The parameters $k$ and $m$ should be chosen such that the probability of a false positive (and hence a false hit) is acceptable.
• Regarding the summary cache, what are false missies and false hits? Why the errors only affect the total cache hit ratio or the inter-proxy traffic, but do not affect the correctness of the caching scheme? Why reducing false misses is more important than reducing false hits in the design?

• The summary cache delays the update of summaries until the percentage of cached documents that are “new” (that is, not reflected in the summaries) reaches a threshold. How would the increase of the threshold affect the number of false missies and the number of false hits? Why?

• Could you explain this statement?

“For the summary cache, the network overhead is determined by the frequency of summary updates and by the number of false hits and remote hits. The memory requirement is determined by the size of individual summaries and the number of cooperating proxies.”

• In the summary cache, after a proxy builds a Bloom Filter from the list of URLs of cache documents, it will send the bit array/vector to other proxies. Do you think that it also needs to send the specification of the hash functions to other proxies? (assuming different proxy may use different set of hash functions.)