Adaptive Web Caching

Sally Floyd
Lawrence Berkeley National Laboratory
http://www-nrg.ee.lbl.gov/floyd/

Joint work with Lixia Zhang (UCLA)
and Van Jacobson (LBNL)

Cache Workshop '97
June 10, 1997

Overview of talk:

- The vision - a foundation for a global data dissemination architecture.

- Two main components:
  - The underlying communication paths between neighboring web caches.
  - The flow of requests for data along these paths.

- Important but orthogonal issues: caching decisions by individual caches, data integrity as a property of the data, unique names for data.

- Is there a path for incremental deployment from reality to vision?

The vision:

- Popular pages `spread out' from the origin server to neighboring caches, following demand-driven diffusion.

- New caches are easily added to the caching infrastructure in a self-configuring manner.

- Data integrity is a property of the data itself, and does not rely on authentication of the caches themselves.

Self-organization of the data flow:

- Individual web caches are autonomous agents.
- The forwarding of requests for data among the caches is based on local information.
- The global behavior that results
  - avoids manual configuration,
  - is adaptive to changes in topology, load, etc.,
  - routes requests to the nearest cache likely to have the data.

Forwarding URLs: metrics for evaluating neighboring caches or cache groups:

- Distance in hops from the origin server.
- Distance in Cache Group hops from the origin server.
- In the general direction of the origin server (using information from the routing tables of the nearby router?).
- Past success rate/declared willingness in resolving requests of this type (e.g., replicated servers, the Japan cache in the U.S.).
- Past success rate/declared willingness in resolving requests in general (e.g., the large regional cache in the other direction from the transoceanic link).
- PICS (Platform for Internet Content Selection) as a tool in selecting the metric?

Overlapping multicast groups of web caches: