

EURIM BACKGROUND PAPER

NETWORK GOVERNANCE WORKING PARTY

November 1999

INTERNET TECHNOLOGY EXPLAINED: HOSTING, CACHING AND MIRRORING

The draft electronic commerce directive makes substantial use of the terms “hosting”, “caching” and “mirroring” and reference is also made to them in other draft European and UK legislation. These words are applied in a number of contexts, but in that of the Internet they relate to very specific processes, which seem not to be fully understood by some of those working on the legislation. This paper sets out to remove the mystery and confusion and to explain the relevance of hosting, caching and mirroring to discussions on Internet regulation.

1. How Data Is Transmitted Over The Public Internet

Technically, the Internet is those networks covered by tcp/ip protocols. These include those for the secure supply chain and dealing systems of the US aerospace and financial services industries. In practice, however, what most consumers, academics and small firms have access to is only any-to-any packet switching. This works as follows:

- All types of information - whether a text, a conversation, or a photograph - are transmitted in the form of bits (1s and 0s of computer code).
- As data enters the network, the stream of bits representing a message is divided into a series of short sections or "packets".
- A "label" is attached to each packet indicating its final destination, and its place in the message as a whole.
- As it travels through the network, each packet passes through a series of specialised computers known as "routers". Each router decides which is the most efficient path for the packet to take on the next stage of its journey. This decision is taken on the basis of the router's understanding of network connections and loading patterns at the moment of the packet's arrival at the router. The process is then repeated at each router the packet passes through.

- This decision that the router takes will change over time as the router's knowledge changes. As a result, the thousands of packets which comprise a particular message may each take different paths through the network. They are only re-assembled in the proper order at their final destination.
- The destination will acknowledge the receipt of each packet to the sending system. The sending system will resend any lost information.
- Packets are also acknowledged across the communications paths that link the routers.

The key point is that routers must make a copy of each packet in order to read and direct it. The life-span of these copies is limited to a fraction of a second. The routers also copy each packet as they send it on, so that the packet can be retransmitted if it is lost. This copy is retained until the packet's receipt is acknowledged, which could be up to a few seconds. These copies represent only an unintelligible fragment of a whole message. They are nevertheless essential for the functioning of the Internet.

2. Hosting

This describes the process whereby information is permanently stored on a computer called a Web server for distribution via the Internet.

The combination of information and computer is called a Site, and the process of delivering the information to users is known as serving.

Most hosting is in the form of Web Sites. Some Web Sites are owned and managed by the information providers and connected to the Internet by leased lines purchased from ISPs; while others are managed by the information providers on shared computers provided for the purpose by ISPs.

An ISP will allocate a Web Site with a Domain name, which is the address that is used to access the Web Site on the World Wide Web (WWW). The Domain name is an alias for a number which is what the computer uses to find the location of the Web Site, for example `www.1acc.com = 207.159.89.66`. Once the Web Site has a Domain name it can then be accessed by any user on the WWW. There are special Internet protocols that allow a Domain name to be converted to an address or vice versa.

Information contained in the Web Site is altered or updated without the knowledge of the hosting ISP.

3. Caching

One of the main ways in which the Internet benefits society is through the abolition of geographic constraints on the distribution of information. Material which is stored on a computer on the other side of the world may be accessed as easily as material located in the next street.

This is possible because the network is organised so that it is not always necessary to transmit data packets across the full distance separating the end-user from the place where information is originally placed on the network. If this were to be necessary, the network would quickly become overloaded. Access for end-users would become intolerably slow, and often completely impossible.

Network breakdown as a result of overloading is avoided through the technique of caching. Caching takes place both on specialised servers within the network and on end-users' computers.

Caching involves the automatic storage of material when first requested by a user. Each Web page consists of a set of objects, and it is these objects that are cached. A network cache does not necessarily contain copies of all objects needed to build a page. Thus, the next time an object is requested by a user, it is not necessary to transmit the data packets all the way from the original hosting computer, and very little demand is placed on capacity in the core network. The only long distance transaction that is required is to check that the cached object is still up to date, and to get the uncached objects

Space in a cache is finite, so it cannot store all material that has ever been requested. Most caches use a rule whereby, when freeing up space, the "least recently used" item is deleted, which results in popular, or often-requested, items remaining in the cache, while less popular ones are flushed out.

Caching is an essential part of "browsing" the Internet. As an end-user leaves one Internet page for another, the information relating to the first page is not deleted from the memory of the user's computer. Instead, it is stored temporarily on the computer's hard disk so as to permit quick access if the user decides to return to this page. In the event that the user does return to the page, data does not have to be retransmitted across the network.

The copies made as part of caching on network servers and within user computers are absolutely essential in order to guarantee efficient functioning of the Internet. In a context where user numbers and traffic volumes are increasing exponentially, transmission capacity constraints are already a serious problem. Such constraints will become insurmountable if legislation in any way limits possibilities for caching.

Caching of material can be managed by the use of meta tags, which are words hidden at the top of a Web Site page, or associated with individual objects, that the reader cannot normally see but search engines can. The authors of documents can put meta tags in a document's <HEAD>_section that describe its attributes and which can create a set of instructions for a search engine or Web server. Meta tags can be used to mark a document object as uncacheable - so that a Web server will not cache the document - or as only valid for a period of time - so it is not used when it is too old.

4. Mirroring

Mirroring is the process by which separate, but identical, host sites are set up on different servers. A mirror site is a hosting site that is set up with the authority of (and quite often at the cost of) the content provider, in order to provide better and more resilient access to their information.

For example, a very busy site in the USA might have a mirror in Europe. Because these sites are no more and no less than alternative hosts for the information, all the accepted rules for hosts apply. Mirror hosts normally contain the whole site, permanently.

An unauthorised, or pirate, copy of a site is still a host. It is incorrect to use the term "mirror cache" to describe a mirror site (even a pirate one). Such a thing has none of the attributes of a cache, because the copies it holds are neither transferred to the cache individually as a result of end users requesting the pages, nor are they routinely flushed out as more timely material is added.

(The working part is grateful to Scott Broadley of BT for help in the preparation of this paper)

Copyright EURIM 2000. All Rights Reserved. For written permission to reproduce any part of this publication please contact the Administrative Secretary, EURIM, 5 Kingfisher House, New Mill Road, Orpington, Kent BR5 3QG. This will normally be given provided EURIM is fully credited. Whilst EURIM has tried to ensure the accuracy of this publication, it cannot accept responsibility for any errors, omissions, mis-statements or mistakes.