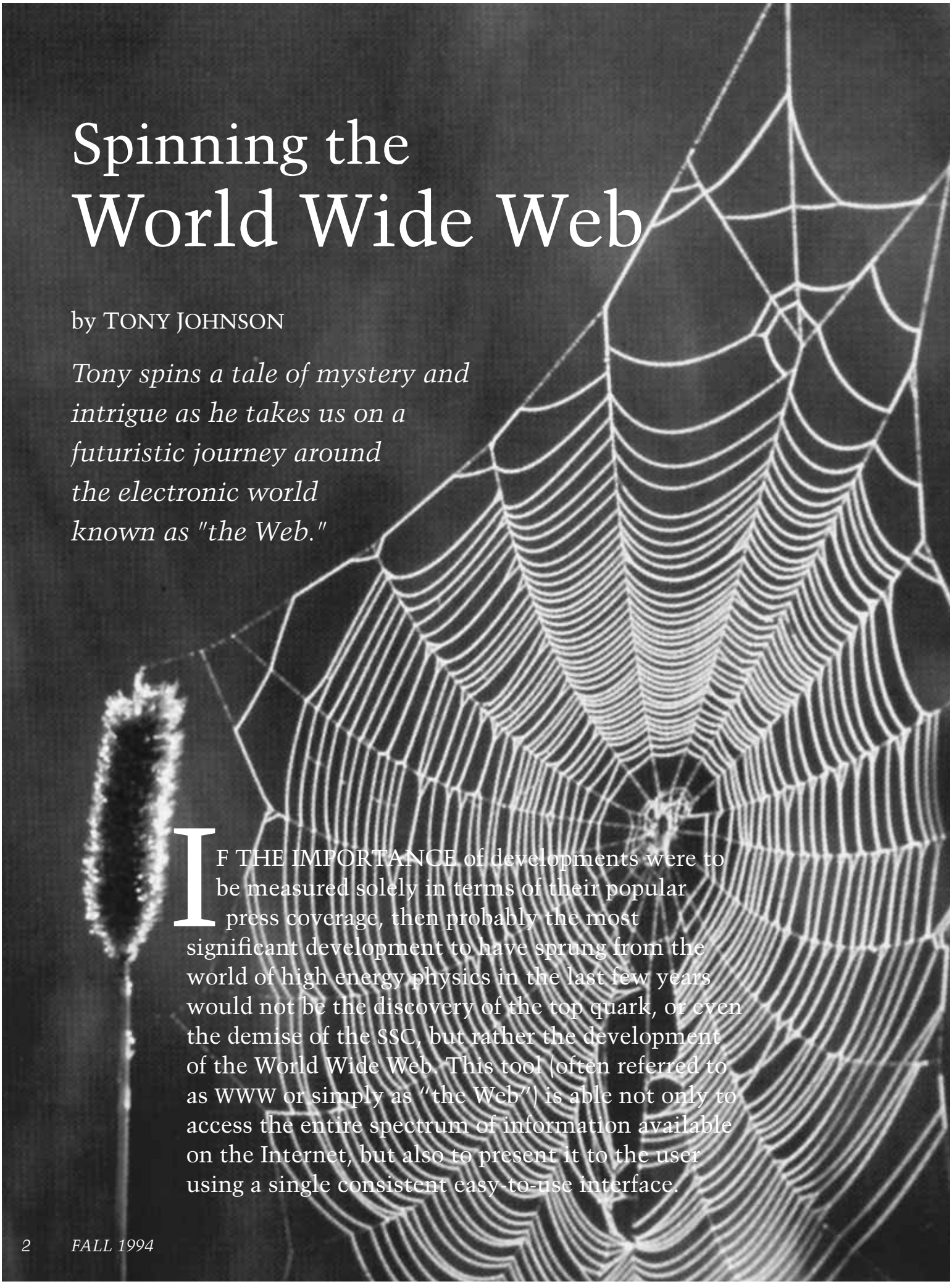
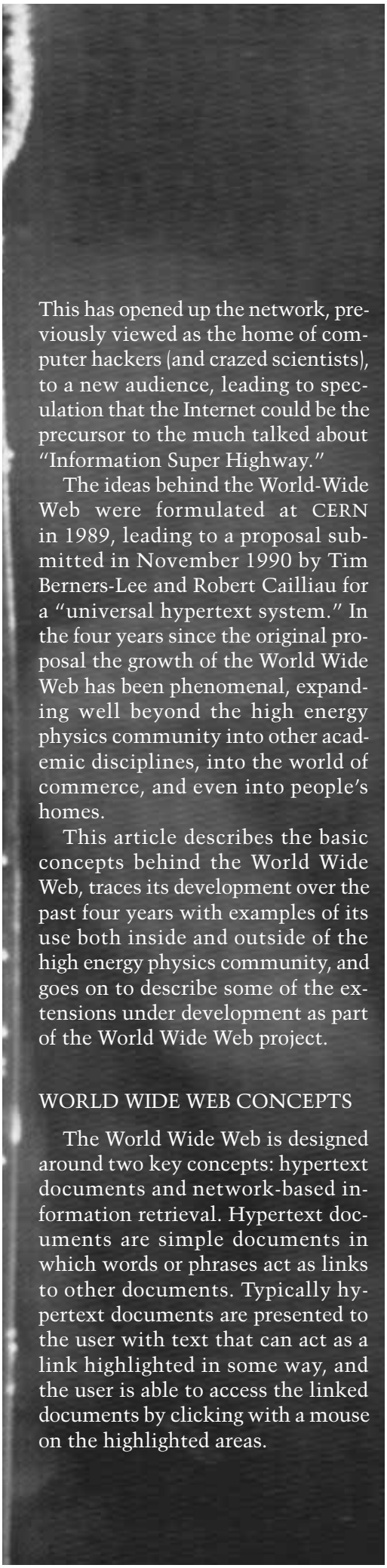
Spinning the World Wide Web

by TONY JOHNSON

Tony spins a tale of mystery and intrigue as he takes us on a futuristic journey around the electronic world known as "the Web."



IF THE IMPORTANCE of developments were to be measured solely in terms of their popular press coverage, then probably the most significant development to have sprung from the world of high energy physics in the last few years would not be the discovery of the top quark, or even the demise of the SSC, but rather the development of the World Wide Web. This tool (often referred to as WWW or simply as "the Web") is able not only to access the entire spectrum of information available on the Internet, but also to present it to the user using a single consistent easy-to-use interface.



This has opened up the network, previously viewed as the home of computer hackers (and crazed scientists), to a new audience, leading to speculation that the Internet could be the precursor to the much talked about "Information Super Highway."

The ideas behind the World-Wide Web were formulated at CERN in 1989, leading to a proposal submitted in November 1990 by Tim Berners-Lee and Robert Cailliau for a "universal hypertext system." In the four years since the original proposal the growth of the World Wide Web has been phenomenal, expanding well beyond the high energy physics community into other academic disciplines, into the world of commerce, and even into people's homes.

This article describes the basic concepts behind the World Wide Web, traces its development over the past four years with examples of its use both inside and outside of the high energy physics community, and goes on to describe some of the extensions under development as part of the World Wide Web project.

WORLD WIDE WEB CONCEPTS

The World Wide Web is designed around two key concepts: hypertext documents and network-based information retrieval. Hypertext documents are simple documents in which words or phrases act as links to other documents. Typically hypertext documents are presented to the user with text that can act as a link highlighted in some way, and the user is able to access the linked documents by clicking with a mouse on the highlighted areas.

The World Wide Web extends the well-established concept of hypertext by making it possible for the destination document to be located on a completely different computer from the source document, either one located anywhere on the network. This was made possible by exploiting the existing capabilities of the Internet, a world-wide network of interconnected computers developed over the preceding 20 years, to establish a rapid connection to any named computer on the network.

To achieve this, the World Wide Web uses a client-server architecture. A user who wants to access information runs a World Wide Web client (sometimes referred to as a browser) on his local computer. The client fetches documents from remote network nodes by connecting to a server on that node and requesting the document to be retrieved. A document typically can be requested and fetched in less than a second, even when it resides on the other side of the world from the requester. (Or at least it could be in the early days of the Web; one of the drawbacks of the enormous success of the Web is that sometimes transactions are not as fast now as they were in the earlier, less heavily trafficked days. One of the challenges of the Web's future is to overcome these scaling problems.)

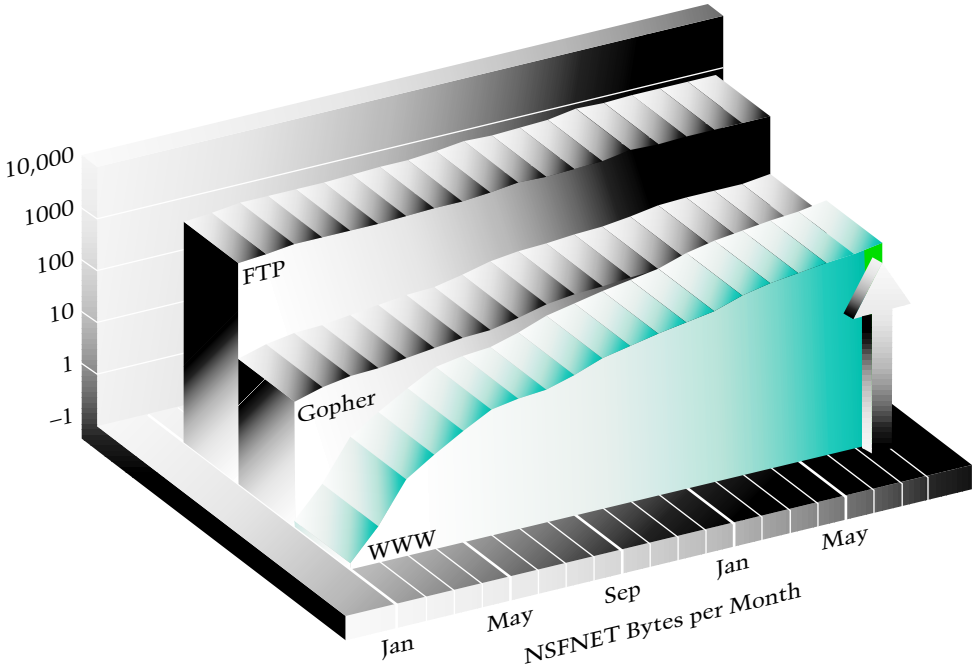
The client-server model offers advantages to both the information provider and the consumer. The information provider is able to keep control of the documents he maintains by keeping them on his own computer. Furthermore the documents can be maintained by the information provider in any form, so long as they can be transformed by the server software into the format the client software expects to receive. This model can naturally be extended to allow documents to be dynamically created in response to a request from users, for example by querying a database and translating the result of the query into a hypertext document.

From the information consumer's perspective, all the documents on the Web are presented in the form of hypertext. The consumer remains blissfully ignorant of how the documents are maintained by the information provider and, unless he really wants to know, from where the documents are being accessed.

GROWTH OF THE WEB

The initial implementation of the Web client at CERN was for the NeXT platform. This earliest browser was able to display documents using multiple fonts and styles and was even able to edit documents, but access was limited to users fortunate enough to have a NeXT box on their desks. This was followed by development of the CERN "linemode" browser, which could run on many platforms but which displayed its output only on character-based terminals. These early browsers were followed by the first browsers designed for X-Windows, Viola developed at the University of California, Berkeley, and Midas developed at the Stanford Linear Accelerator Center.

Initially the growth of the World Wide Web was relatively slow. By the end of 1992 there were about 50 hypertext transfer protocol (HTTP) servers. At about the same time,



The dramatic increase of World Wide Web usage over the past year and a half is illustrated. While the growth rate is phenomenal, more traditional uses of the network such as file transfer and e-mail still dominate.

Gopher, a somewhat similar information retrieval tool to WWW but based on menus and plain text documents rather than hypertext, was expanding rapidly with several hundred servers.

During 1993 the situation changed dramatically, driven in large part by the development of the Mosaic client by a talented and extremely enthusiastic group at the National Center for Supercomputer Applications (NCSA) at the University of Illinois in Champaign-Urbana. The Mosaic client for World Wide Web was originally developed for X-Windows under Unix, with subsequent versions released for both the Macintosh and PC platforms.

The Mosaic client software added a few new key features to the World Wide Web: the ability to display embedded images within documents, enabling authors to greatly enhance the aesthetics of their documents; the ability to incorporate links to simple multimedia items such as short movie and sound clips; and the ability to display forms. Forms greatly enhanced the original search mechanism built into WWW by allowing documents to contain fields that the user could fill in, or select from a list of choices, before clicking on a link to request further information. The introduction of forms to the WWW opened a new arena of applications in which the World Wide Web acts not only as a way of viewing static documents, but also as a way of interacting with the information in a simple but flexible manner, enabling the design of Web-based graphical interfaces to databases and similar applications.

During 1993 the usage of WWW began to grow exponentially. As new people discovered the Web they often became information providers themselves, and as more information became available new users were attracted to the Web. The graph on this page shows the growth in World Wide Web (or more accurately HTTP) traffic over the National Science Foundation backbone since early 1993, in comparison to Gopher and FTP traffic during the same period (FTP—file-transfer protocol—was one of the earliest protocols developed for the Internet, and is still the most widely used for transferring large files). While the growth in WWW traffic is enormous, it is worth noting that it is still not the dominant protocol; in fact, FTP, e-mail and NNTP (Network News transfer protocol) traffic are all substantially larger.

Owing to the distributed management of the Internet and the World Wide Web, it is very difficult to obtain hard numbers about the size of the Web or the number of users. (The number of users on the Internet, often estimated to be in the tens of millions, is itself a contentious issue, with some estimates claiming this number to be an overestimate by perhaps as much as an order of magnitude.) One illustration of the size of the Web came in early 1994 when a server was set up to provide information and up-to-the-minute results from the Winter Olympics being held in Lillehammer, Norway. The implementation of the server wasn't started until the day before the Olympics were scheduled to start, but two weeks later the server (together with a hastily arranged mirror server in the United States) had been accessed 1.3 million times, by users on somewhere between 20,000 and 30,000 different computers in 42 countries.

NCSA now estimates that more than a million copies of the Mosaic software have been taken from their distribution site, and approximate counts of the number of HTTP servers indicates there are more than 3000 servers currently operating (Stanford University alone has over 40 HTTP servers, not including one for the Stanford Shopping Center!).

HISTORY OF THE WORLD WIDE WEB

As the size of the Web has increased, so has the interest in the WWW from outside the academic community. One of the first companies to take an active interest in the World Wide Web was the publisher O'Reilly and Associates. For over a year they have provided an online service, the Global Network Navigator, using the World Wide Web. This includes regularly published articles about developments in the Internet, the "Whole Internet Catalog," an index of information available on the Web, a travel section, business section, and even daily online comics and advertising, all illustrated with professionally designed icons.

The Global Network Navigator is now only one of many examples of commercial publishers making information available on the Web, including a number of print magazines and newspapers which are available partially or in their entirety on the Web.

Another interesting example of commercial use of the World Wide Web is the CommerceNet organization. This organization, based in northern California and funded by a consortium of large high technology companies with matching funds of $6 million from the U.S. government's Technology Reinvestment Project, aims to actively encourage the development of commerce on the Internet using WWW as one of its primary enabling technologies. CommerceNet aims to encourage companies to do business on the Internet by making catalogs available and accepting electronic orders, and also by encouraging electronic collaboration between companies.

One specific way that CommerceNet is enhancing WWW is by the proposed introduction of a "secure-HTTP," which would enable encrypted transactions between clients and servers. This would ensure privacy, but perhaps more interestingly would also enable the use of digital signatures, effectively ensuring that when you fill in an order form on the Internet and submit it, it really goes to the company you believe you are ordering from (and only them), and that they know when they receive the order that it really came from you (and can prove it at a later date if necessary). This mechanism also begins to address a problem of great interest to commercial publishers—that of billing for information accessed through the Web. CommerceNet has ambitious plans to incorporate thousands of member companies in the first year or two, primarily in Northern California, but eventually to expand towards the much broader horizons of the Internet.

USES OF WORLD WIDE WEB IN HIGH ENERGY PHYSICS

While the Web has spread far from its original HEP roots, it remains an extremely useful tool for disseminating information within the widely distributed international high energy physics community. One example of the use of World Wide Web within HEP is the access provided to the SPIRES databases at SLAC, a set of databases covering a wide range of topics of relevance to HEP such as experiments, institutes, publications, and particle data.

March 1989
First proposal written at CERN by Tim Berners-Lee.

October 1990
Tim Berners-Lee and Robert Cailliau submit revised proposal at CERN.

November 1990
First prototype developed at CERN for the NeXT.

March 1991
Prototype linemode browser available at CERN.

January 1991
First HTTP servers outside of CERN set up including servers at SLAC and NIKHEF.

July 1992
Viola browser for X-windows developed by P. Wei at Berkeley.

November 1992
Midas browser (developed at SLAC) available for X-windows.

January 1993
Around 50 known HTTP servers.

August 1993
O'Reilly hosts first WWW Wizards Workshop in Cambridge, Mass. Approximately 40 attend.

February 1993
NCSA releases first alpha version of "Mosaic for X."

September 1993
NCSA releases working versions of Mosaic browser for X-windows, PC/Windows and Macintosh platforms.

October 1993
Over 500 known HTTP servers.

December 1993
John Markov writes a page and a half on WWW and Mosaic in the New York *Times* business section. *Guardian* (UK) publishes a page on WWW.

May 1994
First International WWW Conference, CERN, Geneva, Switzerland. Approximately 400 attend.

June 1994
Over 1500 registered HTTP servers.

July 1994
MIT/CERN agreement to start WWW Organization.

October 1994
Second International WWW Conference, Chicago, Illinois, with over 1500 attendees.

The largest of the SPIRES databases is the HEP preprints database, containing over 300,000 entries. In 1990 the only way to access the SPIRES databases was by logging in to the IBM/VM system at SLAC where the database resides, or by using the QSPIRES interface which could work only from remote BITNET nodes. In either case to access information you had to have at least a rudimentary knowledge of the somewhat esoteric SPIRES query language.

Since 1990, the introduction of the World Wide Web, coupled with the widespread adoption of Bulletin Boards as the primary means of distributing computer-readable versions of HEP preprints, has revolutionized the ease of access and usefulness of the information in the SPIRES databases.

The SPIRES WWW server was one of the very first WWW servers set up outside of CERN and one of the first to illustrate the power of interfacing WWW to an existing database, a task greatly simplified by WWW's distributed client-server design. Using this interface it is now possible to look up papers within the database without any knowledge of the SPIRES query language, using simple fill-out forms (for SPIRES aficionados it is possible to use the SPIRES query language through the Web too). Access to more advanced features of SPIRES, such as obtaining citation indexes, can also be performed by clicking on hypertext links. Since the access to the database is through WWW it can be viewed from anywhere on the Internet.

In addition, by linking the entries in the SPIRES databases to the computer-readable papers submitted to electronic Bulletin Boards at Los Alamos and elsewhere, it is possible to follow hypertext links from the database search results to access either the abstract of a particular paper, or the full text of the paper, which can then be viewed online or sent to a nearby printer.

The WWW interface to SPIRES has now been extended to cover other databases including experiments in HEP, conferences, software, institutions, and information from the Lawrence Berkeley Laboratory Particle Data Group. There are now over 9000 publications available with full text, and more than 40,000 accesses per week to the SPIRES databases through WWW.

Another area in which WWW is ideally suited to HEP is in providing communication within large collaborations whose members are now commonly spread around the world. Most HEP experiments and laboratories today maintain Web documents that describe both their mission and results, aimed at readers from outside the HEP field, as well as detailed information about the experiment designed to keep collaborators up-to-date with data-taking, analysis and software changes.

In addition large HEP collaborations provide an ideal environment for trying the more interactive features of WWW available now, as well as those to be introduced in the future. An example is the data monitoring system set up by the SLD collaboration at SLAC. The facility uses WWW forms to provide interactive access to databases containing up-to-date information on the performance of the detector and the event filtering and reconstruction software.

Information can be extracted from the databases and used to produce plots of relevant data as well as displays of reconstructed events. Using these tools collaborators at remote institutes can be directly involved in monitoring the performance of the experiment on a day-by-day basis.

FUTURE DEVELOPMENTS

The size of the Web has increased by several orders of magnitude over the last two years, producing a number of scaling problems. One of the most obvious is the problem of discovering what is available on the Web, or finding information on a particular topic of interest.

A number of solutions to this problem are being tried. These range from robots which roam the Web each day sniffing out new information and inserting it into large databases which can themselves be searched through the Web, to more traditional types of digital libraries, where librarians for different subject areas browse the Web, collate information, and produce indexes of their subject areas. A number of indexes are already available along these lines, or spanning the space in between these two extremes. While these are quite effective, none of them truly solves the problems of keeping up-to-date with a constantly changing Web of information and truly being able to separate the relevant from the irrelevant. This is an active area of research at many sites, together with other problems associated with scalability of the Web, such as preventing links from breaking when information moves, separating up-to-date information from obsolete information, and maintaining multiple versions of documents, perhaps in different languages.

One new area of research is the development of a new Virtual Reality Markup Language (VRML). The idea behind VRML is to emulate the success of hypertext markup language (HTML) by creating a very simple language able to represent simple virtual reality scenarios. For example, the language might be able to describe a conference room by specifying the location of tables, chairs, and doors within a room. As with HTML the idea would be to have a language which can be translated into a viewable object on almost any platform, from small PC's to high-end graphic workstations. While the amount of detail available would vary between the platforms, the essential elements of the room would be the same between the platforms. Users would be able to move between rooms, maybe by clicking on doors, would be able to see who else was in the room, and would be able to put documents from their local computer "on to the conference table" from where others could fetch the document and view it.

This type of model could be further enhanced by the ability to include active objects into HTML or VRML documents. Using this technique, already demonstrated in a number of prototypes, active objects such as spreadsheets or data plots can be embedded into documents. While older browsers would display these objects merely as static objects, newer browsers would allow the user to interact with the object, perhaps by rotating a three dimensional plot, or

World Wide Web Protocols

TECHNICALLY the World Wide Web hinges on three enabling protocols, the HyperText Markup Language (HTML) that specifies a simple markup language for describing hypertext pages, the Hypertext Transfer Protocol (HTTP) which is used by Web browsers to communicate with Web clients, and Uniform Resource Locators (URL's) which are used to specify the links between documents.

Hypertext Markup Language

The hypertext pages on the Web are all written using the hypertext markup language (HTML), a simple language consisting of a small number of tags to delineate logical constructs within the text. Unlike a procedural language such as Postscript (move 1 inch to the right, 2 inches down, and create a green WWW in 15 point bold Helvetica font), HTML deals with higher level constructs such as "headings," "lists," "images," and so on. This leaves individual browsers free to format text in the most appropriate way for their particular environment; for example, the same document can be viewed on a Mac, on a PC, or on a linemode terminal, and while the content of the document remains the same, the precise way it is displayed will vary between the different environments.

The earliest version of HTML (subsequently labeled HTML1), was deliberately kept very simple to make the task of browser developers easier. Subsequent versions of HTML will allow more advanced features. HTML2 (approximately what most browsers support today) includes the ability to embed images in documents, layout fill-in forms, and nest lists to arbitrary depths. HTML3 (currently being defined) will allow still more advanced features such as mathematical equations, tables, and figures with captions and flow-around text.

Hypertext Transfer Protocol

Although most Web browsers are able to communicate using a variety of protocols, such as FTP, Gopher and WAIS, the most common protocol in use on the Web is that designed specifically for the WWW project, the Hypertext Transfer Protocol. In order to give the fast response time needed for Hypertext applications, a very simple protocol which uses a single round trip between the client and the server is used.

In the first phase of a HTTP transfer the browser sends a request for a document to the server. Included in this request is the description of the document being requested, as well as a list of document types that the browser is capable of handling. The Multipurpose Internet Mail Extensions (MIME) standard is used to specify the document types that the browser can handle, typically a variety of video, audio, and image formats in addition to plain text and HTML. The browser is able to specify weights for each document type, in order to inform the server about the relative desirability of different document types.

In response to a query the server returns the document to the browser using one of the formats acceptable to the browser. If necessary the server can translate the document from the stored format into a format acceptable to the browser. For example the server might have an image stored in the highly compressed JPEG image format, and if a browser capable of displaying JPEG images requested the image it would be returned in this format; however, if a browser capable of displaying images only if they are in GIF format requested the same document the server would be able to translate the image and return the (larger) GIF image. This provides a way of introducing more sophisticated document formats in the future but still enabling an older or less advanced browser to access the same information.

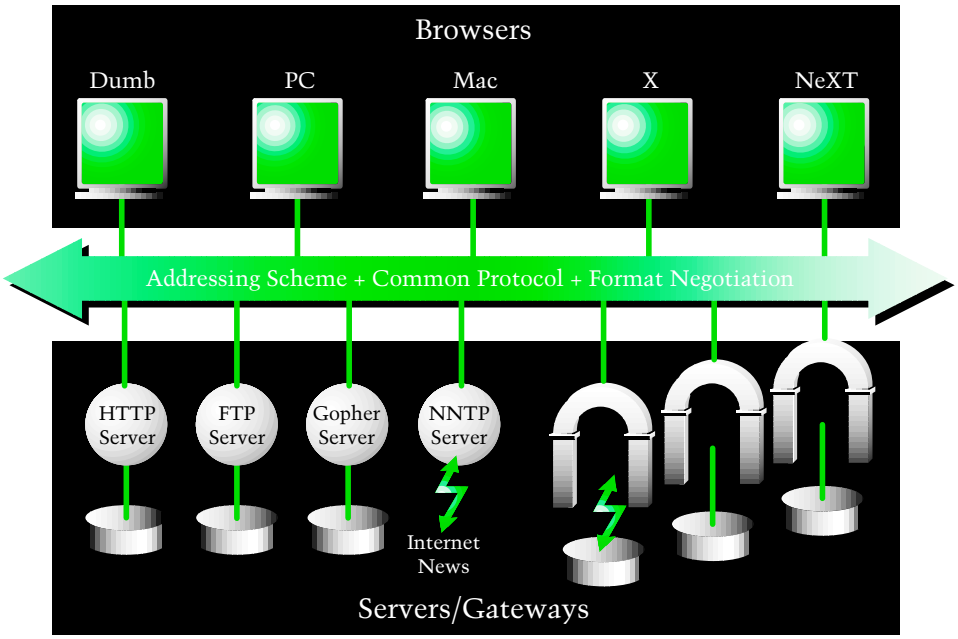
In addition to the basic "GET" transaction described above the HTTP is also able to support a number of other transaction types, such as "POST" for sending the data for fill-out forms back to the server and "PUT" which might be used in the future to allow authors to save modified versions of documents back to the server.

Uniform Resource Locators

The final keys to the World Wide Web are the URLs which allow the hypertext documents to point to other documents located anywhere on the Web. A URL consists of three major components:

<protocol>://<node>/<location>

The first component specifies the protocol to be used to access the document, for example, HTTP, FTP, or Gopher, etc. The second component specifies the node on the network from which the document is to be obtained, and the third component specifies the location of the document on the remote machine. The third component of the URL is passed without modification by the browser to the server, and the interpretation of this component is performed by the server, so while a document's location is often specified as a Unix-like file specification, there is no requirement that this is how it is actually interpreted by the server.



The ability of Web browsers to communicate and negotiate with remote servers allows users on a wide variety of platforms to access information from many different sources around the world.

expanding and rebinning a particular area of a data plot.

Currently the Web is viewed mainly as a tool for allowing access to a large amount of "published" information. The new features described here, together with the encryption features described earlier that will allow more sensitive data to be placed on the Web, will open up the Web to a whole new area, where it will be viewed more as a "collaborative tool" than purely an information retrieval system. Ideally it will be possible to take classes on the Web, to interact with the instructor and fellow pupils, to play chess on the Web, to browse catalogs and purchase goods, and to collaborate actively in real-time with colleagues around the world on such tasks as document preparation and data analysis.

CONCLUSION

Over the previous year the characteristics of the average Internet user have changed dramatically as many new people are introduced to the Net through services such as America Online, aimed primarily at home users. The current Web usage is likely to be insignificant in comparison to the potential for usage once the much vaulted "Information Super Highway" reaches into peoples' homes.

It is perhaps unlikely that the services eventually offered domestically on the Information Super Highway will be direct descendants of the World Wide Web, but what is clear is that WWW offers an excellent testing ground for the types of services that will eventually be commonplace. As such, the WWW may play a key role in influencing how such systems develop. At worst such a system may just become a glorified video delivery system and integrated home shopping network with a built-in method of tracking your purchases and sending you personalized junk e-mail. At its best such as system could provide truly interactive capabilities, allowing not only large corporations and publishers but also individuals and communities to publish information and interact through the network, while maintaining individual privacy. The outcome will have a major impact on the quality of life in the 21st century, influencing the way we work, play, shop, and even how we are governed.

ELECTRONIC SOURCES

THE SPIRES database and SLD information featured in this article can be accessed from the SLAC home page at:

http://www-slac.slac.stanford.edu/FIND/slac.html

The illustrations on the Web in this article show the Midas WWW browser developed at SLAC. Information on obtaining and using this browser is available from:

http://www-midas.slac.stanford.edu/midas_latest/introduction.html

Pointers to other pages mentioned in this article:

Global Network Navigator:
http://nearnet.gnn.com/gnn/GNNhome.html

CommerceNet:
http://www.commerce.net

Stanford Shopping Center:
http://netmedia.com/ims/ssc/ssc.html