Internet pioneer urges overhaul

Jared Shear / MSNBC

TCP/IP co-inventor Kahn sees a new Net where content is truly king

ANALYSIS
By Elliot Zaret
MSNBC

Nov. 21 — **Does the Internet need an overhaul? The sheer volume of information whizzing across the Net has skyrocketed with its mainstream adoption. But as everything from books, music and movies to corporate documents and medical data are increasingly accessed over the Web, there may just be just too much information to keep track of.**

THAT'S WHY Dr. Robert Kahn — one of the inventors of the Internet — wants to revamp the way information is handled over the Net. And his new system, which is increasingly being adopted by publishers and computer giants, may provide as much of an information revolution as the Internet itself.

To really wrap our minds around what that means and why it matters, we're going to have to back up the virtual truck here (Meep! Meep! Meep!) and briefly look at what this Internet thing really is.

While it may seem like the Internet suddenly appeared in the late 90s and exploded overnight into a world populated by the likes of Yahoo!, Amazon and Ebay, that's not what happened at all. First of all, the Web — and all the rich resources on it — isn't really the Internet. Neither is the series of wires, routers and switches that bring these things to your computer.

In actuality, the Internet is an "architecture," or roadmap, for connecting networks of computers to each other. It is the system of breaking down information into packets, the destination server, and shooting them from server to server until they reach the destination where they are re-assembled. This is what happens every time you send an email, read a Web page or download a song off Napster.

The concept of "packet switching" was pioneered in the 1960s and became a reality in the early 1970s when the U.S. Defense Advanced Research Projects Agency (known both as DARPA and ARPA) set up the ARPANET, which was able to connect a whopping 23 university and government computers by the end of 1971.

As the number of computers grew through the mid-1970s, it became clear that the computers in the network had to have a common language for transmitting the information through this packet system. Kahn, who was the director of DARPA's Information Processing Techniques Office, and Vinton Cerf, came up with this language called TCP/IP - which stands for Transmission Control Protocol (TCP) and Internet Protocol (IP).

The Internet as we know it was born January 1, 1983, when ARPANET converted to this TCP/IP system.

But everything about TCP/IP is about the "where" and not the "what." The protocols are deliberately blind to the contents of the packet of information. The whole point was to find an answer to the very tricky riddle of how you get the information from here to there.

"At the time we were building the original Internet, the question we were dealing with was how do you get these networks together," Kahn said.

In order to deal with the "where," every server is assigned a unique IP address - which is a series of four numbers each ranging from 0 to 255. For instance, MSNBC.com's server's actual IP address is 207.46.238.109. If you don't believe me, type the number into your web browser and try it.
Of course you never have to know that - thanks to a data base of all these numeric addresses and their associated host names, called a Domain Name Server (DNS). When you type in the name of a Web site, your server will check its DNS to convert the name into the number and then start firing off the packets of information to that address.

But an IP address only takes you to the home page of the server. The addressing system we all know and love - called "universal resource locators" or URLs - actually have two components: the host name and a path to the file.

For instance, the address to my last column "How the Net can Transform Voting" is http://www.msnbc.com/news/486207.asp. When you type that into your Web browser, it gets broken into two parts.

First, as explained earlier, the www.msnbc.com tells the server to go to the computer at the IP address 207.46.238.109. Everything after the first "slash" tells my browser where on that server to go - in this case, go to the directory called "news" and find the file called "486207.asp".

So why am I telling you all this? It turns out that while TCP/IP is a great way to get computers talking to each other, it's a lousy way to manage the information that's on these computers.

The most basic problem is finding the information. Even if you know the exact URL and have it bookmarked or written down - there's no guarantee it will work next year, let alone next week. If MSNBC decided to move that story to a different directory, your Web browser would get an error when you tried to go to the URL you have. With my stories, some may argue that's just as well; but other information is important to be able to come back to.

THE COPYRIGHT CONUNDRUM

Another problem is with copyrights and other protections of intellectual property. As we have learned from the recent Napster battles, this can be a real sticky wicket on the Net, where users can so easily and freely trade files regardless of any such protection. Because the current system is intentionally oblivious to what's in those Internet packets being transferred, there's no easy way to protect copyrighted data.

With all that in mind, Kahn decided it was necessary to develop a new framework for dealing with all the information on the Internet that would act as a layer above the existing infrastructure but deal with the "what" as much as the "where." So through the 90s, while most of the world was just discovering the Internet, Kahn was working on how to reinvent it.

His new system is called the "Handle System." Instead of identifying the place a file is going to or coming from, it assigns an identifier called a "handle" to the information itself, called "digital objects." A digital object is anything that can be stored on a computer: a web page, a music file, a video file, a book or chapter of a book, your dental x-rays - you name it.

Similar to the way a host name is resolved to an IP address, the handle will be resolved into information the computer needs to know about the object. Only since the information is now about the object, the location of the object is just one of the bits of information that is important. The handle record will also tell the computer things like what kind of file the object it is, how often it will be updated and how the object is allowed to be used - whether there are any copyright or privacy protections. The record can also have any industry-specific information about the object, like a book's International Standard Book Number (ISBN) code.

There are two other crucial things about these records: First, each handle can have multiple records associated with it - allowing multiple copies of the same information to be stored on different servers or for different systems. Second, the handle record is updated by the owner of the information - something in stark contrast to the host-name data, which is updated by central repository companies like Network Solutions. This will make things like changing the location of the file much more seamless, rather than waiting days for a new IP address to be updated in the DNS server (and anyone who has ever changed their server's host company or IP address knows how painful this can be.)

"You could now move it on a daily basis and a million users could come in and find it," said Kahn. "This identifier really becomes the hooks and handles of where you will go in the future."

So what does one of these handle identifiers look like? A handle consists of two sections: a prefix - or "naming authority" - and a "local name."

Under the handle system, my last column might have an identifier like: "10.12345/nov0700-zaret". In this fictional example (since MSNBC doesn't use the handle system), "10.12345" is MSNBC's naming authority, and "nov0700-zaret" is the name of the object. MSNBC would then keep a record in its handle registry that told the computer what server the object is on, what file it's stored in, as well as the copyright information and anything else it may want in that record. No matter where the file is moved, you would be able to use the handle to get to it, as long as the record is updated.

Further, like with the DNS system, the numeric address of the naming authority might be represented by an easier-to-remember word. So instead of all those numbers, MSNBC's naming authority might be "msnbc-news" - so the handle for my column would be "msnbc-news/nov0700-zaret."

Now all of this isn't just some sort of esoteric exercise - this is the real world, baby! The handle system is already starting to gain traction.

In 1994, the Association of American Publishers saw the whole e-book revolution on the horizon and realized they needed to have some way to identify their content (as opposed to the server it may reside on) and set up some basic means of protecting their copyrights. By 1997, they were working with Kahn and the CNRI to adapt the Handle System to their needs. The result is the "Digital Object Identifier" (DOI) directory.

"The DOI is not a silver bullet that will eliminate the urgent need for publishers to get their houses in order with respect to intellectual property management - although, with the establishment of a mini-industry of technology vendors building DOI-compatible electronic publishing solutions, they should find their burdens eased and their time to market shortened," wrote Bill Rosenblatt in the Dec. 1997 Journal of Electronic Publishing. "At a minimum, the DOI helps ensure that the publishers will dance to the same rhythm. By adopting the DOI, publishers are making a statement akin to the one the great publisher Benjamin Franklin made at the signing of the Declaration of Independence, 'We must all hang together, or assuredly we shall all hang separately.'"

The computer industry has joined in as well - Microsoft, Hewlett Packard and Xerox have all signed on to DOI for their electronic book ventures.

The Handle System is also being used by the Library of Congress to identify the materials in its National Digital Library Program. According to its web site, one of the reasons it went with the Handle System was the ability to give free access to copyrighted materials, while maintaining the protection.

(The) architecture can not write the law, but it provides a technical design that matches the legal structure that is expected to emerge," the Library of Congress says on its Web site. "The architecture respects the creators and owners of intellectual property. It allows the preservation of rights that can last for more than one hundred years, and recognizes that digital works may include material from many sources, with separate property rights."

Kahn's system is also being adopted by his old friends at DARPA, which is using it to build the Defense Virtual Library.

There are many other potential applications besides libraries. A digital object could be your current telephone number, and the handle could give you a permanent number
that would automatically be directed to your current phone. And because the handle system will allow numerous copies of the same file to be co-existing on many servers, it could allow for the kind of collaborative network that peer-to-peer proponents are pushing for.

"It's a very interesting reconceptualization of the Internet," said Kahn. "Treat it like it's one data base."

Of course no one - especially not Kahn - expects this to become mainstream overnight. But Kahn believes different industries will follow the book publishers' model and develop their own flavors of the handle system as need determines, and eventually it will become the norm.

Which is sort of what happened with the Internet - which took nearly 30 years to meander from the laboratory to the mainstream.

"When I first started talking about the Internet, some people had trouble understanding that too," said Kahn. "Very few understand the power and validity of this either."