The COOK Report on Internet

# Domain Name Service Under Stress

## Can IAHC Solution Work or Is NSI Unassailable? Could Legal Action Challenge Authority of IANA?

### We Survey NSF Policy, & Rise of IAHC and Ask Will Opposing Camps Compromise or Sue?

**Editor's Note**: We have written almost nothing about the debate of the last 18 months that has focused on NSF's authorization for Network Solutions (NSI) to begin charging for domain names. With the IHAC final report out and some speculation about whether the cooperative agreement with NSI will live out the full five years we decided that a summary of events might be useful. From our own files and a network of sources we compiled the following survey and analysis.

## Introduction

In 1992 NSF issued a competitive solicitation for "Network Information Services Manager(s)" inviting proposals to provide support for the non military portion of what was beginning to be called "the Internet" in three areas: Registration Services, Directory and Database Services and Information Services. Organizations were encouraged to submit proposals in any or all of the three service areas. It should be noted that prior to NSF's solicitation, registration services had been supported by DoD as a part of the continuing ARPANET and MILNET efforts and that limited information services for the non-military Internet community (then largely academic institutions) were provided by the National Network Service Center run by BBN and supported by the NSFNET program.

As a result of the solicitation process, outstanding proposals in each service area emerged (that of Network Solutions: for Registration Services, AT&T for Directory and Database Services, and General Atomics/CERFNET for Information Services). NSF, as a part of its review process asked these three organizations to develop a plan to insure that the NSFNET community (academic users) would not have to deal with multiple organizations but would perceive the three as having one "friendly" face. The three organizations developed a concept to do this which they called the "InterNIC" and submitted to NSF revised proposals which included this concept.

In 1993 Network Solutions, AT&T and General Atomics each became the recipient of a cooperative agreement for services within the respective areas delineated above and requiring that they provide their respective services collaboratively consistent with the "InterNIC" concept for a uniform interface which they had developed. During the early stages of InterNIC activity, the primary point of entry into the "InterNIC" was via a toll free phone number at General Atomics from which callers were either served locally (if their request was for general information) or referred to Network Solutions or AT&T dependent on their specific needs. We note that despite recent confusing assertions to the contrary, reference to the InterNIC concept to becoming the "NIC of NICs" referred ONLY to the General Atomics Information Services activity being a resource NIC for the information service centers being developed on academic campuses throughout the country (rather than providing end-user services directly to students and faculty which would inhibit such local developments).

The specific responsibility of NSI under their cooperative agreement was to provide registration services, including assignment of IP and autonomous system numbers, registration of domain names and provision of INADDR.ARPA services, for the non-military Internet. The authority to do these things were delegated to NSI by the Internet Assigned Numbers Authority (IANA) supported by DoD. The responsibilty for funding came from the NSF under the Cooperative Agreement.

Vol. VI, No. 1, April 1997
© Cook Network Consultants
ISSN: 1071-6327

As originally awarded, the cooperative agreement to NSI to provide these services was written on a cost-plus-fixed fee basis and was estimated to cost NSF approximately $5,000,000 over a 5 year period (or approximately $1,000,000/ year) with the specific amounts for each year being negotiated based on experienced and anticipated growth in demand for the services. Based on past history, continued moderate growth in the size of the activity was anticipated. Under such an arrangement, if NSI's expenses exceeded the budgeted amount because of increased demand, either the NSF

On the Inside:

would be required to make up the difference or NSI would not be required to provide services in excess of the costs funded. (As a point of reference, when the NSI award was made in January 1993, the total global population of the domain name database was around 4000 and new name registrations were running a few hundred a month, most in the .EDU domain).

The solicitation for the award came out in 1992 at a time when few people realized there was an Internet and even fewer realized that it would become commercial but no one had any serious inkling how fast it would grow. For example, by fall of 1994, new name registrations had increased to about 2,000 per month, the database population approximated 20,000 and the majority of names being registered were those of commercial companies. It was only at this point that the Web began to make a serious impact on growth rates.

In September of 1994, the IEEE held a workshop to discuss the issues related to domain name registration which recommended that NSF find a way to move to a user fee for registration services. (gopher://ds0.internic.net:70/00/nsf/cise/workshop.asc) In November, 1994, NSF conducted a merit review of the Network Information Services Managers project. That review (http://rs.internic.net/nsf/review/review toc.html) recommended the early termination of the General Atomics portion of the activity (due to disappointing results) and continuation of the AT&T activity. It was highly laudatory of the efforts of Network Solutions but recommended that NSF extricate itself from funding the registration for commercial organizations as rapidly as possible.

From January 1, 1993 through March 31 1994 (15 months) the NSI DNS cost to the NSF award budget was $1.1M. From April 1, 1994 through March 31, 1995 (12 months) the cost was $1.3M From April 1, 1995 through September 30, 1995 (six months) the cost was $1.6M. From October 1, 1995 through March 31 96 (6 months) cost was estimated at $4M. There had been an increase in registrations from September 1994 which ran 2,000 a month to 20,000 a month in September 1995. The growth in registration of .com domains was on an upward curve such that the expense paying NSI the required projected registration costs would soon hit $12 million dollars per year or 1/3 of the NSFNET budget of DNCRI (the Internet arm of the NSF with oversight responsibility for the CA).

Unfortunately things began to unravel fast. In November, 1995, the first NSF sponsored symposium on "Internet registration, governance, supportability and sustainability" addressed those issues only tangentially. The focus of the workshop became the NSI "monopoly position" and the possibility of NSI reaping "unconscionable profits" as a result.

This was the first unexpected fall out from the explosive growth of the net. The second was the beginning of trademark related lawsuits by companies which learned that the .com domain name they sought to obtain had already been registered by someone else. In late 1994 officials at the NSF found themselves named in such a suit for the first time. They also realized that a significant portion of the NSF DNCRI budget would soon be eaten up by the costs of supporting charges for commercial domains, clients whose operations had nothing to do with the research and education mission that congress had chartered the NSF to promote.

## Charging for Services

The original solicitation had warned that the recipient of the CA might be required to start charging for its services during the term of the CA. Seeing that DNCRI would be swallowed alive by the monetary and legal fallout from the DNS explosion, DNCRI requested in the first half of 1995 that NSI develop and submit for NSF approval a plan to begin to charge an annual fee for registration and maintenance services related to domain name registration. The agreed upon amount of the fee was set at $50 a year with new names being charged initially for two years maintenance or $100 and 30% ($15 of each $50) was to be set aside by NSI in escrow for the support of the "intellectual infrastructure" of the Internet. NSF's announced hope was an extremely laudatory one - namely that the Internet community would identify new or unite behind existing organizations (such as the IETF, ISOC or IAB) to receive the funds which could have been used for supporting the IETF secretariat, the IAB or an RFC secretariat or the when the current federal support for these activities was to be discontinued.

On September 14, 1995 NSI began to charge for domain names having entered into an agreement with NSF that freed it both from the financial burden of supporting .com and from being named a defendant in a lawsuit. When the fees were introduced, it was stated that they should be viewed as a short-term expedient to solve the funding problem. The NSF asked the Internet community to address longer term issues of "governance, supportability and sustainability" and indicated its interest in supporting appropriate parties to hold workshops to facilitate the discussion of these topics.

Unfortunately things began to unravel fast. In November, 1995, the first NSF sponsored symposium on "Internet registration, governance, supportability and sustainability" addressed those issues only tangentially. The focus of the workshop became the NSI "monopoly position" and the possibility of NSI reaping "unconscionable profits" as a result. The only agreement at that workshop seemed to be that (domain) "names registration and (IP) number assignment should be separated".

Neither that workshop nor it's successor in February of 1996 <http://www.aldea.com:80/cix/agenda.html> produced any consensus on either future models for sustainable and supportable Internet Governance or the use of the infrastructure funds. Instead, what began to emerge was an agreement that competition in domain name registration was desirable and some consensus that name registration and number assignment should be separated (to eliminate having a single organization in charge of both) and that the European and Asian-proved model (of service provider-based associations) for managing IP space appeared sound. The larger issues, however, continued to be largely unaddressed.

At the same time many network hotheads, in the opinion of their detractors, motivated to "make a killing" by providing name registration services of their own, began a concerted campaign against the "evil" NSF for having granted NSI a huge money making monopoly as though NSF had "plotted" from day one to bring this to pass. We find these charges completely unfounded.

## The Dissidents

By June of 1996 a group called the Alternic had coalesced around an individual named Eugene Kashpureff. (On mail lists the most vocal attackers were probably Bob Alistat and Jim Fleming). The Alternic group announced their own competitive domain name service and started collecting money for domain names that they issued. Rarely ever mentioned in their discussions was the premise that unless 90% or more of the network's root domain name servers were pointed to take their data from the Alternic domain name database, a web site, or an individual using a domain name assigned by them, would not be reachable from the vast majority of the rest of the network. Some on the network called these people "DNS pirates". Failing to see how they could claim more legitimacy than that of the groups they were railing against, we are inclined to agree. In the words of their critics: "Surveys show Alternic penetration is less than 0.5% so they are notvisible to 99.5% of the net. The bottom line is that these folks are attemptingto do a land grab of a valuable public resource."

We think it not at all surprising that the commercial providers operating the dozen or so root level DNS servers around the Internet continued to fill them with the NSI database that, by early 1997, contained upwards of 800,000 names.

## The Community's Solution

Meanwhile in the late spring of 1996 the IANA (Jon Postel) published an Internet draft serving notice that it intended to sanction the formation of several competitive groups to issue new competitive top level domain names. Left unaddressed by Postel was the question of how these names would be integrated into the network root nameservers. Perhaps he thought that they would replace NSI in a nice orderly transition when the NSI cooperative agreement ended on April 1 of 1998?

Postel, in his draft, was acting on behalf of the Internet Society (ISOC). By the early fall of last year ISOC announced the formation of IAHC (Internet International Ad Hoc Committee) This committee headed by ISOC President and CEO Don Heath was liberally sprinkled with lawyers and, in addition a few representatives from long time Internet users, had representatives from the ITU (International Telecommunication Union), INTA (International Trademark Association) and WIPO (World Intellectual Property Organization). Domain names as trademarks were well looked after by the IAHC group. But a noticeable absence from the IAHC group was any representation of any major player from the commercial Internet (for example Netscape or any of the largest service providers like MCI, Sprint, UUNET, BBN etc) - a remarkable omission when one considers the companies and monies powering the Internet's growth.

In January 1997 IAHC came out with a plan that called for the creation of seven new international top level domains and for up to 28 groups to apply to IAHC/ISOC to for the right to be competitive registries for the seven new top level domains using a shared data base under the control of a Council of Registrars known as CORE. The idea is to let the registrars compete with each other for the cost effective issuance of names. In theory at least, all the registrars will be able to register all the seven new top level domain names domains into a shared data base run by the Council of Registrars. The names would be time stamped and first one to hit the shared data bases becomes authoritative.

Unfortunately several critical unknowns are lurking in the new IAHC process. First the time for the construction of the data base and the infrastructure needed to link it together is very short. Secondly it is unlikely that top new top level domains will solve anyone's problem. After all who thinks that any of the 2000 to 5000 of the world’s largest commercial organizations which currently have .com registrations are going to sit back and let other companies of any size register their trademarked .com name as a .firm name? Would IBM let any other company anywhere in the world register as IBM.firm? Not likely. The new top level domains, it seems to us, do nothing to satisfy the issue of one uniquely identifying name for a company large or small anywhere in the world.

It is possible that international top level domain names will soon be good for only one end and that is to place their holders at risk from legal action against them from somewhere in the world. (For more on this see the discussion of arbitration below.) Not surprisingly, IAHC has a different point of view. A source pointed out to us: "You're picking the extreme case of a globally known trademark - now try the exercise with a trademark like 'united' or 'genesis'. Many companies are essentially 'locked out' of .com because someone else has their name. We are trying to faciliate access for them to the DNS. Access is a big issue - note how Gateway Computers has just sued some poor soul who has held gateway.com for many years."

At a minimum, if one moved to geography based versions of .com domains, -- .com.us for example but something that would not solve the problem of large multi nationals -- one would at least limit the realm within which the lawyers could ply their trade. Unfortunately with the rancor and acrimony generated by the disputes of the past year, no one has been looking at these kinds of problems. Moreover, domestically, the alternative of using the geography based US domain has been rendered less and less feasible by the parcelling of that out by the IANA over the last 2 years to a number of groups authorized to register state and city areas and charge for doing so. People are beginning to complain about being told the name and number of the registry for their area and never being able to get a human to answer the registry's phone.

## The Competing and Hostile Groups

What we have now is a dangerous situation of three competing groups who want to register domain names. First there is NSI which has current de facto and legal control of the current top level names (including all powerful .com) having been given that right by the IANA following their selection by the NSF in an open fair and competitive bid process nearly five years ago. NSI has the infrastructure and the cash flow to keep on trucking. Unfortunately it has the hatred of many network users who simply do not understand with sufficient clarity how we arrived at the current state of affairs. Second there is IAHC which represents the solution of the Internet community. Some consider - perhaps unfairly IAHC to represent the interests of the Internet "old boys" from the research and education community. Third there is the Alternic crew which although loud and vociferous has a very limited subscriber base and limited credibility with those who understand the chain of events outlined in this article.

IAHC has let NSI know that it wants to be able to start registering .com domains. NSI, in an interview with its new PR director, has somewhat arrogantly replied that it will think about it, but only if IAHC can show in advance that it can meet the appropriate standards. IAHC believes that it represents the will of the Internet community. Many segments of community generally dislike NSI and indeed it is fair to say that some segments of it despise NSI. The question becomes one of whether the two sides will be able to work out their differences amicably or whether they may face each other in court.

Since the authority for NSI's charter is embedded more in custom than in law, the possibility of some one attacking that authority by legal means is great. Also the Alternic could become embroiled with either of the first two, or customers with any of the three.

In evaluating at IAHC's chances of success one must ask two things. Where are we to assume that it will get the considerable sum of money required to build its infrastructure and make it work in a very short period of time? Second who will register what domains and how quickly will the data from the new Registrar's IAHC sponsored seven top level domains start being added to the data bases of the root DNS servers? The IAHC side has indicated that it would like NSI to join CORE and add .com to the domains registered by the CORE Registrars. As soon as the IAHC CORE system has an operational shared database, we'd like to see NSI join and do just this. However when one looks at the stark reality that NSI would have to give up control over its pre-eminent cash cow in order to do so, we think it very unlikely that this will occur in the short term. After all the shared database system has to be built first and it is not clear where the sponsors are going to find the money necessary to do so. Furthermore if NSI goes ahead with its IPO, it is unlikely that its stockholders would ever permit such and action by it. Having NSI and CORE registering .com independently of each other of course could not work because there would be no single authority files. However, in the longer term, if CORE does succeed in building a working system that gains acceptence, IANA and community pressure could force some changes.

## Dispute Resolution

The IAHC Final report seeks to avoid using the legal system wherever possible. This is a goal that we certainly endorse. However given some of the players and issues at stake we wonder how well the new system will succeed.

Section 7.1.2 of the IAHC final report calls for "Administrative Domain Name Challenge Panels will be established to administer the policy that second level domain names which are identical or closely similar to names which are, for the purposes of the Internet, internationally known, and for which demonstrable intellectual property rights exist, should only be held by, or with the authorization, of, the owners of such demonstrable intellectual property rights." It goes on to say: "No decision of an administrative domain name challenge panel would affect the power of the appropriate national or regional soveriegn court to hear cases interpreting and enforcing intellectual property rights that fall within its jurisdiction. Likewise, this procedure would not prevent any party from bringing a case before such national or regional sovereign court, nor from initiating arbitration or mediation procedures that are otherwise available."

In evaluating at IAHC's chances of success one must ask two things. Where are we to assume that it will get the considerable sum of money required to build its infrastructure and make it work in a very short period of time? Second who will register what domains and how quickly will the data from the new Registrar's IAHC sponsored seven top level domains start being added to the data bases of the root DNS servers? The IAHC side has indicated that it would like NSI to join CORE and add .com to the domains registered by the CORE Registrars. As soon as the IAHC CORE system has an operational shared database, we'd like to see NSI join and do just this. However when one looks at the stark reality that NSI would have to give up control over its preeminent cash cow in order to do so, we think it very unlikely that this will occur in the short term. After all the shared database system has to be built first and it is not clear where the sponsors are going to find the money necessary to do so.

Section 7.1.1 calls for "On-Line Mediation and Expedited Arbitration. "Each application for a domain name under a shared gTLD administered by CORE (whether or not a voluntary 60-day waiting period isrequested) will contain a clause whereby the applicant agrees: To participate in on-line mediation under the mediation rules of the Arbitration and Mediation Center of the World Intellectual Property Organization (Geneva), if such mediation is initiated by a right holder who wishes to challenge the domain name applicant's right to hold and use the second level domain name. To participate in binding expedited arbitration under the corresponding rules of the WIPO Arbitration and Mediation Center, if such arbitration is initiated by a right holder who wishes to challenge the domain name applicant's right to hold and use the second level domain name."

Some critics have viewed the language of 7.1.1., as forcing applicants to agree that in the event of a dispute over their domain name, they will take such dispute to arbitration rather than to court. Before thinking that this buys something, one should remember that conflicts are normally between domain name holders and trademark holders who are in no way bound to accept the results of any arbitration proceeding - especially if the aggrieved party was not in any a participant in the IAHC registry process and found out about it only after the fact. Virtually, every major case so far has been disputed by a trademark holder who comes in and wants to register, and finds the name already taken. In this case the trademark holder is unlikely to agree to arbitration.

A supporter of the IAHC process told us: "The applicant would agree that, in a case where a third party wanting to challenge the domain name chose to submit the dispute to mediation or arbitration at WIPO, the applicant would also submit to such mediation or arbitration." Looking at this a critic responded: "In other words, the applicant must accept ajudication in a forum (WIPO) which is historically friendly to traditional intellectual property holders (read "trademark owners") if they (mark owners) select that forum. Remember WIPO and the database treaty in December 1996?" [**Editor**: WIPO was pressing for unprecedented increases in copy right coverage.]

The policy area for dispute resolution is another problematic area. The predominant opinion on IAHC mail list discussions is that Registrars form their own policies and procedures which critics see as tantamount to letting them do whatever they want. IAHC has not yet decided to have the Council of Registrars create a uniform policy for dispute resolution -- one that all Registrars must accept. (An IAHC source tells us that there is a policy binding on CORE - but not as far as we can tell on each Registrar at the Registrar's operations level.) This is a matter of concern because some critics maintain that some of the non IAHC members involved in these discussions have failed to take the question of registering trademarked names with sufficient seriousness.

It is hard to see what has been created, apart from the Administrative Domain Name Challenge Panels, that will keep the Registrars from out of trademark policy trouble. The approach to trademark issues on the part of the people running the registrars will be critical. It will take very careful and astute registrars to avoid being hit by a suit. Some suggest legal action will occur regardless.

When the litigation happens, it will very likely not only hit the errant registrar but also be directed at the IANA, the ISOC and the IAHC as the organizations that brought the Registrars into being. One hopes that all of the board members and committee members involved have excellent liability insurance policies - insurance that, given the issues and exposure of involved individuals, is likely to be quite expensive.

One of the things that the IANA said was that the purpose of the competing registries was to collect money to provide liability insurance for the IANA, the ISOC, and the registries. However, until the registries have been in operation for some number of months there will be no cash flow generated that is capable of paying for insurance. Postel's July 9, 1996 draft said "this document describes policies and procedures to allow open competition for domain name registration in the ITLDs and provide the IANA with a legal and financial umbrella . . ." Under the Section Goals: at sub section 2.1 it says: "to provide the IANA with the international legal and financial umbrella of the Internet Society."

The final IAHC report it says: "the IAHC was formed at the initiative of the Internet Society at the request of the IANA. . . ." We are unable to find anything in the report about liability insurance. It seems quite possible that the IAHC will realize that it is so exposed that it may fold its tents and go home before it ever really opens for business - i.e. gets itself in trouble. Of course if one is looking for resources on the IAHC side of things, the International Telecommunication Union does have resources that it can commit should it choose to do so. These considerations would seem to favor the viability of NSI as the incumbent and make the chances of success for the IAHC small.

On another level, the situation gets more complicated because while Postel is granting the authority for operation to NSI, in the new registries draft and the IAHC documents he also grants it to ISOC. Don Heath has said on the IAHC mail list that he wants the ability of NSI to continue to function independently of ISOC and IAHC terminated as soon as possible. One would infer so that the monies it collects flow into the control of the organs established by IAHC.

## Technology Problems

A problem that emerges from having multiple entities building databases even if they are intended to be components of a shared database is the question of who keeps the authority files for the data base? Who has the responsibility of making sure the records in the database are uniform and free from conflict if they are created by disparate entities? This problem emerged in an interesting way in the last week on the net-time mail list, a closed list populated mainly by european artists and other intellectuals. There it seems that an American named Paul Garrin who lives in Newark NJ is perpetrating his own mischief. We publish as a side bar a February 16, 1997 exchange with him for several reasons. The exchange indicates the potential for mischief making by dissident groups that network policy makers are completely unaware of. It indicates a lack of awareness on the part of the dissident [Garrin] of the importance having a single authoritative copy of the DNS database. And it indicates Garrin believes that law enforcement officials (and by inference courts) may also lack that fundamental understanding.

## Who Goes to Court First?

The Garrin effort is another example of the fear uncertainty and doubt that is flourishing in the midst of the Internet governance vacuum that has allowed the DNS situation to billow into a dangerous controversy.

Why dangerous? Well it is easy to imagine that one of these people may be able to go to court and get a judge who doesn't know any better to issue a restraining order against NSI. One that might block all NSI's actions for some period of time. Hidden beyond the view of 99.9% of those involved with the Internet is the fact that NSI performs a second function for which it does not charge. It hands out blocks of IP numbers for ISPs. For a machine to have a reachable address on the Internet it must have its own unique IP number. If it does not it cannot be reached by anyone and in effect cannot be considered to have a direct Internet attachment.

If DNS becomes embroiled in legal chaos, the Internet will be hurt. People may have to start using IP numbers in web browsers and in worst case with email. But the net would still work. Break the IP registry in North America and you effectively stop the expansion of the Internet in this part of the world. Over time you would also impair the ability of those with IP numbers assigned to function. Ultimately THIS kind of breakdown would threaten the ability of the network to function.

And this is a little stated reason why those who realize the potential for trouble are moving to disentangle the allocation of IP numbers from NSI and create an independent 501(c-6) entity called the American Registry for Internet Numbers (ARIN) - one that will be legally vested with the right to distribute IP numbers to networks in the Americas. ARIN was introduced at the December IETF and an Internet discussion list on the topic has been ongoing since that time. The proposed model is similar to those currently operating in Europe and Asia. The IANA has indicated support for the initiative. The NSF has indicated that it "considers it encouraging to see a community-based consensus emerging around a sustainable model patterned on the RIPE and AP-NIC model endorsed at recent workshops." With ARIN legally chartered, if NSI and the whole DNS process does get caught up in a legal tangle, at least the IP number registry and database will be free of the problems entangling its DNS counterparts.

## It's Only a Matter of Time - the Question is How Much?

Within the past year both the Federal Networking Council Advisory Committee (FNCAC) and NSF review committee have advised the NSF to extricate itself from the NSI Cooperative agreement as soon as possible. Declare the process self-sustaining and fully commercialized and the cooperative agreement terminated. Whether the NSF will take such a step at all or any time soon is unknown. With the IAHC solution defined, the NSF might see that as a possibility that would enable it to wish IAHC luck, pass the torch and extricate itself from a maelstrom of criticism leveled at it by the network

There is one very weak link in the whole process: The authority chain. In other words the IANA - Jon Postel. The IANA's authority endorsing it is what gives NSI's registry value. The operators of the root DNS servers are willing to accept Postel's recommendations as authoritative. They carry the registry database(s) that Postel asks them to.

But what if a court were ever effectively to say to NSI: the authority of the IANA is in dispute and/ or no longer valid? Therefore you may no longer use the IANA as your authority in asking the root servers to carry your database. Or perhaps more likely were a court to say to the operators of the root servers: you may no longer restrain trade by accepting only data that the IANA deems authoritative? In a worst case scenario, you then might then find multiple groups with databases of questionable quality insisting that they be added to the root servers. If the databases conflict and the system falls apart, too bad. This probably won't happen. But it could. Being fully aware of the dangers may be the best way for all parties to avoid disaster.

hot heads.

While the question sooner or later could become one of whether IAHC and NSI get into a legal tangle, other developments seem more likely. The chances of the entire domain name system including NSI getting tied up in knots may well be slim. The chance of IAHC getting clobbered is rather large. In order to stop NSI from issuing names you would have to prove harm. But NSI has been doing this for four years without serious allegation of harm.

However, there is one very weak link in the whole process: The authority chain. In other words the IANA - Jon Postel. The IANA's authority endorsing it is what gives NSI's registry value. The operators of the root DNS servers are willing to accept Postel's recommendations as authoritative. They carry the registry database(s) that Postel asks them to.

But what if a court were ever effectively to say to NSI: the authority of the IANA is in dispute and/ or no longer valid? Therefore you may no longer use the IANA as your authority in asking the root servers to carry your database. Or perhaps more likely were a court to say to the operators of the root servers: you may no longer restrain trade by accepting only data that the IANA deems authoritative? In a worst case scenario, you then might then find multiple groups with databases of questionable quality insisting that they be added to the root servers. If the databases conflict and the system falls apart, too bad.

NSI has had four years to develop quality control. How good would be that of registry "x" with four months? If you are going to be certain to avoid conflicts, then you must have a single source of authority. But if you don't care about conflict then you might well use databases from multiple registries. If this disaster ever came to pass, the major corporate web sites would go to Congress and ask it to regulate the Internet because, at this point, only the entry of a correct IP number could get someone to a web site.

One thing is certain. At some point within the next year the DNS wars are likely to go into overdrive. The Internet has faced its first major test of self-governance of the commercial era and not done well. We see no conspiracies, no men or women consciously doing evil deeds. But we see not much leadership either. We have some very strong opinions about some of what seems to be emerging from NSI, including its plans for an initial IPO. We will defer expression of them until we see the moves that the players make during the remainder of the Cooperative Agreement.

All in all it is becoming clear that what is left of the cooperative culture of the Internet faces a major stress test, with the outcome unknown. The legal meltdown that we have described probably won't happen. But it could. Being fully aware of the dangers may be the best way for all parties to avoid disaster.

**Side Bar: Another Challenger**

## Paul Garrin Has a New Registry Called Net.Space

On February 16 we wrote to the Net time list: Paul Garrin says that he has been told by the US Department of Justice: Under US Law, the current root server hosts must provide access to their facility by their competitors on a non discriminatory basis. (We just need to ask them for it!) The U.S. Department of Justice, Anti trust division has confirmed this to the name.space legal counsel. "Your case is a carbon copy of MCI vs. AT&T" they said.

***COOK Report***: Sorry I simply don't believe that they would say such a thing. Especially "Your case is a carbon copy of MCI vs. AT&T" . How about scanning the letter putting it up on a web site so that folk can call THEM for verification? Lawyers generally don't stick their necks out which such a definitive statement. What you claim is NOT CREDIBLE. Besides the analogy is wrong. Yes the LECs have to let their customers have access to any long distance carriers, but the North American Numbering Plan does not have to be opened to any third party entity who wants to assign phone numbers from it.

**Garrin**: If the Root servers refuse, they are in violation of the law and subject to Anti-Trust violations. According to the US DOJ representative, There is no argument in this case. The law is clear in their opinion. (the case begins this month).

***COOK Report***: If the root servers refuse what? To run YOUR distributed data base? Instead of IAHCs? Or are they going to run yours on even days of the month and IAHCs on odd? And if Kashpurev hears about this, then he wants in and if they allow YOU in why won't they have to allow Alternic in? *then you get to run every third day?*

The root servers are there to give *authoritative service. ie run the one copy of dns database that is accepted as legitimate by the whole* Internet. This is by definition a *monopoly* situation. It has nothing to do with a phone network that can be used by multiple companies. What you want to do in building a distributed database sounds like what IAHC is doing. What makes you think the Internet is compelled to adopt YOUR solution rather than its own?

IAHC was formed by as close to a treaty process as the Internet has. Postel working with the Internet Society putting together folk from ITU, WIPO and other interested organizations. What gives you the right to assume that *your* solution is to be accepted and not that of ISOC, IANA, ITU and WIPO? Or are you asking that people have

**con't p. 24**

# Internet Routing Technology: Network Growth Brings  Stability & Scaling Problems

## Noel Chiappa Interview Offers Routing Technologies Tutorial & Explains Why BGP Replacement Needed

**Editor's Note**: Noel Chiappa had invited us to query him about problems in Internet routing that he sees as having the potential to seriously inhibit the continued scalability of the Internet over the next few years. We checked some other sources who agreed that the problems he had in mind were quite real. Consequently we decided to see what he was thinking about and in doing so got quite an education.

**Chiappa**: I read the article you did about you view of the state of the internet and challenges that it faces to its continued growth. It struck me that the one thing you really missed was the scalability to some degree of both the routing architecture and the current deployed routing technology. This is important because, if there is a problem in the technology, we are looking at a relatively long lead time to fix the problem.

Why? Because the way the community works, it will take a year or so to agree on how to approach the problem and another year to get all the documentation written. Then it takes a year or so at a minimum for all the vendors to  get it deployed, in their stuff and out in the field. To be totally honest, unless everyone just goes into total panic, you know, the Huns are attacking mode,it's going to take four to five years.

### A Task for IETF

***COOK Report***: Is the problem as much of one with software and standards then as of hardware?

**Chiappa**: Yes. Take a typical problem like lack of capacity. You can solve this kind of problem on a piecemeal basis. Each ISP just goes out and installs more. The solution to the routing problem demands a coordinated response across the whole internet and that is inevitably much more difficult because it is an inter organizational problem. You are talking about developing new protocols, turning them into software and getting them deploye7d. That's why you need immense lead time. The IETF is currently the only good place this can get accomplished. If the routing architecture isn't scaling well, we'd best get cracking because it is going to take quite a while to fix it.

***COOK Report***: What leads you to believe it is not scaling well?

**Chiappa**: First, the current system will simply not scale as well as some alternatives. Second, we are experiencing some problems with the current routing in the Internet. But the question to the answer is are these problems there because we have reached some fundamental limit, or is it because we have various implementation and configuration errors in various routers that are contributing to the problems we see with routing? (See remarks of Curtis Villamizar on page 16 of this issue.) There definitely is some evidence that some of our problems are due to this kind of error.

***COOK Report***: How would you define this kind of problem?

**Chiappa**: I can speak only from my view of one corner. Perhaps a corner or less of the very huge Internet, but a well traveled quarter, no less. I sit here in Virginia and my mail is kept on a computer at MIT, in Cambridge. Normally my packets go from a local ISP here to MAE East, then through a major ISP to a regional ISP in Boston. I have seen a lot of routing problems including many routing loops between two routers in a large ISP at the MAE. It is rather hard to tell what was causing the flapping out of the large Washington based ISP with traffic on its way to the MAE.

### Why Must Routing Demand Ph.D. Level Talent?

***COOK Report***: In your examination of this were you able to eliminate human error as a contributing factor to the flaps?

**Chiappa**: The problem is that it could have been very subtle. I don't know what all the knobs are on current routers. It could be that the router had been configured to be too sensitive. Configured to be less sensitive, it might not have picked up the instability. It is hard to do anything more than an educated guess.

A part of the problem is that a lot of the smaller ISPs probably don't have any technical people who understand the problems and subtleties of large scale routing. Heck I am not even sure that many of the larger ISPs have people who understand the full depth of the routing issues.

***COOK Report***: What would be some of those subtleties?

**Chiappa**: Things like dampening. Some ISP operations people worked with Cisco to implement some dampening knobs that allowed them some control. As a result they became very knowledgeable about how to use some of the more sophisticated tools to control routing instability. [Editor: Sean Doran is the principal such person of who we are aware.]  It is not clear that "Joe Local ISP" has anyone like that. I am sure that many local ISPs don't have anyone who has a clue. But you sort of understand this. What I found a little more astonishing I that I am not sure that the people at the large Washington area ISP understood all these issues.

I saw behavior in the Washington ISP that was simply inexplicable if they had anyone technically competent. When they lost the route to MIT, what would happen is that two of their routers would go into a two element routing loop.

***COOK Report***: Ping ponging it back and forth?

**Chiappa**: Yes and this would last for quite a while. I could never figure out what their technical people had in mind when they set up the routers in such a way as to allowing that to happen.

***COOK Report***: Well then, one definition of the problem seems to be a human resources situation where the network has grown faster than our ability to train newly competent people. But, how do these problems relate to the more fundamental issue of whether new protocols are needed?

**Chiappa**: You have a whole bunch of factors at work. In a system as large and as complicated as the Internet, there is no single cause for the routing instability that we are seen. The personnel cause, in a weird kind of way, does relate back to the more fundamental one of architecture. If it took Ph.D. airplane mechanics to maintain airplanes, we would not be able to afford commercial aviation. Both commercial aviation and the telephone industry work because the infrastructure of each has been designed to fit comfortably within the capabilities of the crafts people who do a lot of the work. Perhaps the routing architecture we have now is fundamentally ill designed, because, if it requires extremely highly skilled technical people all over the place to keep it running, then maybe we have taken the wrong direction. I just don't think that a routing system that needs mega wizards all over the place to keep it running can be described as a well designed system.

***COOK Report***: Do you have an example or two as to why it needs the mega wizard?

**Chiappa**: Operations people at the ISP's could give the best examples. But I think that one good example is the process of setting up the required filters. Everyone's router contains an enormous filter database which describes what routes they will accept and what ones they wont. Meshed into this is likely to be a complex set of dampening decisions. It is all tremendously complicated. The software has had to become complex, in part, because there has not been any new routing technology developed for many years.

## Kinds of Routing Protocols

There have been examples where new protocols were slight improvements over previous protocols. The whole BGP series is an improvement over EGP. Routing is divided into inter domain protocols which operate between organizations and intra domain protocols which operate within a single organization. For the intra domain we have had IGP, and within the IGP class of protocols, RIP or RIP2, and OSPF.

> If it took Ph.D. airplane mechanics to maintain airplanes, we would not be able to afford commercial aviation. Both commercial aviation and the telephone industry work because the infrastructure of each has been designed to fit comfortably within the capabilities of the crafts people who do a lot of the work. Perhaps the routing architecture we have now is fundamentally ill designed, because, if it requires extremely highly skilled technical people all over the place to keep it running, then maybe we have taken the wrong direction.

***COOK Report***: When you said intra domain, we thought you were talking about the territory covered by a single corporate campus network. However can this also apply to a metropolitan area network with a dozen or so pops as hubs? For example, Avi Freedman's Net Access in Philadelphia, where he was installing OSPF in the routers under his control this summer.

**Chiappa**: This is where I have heard different things from different people. OSPF was designed to run extremely large networks with thousands of routers. There have been some people who report to me that they have had problems in using OSPF in extremely large configurations. I never really understood whether it was a problem with the protocol, or whether they simply were trying to use OSPF incorrectly. As far as I am concerned, OSPF is technically the most capable of scaling of all the intra domain (IGP) routing protocols.

***COOK Report***: But can a metropolitan area ISP use OSPF within its autonomous system?

**Chiappa**: Sure. Any time you have a collection of routers connected together and maintained by a single administrative control, you have an intra domain routing protocol application. Let me add that I think that OSPF is the best and most capable of the IGP protocols. But, in a situation where the routing was collapsing in the part of the network *outside* the OSPF user, OSPF might then not be the best protocol for use in such a situation.

## A Mini Tutorial on Routing

Here is the issue. OSPF tries to respond more quickly to changes. Let me offer your readers a mini tutorial in routing. There are basically two ways to do routing. Way one is destination-vector routing. This gives you a vector, that is an array, of destinations, for each destination. It also gives you some information - at the very least, the location of the 'next hop'. Now you will hear a couple of terms mentioned underneath this umbrella: distance vector which means what it gives you as a vector of distances to those destinations; and path vector a list of paths (which is what BGP is). They are all a hard part of the same general concept. This concept is that what you get from your neighboring router is this big long list which is basically a copy of his routing table. It says for destination "a", here is some information, for destination "b" here is some information, and so on for each additional destination that the router knows about. In the distance-vector variant of destination-vector routing, when the neighboring router gives you for each destination the distance from it to that destination, you would take his incoming routing data, which is basically a complete copy of his routing table, and ask if his path for that destination is shorter than the one you already have? If so let me replace my route with the one I just got from my neighbor. So the basic concept of destination-vector is that you get a complete copy, of your neighbor's routing table.

The other basic school of routing technology, is what I call map distribution. In this, instead of getting a routing table, you get a map of the network, and use that to create routing table entries. Here the data you receive has originated all across the network. In the simplest variant, every router, every so often, makes up a packet that says I am router "x" and I am connected to routers a, b, and c. This information gets flooded throughout the network. Every other router gets a copy of such a packet and every router sends such a packet. With such information in hand, you may construct a map of the entire network in terms of who is connected to what.

What happens next is that everyone runs a routing algorithm to figure out how to get from each router to any other place in the network. They use the results of this to fill in their routing tables. If I am router "x" and I am trying to figure out how to get to destination "a" at some far off network point, just by looking at the map, I can tell what the shortest path is from me to "a". Some my routing table says for destination "a" send it to whatever the first hop is on that shortest path.

The map distribution algorithms will inherently respond faster to changes in network conditions. For example when a link goes down in a map distribution system, the two routers on the end of the link notice that the link has failed. They flood through the entire network a very small packet which says our connectivity has changed. Routers are not required to do any computation on that packet before passing it on. Consequently it floods throughout the entire network very quickly. Updates to routing tables take place essentially after the packet has already been handed on.

***COOK Report***: Is the packet like a network co-ordinate? Note the co ordinate, pass the packet on, and then update your map.

**Chiappa**: The point is that when you get that packet in, you can pass it along to your neighbor, before you have done any computations with it. The changed information floods across the network so quickly that everyone effectively updates their routing tables in parallel. Now think about what happens, if a link fails, in a destination vector system. The guys who are next to the failed link have to recalculate their tables and, only when they are done, may they pass on the results to their neighbors who have to recalculate and then pass on and so on. Now you can see that what we have is a wave of change in the routing tables that starts out at one place and spreads like ripples across the surface of a pond after a rock has been thrown in. If such changes in routing are not correctly dampened, it can oscillate and you can get waves of change washing back and forth and colliding with each other.

***COOK Report***: But this is the way the net works now. You have just defined a classic route flap.

**Chiappa**: Absolutely. BGP is destination vector routing. The problem is not that the routing changes when something breaks - you want that, that's how you recover from changes. So a single 'flap' is not a problem. The problem is when you get lots of changes, how does the system respond to that?

***COOK Report***: But, on the other hand, if you did not implement destination map routing with absolute precision, would you have an even worse situation?

**Chiappa**: As long as routing is in a normal operating domain, map based systems are better because you can run a larger system with faster changes by means of a map based algorithm.

***COOK Report***: How would you define "normal operating domain"?

**Chiappa**: A normal operating domain means operation not disrupted by route flaps. However if you get into total overload with your routing going simply bonkers and bananas, the fast responsiveness of the map based systems might work against you. It gets quite complicated because we are running mapped based systems only in areas of the Internet like within a particular ISP or corporation. At the interface between those areas of the network and everyone else, each map based system has to talk with an destination vector system.

***COOK Report***: Is OSPF a map based system?

**Chiappa**: Yes. Map based systems have multiple implementations. RIP which is a distance vector algorithm is a destination vector algorithm as is BGP which is a path vector algorithm. OSPF is an example of what is called a link state algorithm which is one kind of map distribution system. But there are other kinds of map distribution architectures that are nothing like OSPF.

If you have an area that is running OSPF, at the edges of that area, it must interoperate with the overall internet routing architecture which is basically destination vector. Now if the overall destination vector starts to go berserk, its not clear to me that having an extremely fast responding IGP is good because it would propagate the exterior craziness into the inside of your system extremely quickly. Maybe you don't want that. You might prefer a slow responding system if the world out side you is going crazy.

***COOK Report***: Is one of the protocol developments needed something to provide a smooth transition between the OSPF and destination vector segments of the net?

**Chiappa**: Not really because, at the interdomain level, the routing is all destination vector algorithm, while the OSPF issues are really just local ones that have nothing to do with over all Internet routing stability. It's overall Internet routing stability that I'm concerned with here, and that will be an issue no matter whan the IGP technology is". OSPF's advantages are primarily in its responsiveness to changes. You must realize that you have in every router something generically known as the routing interchange box. This is a piece of software that allows you to interchange data between various routing protocols. The operation of this interchange, is I think, not very well specified. If you are talking about a protocol that is written down in a book, such a protocol is, by definition, very well specified and must be interoperable so that router vendor "x" can talk to router vendor "y" over the wire. But the interchange between routing protocols is often much less well specified. It is a situation where different vendors will have different capabilities available.

Having said all this, let me go back to my initial comment about how the routing problems in the internet have many different causes. Since I don't know for sure what is causing current router problems, I don't know for sure to what degree work there would be helpful. But there are a number of reasons to think that as the Internet continues to grow, the current routing architecture is going to become less and less suitable.

Now one of the reasons we have already touched on - that it requires fairly technically savvy people to stoke and feed the current routing system and that the supply of these people is probably not growing as fast as the current internet is.

## Stability Problems in Large Scale Destination Vector Systems

But there is another one that is of much concern to me as a theoretician. This is the existence of a general stability problem with destination vector algorithms at really large scales. Namely tens of thousands of nodes and 50,000 to 100,000 routes. Wide area routing is all destination vector. It is all BGP. When you consider the behavior of the wide area system overall you must define it as a pure destination vector system.

If you draw a map of the topology of the internet and represent all areas that are running an interdomain routing protocols as single dots, all the routing running between those dots will be destination vector technology. What's more, the system you are looking at is a pure BGP system. One of the beauties of the internet is that you can get from anywhere to anywhere else. Everyone knows where everything is. The bug that comes with this is that if in some corner of the Internet the routing is going berserk and generating lots of updates, the current routing algorithm ensures that the erratic behavior will ripple across the entire network.

Now there are routing systems that have been designed to avoid the effect of bounce back. The rock hits the pond. The ripples slosh to the edge where they disappear. But lets assume I am throwing rocks into the pond every second at the same place - one rock after another. It will result n the surface of the pond being permanently disturbed because I am continually adding new disturbances to the routing. Thus routing instability at a single point can affect routing stability over the entire internet. In fact there is some evidence that this is to some degree what we are seeing. People have monitored an individual backbone router and seen how many routing up dates it got in the course of a day and it turns out that a large percentage of them are for a small number of destinations. So to the extent that we have routing instability in the Internet, beanheads may be partly to blame.

***COOK Report***: Do you think that at some point in the not too distant future, people who run backbone routers in the Internet may have to be certified as to their ability to do so without harm to the rest of the network?

**Chiappa**: That could be. This gets back to my earlier point that to the degree we have a routing architecture that takes people with Ph.D.s to feed it, we are suffering from poor design. Obviously the long term five year fix is to make it more robust, but the short term fix might be to require certification.

There has been some attempt to automate the process to some degree. To put an Artificial Intelligence front end on it basically. I am not sure how well its really working.

***COOK Report***: Is this some of what Merit is playing around with?

**Chiappa**: It is probably some of what they are doing. The point is that an individual site is still free not to use the AI tools. The routing we now have is inherently vulnerable to certain kin10ds of bean headedness.

***COOK Report***: If the community doesn't start to develop these protocols, it is easy to imagine how the net might soon wind up with the need for some overall kind of routing authority or even police.

**Chiappa**: I suspect the cure there would be worse than the disease. The stability (dampening) stuff that ISP operations personnel [Editor: Sean Doran] worked on was an attempt on a localized basis to solve this problem. One large ISP [**Editor**: Sprint] had it set to where, if a particular destination flapped more than a certain number of times n a certain time period, it was taken off line for a specified period of time. They were on a unilateral basis attempting to filter out the destinations that were causing a lot of the rout flap. I am pretty sure this capability is in Cisco routers now. All one need do is to turn it on.

***COOK Report***: What happens should the use of these dampening techniques become much more wide spread?

**Chiappa**: It removes a certain class of problem- routing instability cause by a few local destinations that are oscillating up and down.

***COOK Report***: What kinds of problems are left then?

## Stabilization Time Problems

**Chiappa**: Let me step back in to the theory world for a moment. Let's imagine that we are trying to write an equation for a given kind of routing technology that predicts what the stabilization curve is. Now imagine a theoretical network topology, that is to say a bunch of nodes that you connect with wires in some random way. Now imagine that you install on that network routing technology of type X where it might be destination vector, path vector, link state or whatever. You have set all this up on a computer to model it. Now you take one link and cut it out. You then watch and see how long it takes for the routing to stabilize.

Here's what I mean by stabilization. Assume that routing is running smoothly and everyone has a correct non looping path to everyone else. Routing is stable. Now when you snap a link, all traffic flows that don't cross it will continue to operate fine. All traffic that tries to cross the link will fall into a black hole. Now when the link has snapped and a topology change has occurred, we want to fix everyone's routing tables to take the change into account and reestablish good routing for everyone.

Now we have a chaotic period where traffic moves with routes flapping and may fall on the floor. It may even loop for a while if someone gets an update before someone else does and as a result two different routes each think that they are the single best way to a destination. This is usually a fairly temporary situation. Eventually the whole thing stabilizes and everyone has a good route. The time period between the link failure and point where everything is completely back to normal is called the stabilization time.

You can see that, if you take a sample network one to two hundred nodes in size, there are particular topology changes that you can make that will produce a much shorter stabilization time than others - depending on your particular routing technology. So for a given routing technology and a given topology if you serially look at the response to every possible topology change and plot the results on a graph you get a histogram which has the shape of a curve. The histogram is the response curve of that particular routing technology to that particular topology.

Now keep the routing technology the same and generate a different topology. The histogram generated by that configuration will be similar but not quite identical to the first. Now generate all possible routing topologies and average the curves for all of them. What you have generated is a stabilization response curve for that routing technology. It says on average how routing will stabilize in response to a topology change.

The shape of the curve will be dependent on the routing technology. Now one of the nice things about link state routing technology is that, not only is the time for stabilization extremely short, but it is also very bounded. You can even predict the amount of time - so much to flood the packets, and so much to recompute the routes. Do these two things and you are done. Destination vector systems that make corrections sequentially, rather than in parallel, will inherently take longer to stabilize.

## Writing an Equation that Predicts Stabilization Time

We are going to try to write an equation that predicts the shape of the curve or the stabilization time of the network for a particular routing technology. Such an equation would be tremendously complicated. I cannot pretend to tell you exactly what it would look like but I can tell you some of the terms that would be in it. One of these terms that will appear in the equation, in a number of places, is the size of the network, because the bigger the network is, the longer it is going to take to stabilize. Another term that will appear in it is the speed of the links, because you have to ship the information around and, the slower the links, the longer this will take. Another term that will be there is the speed of the routers. And yet another necessary term is the speed of light. Depending on how far routers are apart from one another, it can take significant real time to get the bits from point a to point b.

There are some people who think we can solve the routing growth problem by throwing more bandwidth and CPU power at it. What I say to these people is: write down what you think the equation is for stabilization time and then set bandwidth and computing power to infinite. If your equation predicts zero stabilization time, there is something wrong. Because the speed of light is a constant and has to be in the equation as well.

The stabilization time for the Internet is going to be related to a number of things. One of these is the kind of routing technology that you are using. Another is the size of the network in terms of the destinations that you must track. Now when I say size I should emphasize that, to a routing person, size means two things. One component of 'size' is the number of switches (routers) involved in this very large distributed, dynamic computation. The other factor is the number of destinations being tracked. These two are usually grow at about the same rate.

***COOK Report***: What is the relationship between the number of destinations being tracked and the number of routes advertised at the defaultless core of the Internet?

**Chiappa**: The two are essentially the same. The number of destinations being tracked is equal basically to the number of routing table entries you have. When I said the stabilization time would include the size of the network, it really includes two different sizes. One size is the number of devices involved in a distributed dynamic computation, the other size is the number of entries in the routing table. The current system is a hop-by hop system. For the path between "x" and "y" to stabilize and work, all the routers between "x" and "y" basically have to come to the same idea about what the right path is between "x" and "y". It is a distributed computation, but it is also dynamic, because the routing calculation happens continuously when topology changes occur.

Here is the critical thing about the network. If you draw a graph that says here is the stabilization time of the network as it gets larger, you will discover that such stabilization time starts to increase as the network gets larger. Now we come to the most critical consideration. Topology change is what causes the need to restabilize. What is the time between topology changes as the network increases in size? That obviously gets smaller.

***COOK Report***: Because there are more things that can go wrong in a larger network.

**Chiappa**: Yes. Because as the network gets larger the mean time between the failure of some piece of equipment somewhere in the network gets smaller because the network is made up of so incredibly many pieces of equipment, each of which may fail. The MTBF for something somewhere in the network is going down over time, even if the MTBF for each individual component is staying the same, or even getting better. It's the growth in the number of components that means the MTBF for the network overall is getting smaller."

## Two Intersecting Curves: Time to Stabilize and Time Between Need to Do So

Now you have two curves. One is the time to stabilize which increases as the network gets larger, and the other is the decreasing time between failure of some component of the network. What happens when the two curves cross? After all increased failures lead to the need to restabilize more and more, but the larger the network the longer it takes to dampen the effects of increased failure. One can see the potential of continued hardware failures somewhere in the network causing a need for routing stabilization that is never completed before the next outage causes another need for new stabilization.

Now you have two curves. One is the time to stabilize which increases as the network gets larger, and the other is the decreasing time between failure of some component of the network. What happens when the two curves cross? After all increased failures lead to the need to restabilize more and more, but the larger the network the longer it takes to dampen the effects of increased failure. One can see the potential of continued hardware failures somewhere in the network causing a need for routing stabilization that is never completed before the next outage causes another need for new stabilization.

***COOK Report***: In other words, the whole network crashes.

**Chiappa**: Right. We have seen this happen in networks. It's actually sort of complex, since not all changes affect all destinations. One change might effect a completely different set of destinations from another, in which case they don't interfere. To go back to our pond analogy, it's like having multiple 'virtual' ponds stacked one above another. A rock into one can produce a pattern of ripples which is totally separate from the pattern on another pond. On the other hand, some changes do effect more destinations than others. To continue with the pond model, it's like you have N ponds stacked up, one for each destination. Each rock (i.e. topology changes) causes ripples in a different set of ponds - and sometimes the pond sets affected by two (or more) rocks have an overlap. So you can see it gets pretty complicated pretty quickly.

But if the network gets too large, you can indeed see the routing havingproblems stabilizing. I don't want to sound like the sky is falling. There are solutions. But then we get back to my original position that fundamental changes in technology might be necessary and that these changes take about five years.

***COOK Report***: What then can be done?

Chiappa: We have a number of possible approaches. Different routing technologies have different stabilization curves. Therefore you can move the stabilization curve out a ways by using a different routing technology. I happen to think that we have chosen in destination vector technology just about the worst routing technology. So one approach would be to switch to a different routing technology.

***COOK Report***: In other words, away from BGP?

**Chiappa**: Yes. Another possible approach is to reduce the number of destinations that the routing is tracking. To use more aggregation to reduce the number of routes in the default free core.

***COOK Report***: CIDR in other words?

**Chiappa**: CIDR has always been important. The whole reason for CIDR has been to control the growth in the routing. Remember that I talked about that hypothetical equation. Two of the terms there are the speed of the links and speed of the routers. Obviously this faster technology is getting better and, to some degree this is increasing the size of the network that we can run. Now CIDR is attempting to keep the growth in routing table size to be less than the curve of what is the maximum routing table size that we can run with this particular routing software, in this case, BGP. Such maximum size will grow over time as the hardware accounts more capable. So by using CIDR, you can support a bigger and bigger network over time as the hardware becomes more capable. One of the reasons that we have been pushing on CIDR is that we knew the tables growing so fast that they were outstripping the ability of technology to keep up. We had to implement address aggregation in order to reduce the number of destinations that had to the tracked.

***COOK Report***: We are publishing in this issue, an interview with Tony Bates about the CIDR list in which Tony asserts, at one point, that as the size and complexity of routing grow, human error becomes a factor in introducing routing errors. That we can get leakages of a thousand routes or more into the core backbone by simple human error. It would seem to be the Ph.D. routing engineer factor appearing once again.

**Chiappa**: Absolutely. We get back to the same point at the technology was not designed to scale in terms of the people issues either. The whole Internet is now suffering tremendously from the fact that it was never designed to grow to this size.

***COOK Report***: So even if you solve the technology problem, you are still likely to be bitten by the human interface problem.

**Chiappa**: We probably have a whole range of problems interacting. These may range from buggy implementations to instances of poor configuration due to unskilled personnel. But the network is also probably starting to run close to some fundamental limits given the design of its routing technology. Now assume perfect humans. Even if we do that, for a given set of routing technologies and a given set of hardware capabilities, there is a maximum limit that may be supported in terms of routing destinations.

Remember the two intersecting curves we talked about? For a given routing technology in terms of software and a given hardware base in terms of hardware capabilities, there is a maximum size network that can be supported in terms of number of destinations [routes advertised at the defaultless core], even assuming perfect humans.

***COOK Report***: But let's imagine we had twenty routing gurus in the room. Might not some of them say that one way to fix this is to be prepared to have a very authoritarian mechanism at some point in the future to clamp down on the number of routes that we do allow to be advertised.

## Solution by Aggregation or by New Routing Technology?

**Chiappa**: This issue has been debated for some years now among what I would call the more cognoscenti members of the community. The problem is that clamping down on the number of routes more or less means that people have to renumber when they change where they are connected to the Internet, i.e. change ISP's. Renumbering has been a political hot potato. The feeling of a lot of us is that the right thing to do is to charge for routes in such a way that the charge for a route would be proportional to the scope over which your route was advertised. So if you have a route advertised across the entire Internet, such a route advertisement would cost much more than a route that went no further than your local ISP. If you do this, a lot of people might find that it is more cost effective to renumber than to ask for routing for your old IP numbers. Of course part of the problem is that the current system for changing addresses is a real pain.

***COOK Report***: To what extent can it the argued that route aggregation and charging for routes is an easier, better, or more appropriate way to fix this than developing a whole new routing technology that replaces BGP? These are two alternative approaches. Where do each of them lead over the next three, five, or even ten years?

**Chiappa**: That is an excellent question. In fact, it's probably the central question in this whole issue. Unfortunately the current situation is very complicated. Everything else being equal, it might be valid to simply say we are just going to stick with more aggregation. But there are a number of different factors that come into play.

***COOK Report***: Well, are there limits to the use of aggregation over the long term to solve the problem?

**Chiappa**: None that I know of. In fact the only technical approach that offers us the ability to grow indefinitely is indefinite route aggregation. If I want to build a network of theoretically infinite size, the only technology that I know of that will permit this is aggregation. However there are a number of other factors that come into play here. One is the sheer difficulty of changing addresses. At a certain point you have to say that the cost involved in renumbering may be too high

You asked "To what extent can it the argued that route aggregation is an easier, better, or more appropriate way to fix this than developing a whole new routing technology", which is an excellent question, and I'd like to explore that a bit.

In one sense, the answer is yes, aggregation is The Answer: aggregation of some sort is the only approach that will allow the network to expand basically indefinitely. As such, maybe it's the *only* answer we need, and we don't need to look for anything else. However, there are things that might drive us to needing better route calculation technology.

Remember that I said that the stabilization time equation contains a number of terms, two which are related to the size of the network, those being i)the number of destinations (a.k.a. the number of routing table entries), and ii) the number of boxes participating in the distributed calculation. Now, aggregation (a.k.a. CIDR) is mostly being used in a way which is primarily useful in holding down the first, the number of destinations.

If you want to grow the number of boxes that are involved in the calculations, grow it faster than the growth curve over time that more powerful technology will give you, you sort of have to go to better route calculation technology. Now, growth in the Internet will probably include growth in the number of routers in the default-free zone, so CIDR won't help us particularly if the term(s) in the stabilization equation which depend on the number of routers come to domination stabilization time - so there's a factor that might drive us to new route calculation technology.

This is actually complicated, since the calculations are already sort of divided up, so that only the routers in a particular area do routing calculations for traffic paths inside that area. So, all routers inside an area don't count in the "total number of routers" for the stabilization time for the enclosing area - they are (basically) two separate distributed computations. That's why I referred to the "default-free zone", since that 'area' is likely to be the one that runs up against scaling problems in the absolute number of routers first.

Now, very astute readers will no doubt be saying "ack, the sky is falling, we cannot continue to grow the Internet indefinitely even with i) CIDR and ii) new route calculation technology, since as the Internet grows, *eventually* we will have too many routers in the DFZ, no matter what new technology (hardware or software) we deploy". This is certainly a legitimate point.

What I think it points out is that we will have to find ways to limit the number of "top-level" routers in the system, and not grow that set indefinitely. Things like CIDR can help here, if it allows us to 'demote' routers, from the set that have to be at the top level, to more local router sets, so we can use recursive-divide-and-conquer to make things scale OK.

But that kind of thing, forcing people down to lower levels, might be something we'd want to avoid, and to the extent that we want to continue growing the number of top-level routers, that also might be something that would force us into deploying new route calcuation technology.

However, there is a limit to how much we can get out of new technology - at a certain point, we will have to limit the number of routers in any 'area', and use aggregation technique to do that - and that will include the "top level" area.

## Defining an Abstraction Hierarchy

I want to define what I call an abstraction hierarchy. Draw four dots in a square and label them 1, 2, 3, 4 going either clockwise or counter clockwise but not diagonally. Draw the actual square connecting the four together. The dots are routers and the lines connecting them are links between them. Now draw a circle around one and two and label it "a" and draw a circle around 3 an 4 and label it "b". You are drawing two different things on one piece of paper. The square with the four dots is the physical topology. The other is that you are starting to draw the addressing system that we use to name them. Assume "a" is an area and "b"is an area. Then the nodes have addresses in the form of a.1, a.2, b.3, and b.4.

Next draw an even larger circle around the whole thing and label that "x". You can draw an addressing hierarchy abstraction diagram which shows the relationship of the various naming abstractions which are represented by the circles. Have you ever seen in Unix handbooks the pictures of unix tree structures file systems? Take the first diagram and next to it make what will become our abstraction hierarchy. Start with a large "X" at the top draw a line on each side extending out from the big "X" at a 45 degree angle so that you have a pair of "legs" to support the "X". Label the first leg "a" and the second "b". Now forking off from the bottom of each put two more legs labeling the first "1", the second "2", the third "3" and the last "4". On the one side of the hierarchy you can get a node with the hierarchical address represented as "x.a.1" and on the other "x.b.1", "x.b.2" and so on. This represents the hierarchy of the naming abstractions for this particular piece of connectivity. This naming tree is related to but is not the same at all as the actual physical connectivity.

Here's the problem. A network is not a static thing. The connectivity within it changes all the time. People put new links in and take old links out. Now one of the reasons we do the aggregation is to reduce the number of destinations and make the routing reasonable. As the topology changes the abstraction hierarchy that you want to use with the topology also changes. But changing the abstraction hierarchy means changing the addresses within many nodes. With the current IP technology this is a particularly painful process.

## Distinguishing Physical Connectivity from the Abstraction Hierarchy

From here on out I think it will be useful to you, and in fact I wish everybody in the Internet would understand this very well. The problem is that people don't distinguish in their minds the difference between the physical connectivity map and the abstraction hierarchy. They will often try to draw the two of them in one diagram and will wind up getting very confused.

***COOK Report***: This happened more than a year ago in our interview with Dave Crocker in which you were involved did it not? Wasn't this the contradiction involved there?

**Chiappa**: Yes. Basically to say a few words about geographic addressing: for routing to scale there has to be a connection between the physical connectivity and the addressing. Either the addressing has to following the connectivity or the connectivity has to follow the addressing. Geographic addressing says we are going to make the connectivity follow the addressing. But fortunately the Internet has no police force that can force connectivity to follow addressing. The only option available to us is to have addressing follow connectivity.

In fact Yakov Rekhter has an argument that says having the connectivity follow the addressing is impossible because people put connectivity in based on traffic patterns. Connectivity is driven by traffic patterns that are in turn driven by real users.

***COOK Report***: Doing it differently would be like telling New York City that if you want to talk a lot with Los Angeles you have to be moved to California in order to do so.

**Chiappa**: Exactly. You hit the nail right on the head. The traffic follows the users and the connectivity follows the traffic. And once the connectivity is set the addressing has to follow the connectivity. The reason I got into all this is that you had an extremely good question which was do we need new routing technology at all in terms of the software. Can't we just simply use aggregation? There are a couple of reasons for thinking probably not. One of them is this issue that the net gets larger you have to keep changing the addressing to match the connectivity and renumbering becomes more and more difficult. The other thing is that I think people are going to want their routing to do more and more for them.

***COOK Report***: Those criteria exist regardless of the number of routes advertised at the defaultless core of the Internet. Correct?

**Chiappa**: Correct. People need more from their routing. For example look at a problem that I had. I was hooked up to a particular ISP. And the way that internet routing worked was to decide to route my traffic over this particular link from MAE East to the ISP on its way to MIT. That link was tremendously overloaded with loss rates of 40% in the afternoons. You cannot run TCP over a link with 40% loss rates. I would work until 11 am and log off and wouldn't log back on until late at night. Because of the way that Internet routing works, I could not force my traffic to take any other path. There might have been alternative paths that would have been perfectly acceptable, but I simply could not get to them because the routing allows me no control over it. To solve my problem, I had to sign up with a different local ISP - one that happened to be connected to another part of the network with the result that the path from me through them to MIT was different.

***COOK Report***: The problem is that the unpredictable way in which the Internet grows means that one month from now or 17 months from now, your good path through the new ISP might be clogged up.

**Chiappa**: That's right. And that is why I kept my account with the original ISP as well so that I could swap back and forth between the two to get different paths as network conditions demanded. This is a really brute force technique. It would be much simpler if I could just say to the routing where it should go.

## User Controllable Routing - How Distant a Goal?

It certainly was not the fault of the first ISP which is very well run, but suffers from a problem not under its control since the congestion was on a link on the other side of the MAE three ISPs away. People are going to want the routing to do more for them anyway. Many things are pushing us toward advanced routing technology.

But your original question was an extremely good one. It is absolutely true that routing can continue to scale for ever by the process of further aggregation. However aggregation will become more difficult. For a lot of human factors reasons people don't want to use fully aggregatable addresses. Not to mention the additional hassles that procedures like multi homing cause.

## Ipv6 to the Rescue?

***COOK Report***: But won't many people who read this discussion say: we have this new protocol called IPv6 that will solve all our problems?

**Chiappa**: unfortunately IPv6 doesn't solve anything at all. There are two things in IPv6 that might help. First it is supposed to allow easier re adressing. That would certainly be a help. Because, if it is easier to change an address, it will be easier to do more aggregation. But what is the payback from doing more aggregation? Unless there is a dollar cost in not aggregating, people won't aggregate. Secondly, addresses must be assigned in a way that has topological significance. IP addresses could simply be handed out one after the other in the same way in which we hand out ethernet card numbers. Now organizing the IP numbers and cramming them all into a 32 bit address is rather tricky. It might be easier to do this organization with a few more bits to work with.

But now that I think about it, it doesn't really matter. The issue is the total number of routing table entries and this is likely to be pretty much the same no matter how long the addresses are.

***COOK Report***: So you are saying that IPv6, other than the ability to renumber more easily, doesn't buy us much at all?

**Chiappa**: Yes, at least for the routing scaling issue.

***COOK Report***: Some research sites have begun to run an IPv6 backbone. If you look at a timetable for a move to IPv6, what is your gut level guess? Two or three years before and significant part changes and 5 or 10 years before the majority changes?

**Chiappa**: My personal opinion is that IPv6 is never going to happen.

***COOK Report***: Because the expense of making the change-over will be very large for every organization that undertakes it? And because an organization may not be willing to undertake the change unless it is certain that nearly everyone else will make the change, so that its effort becomes worth while.

**Chiappa**: Yes indeed. You will hear a lot of things about IPv6 but fundamentally IPv6 is a solution to only one problem, and that problem is running out of 32 bit address space. This is really it's primary benefit. As an engineer, I must sit here and ask: as opposed to the cost and benefits of other approaches like NAT boxes, and private addresses and firewalls, what are the cost and benefits of doing the enormous work of switching to IPv6? Now what you find is that IPv6, as a solution to the address space problem, really only works if almost everyone else converts to IPv6 also. We have the classic chicken and egg problem, because it is not useful until almost everyone has done it, but few will do it until it is useful.

We can draw a lot of insight from the CLNP situation of some years back. What I will define as the "CLNP rule" is that for a new internetworking technology to succeed, it has to offer substantially improved capabilities over what people currently have. Let's assume I want to sell a new text editor. If I get ten percent of the market, that group of people may well be perfectly happy with it, if it does what they want it to do. However, communications protocols are substantially different than text editors, because the whole point of such a protocol is to communicate with other people. An incompatible communications protocol limits the number of people to whom you can talk. Consequently, a communications protocol, with only ten percent of the market, will die very quickly. The only way to break that barrier is to be very, very much better than what is out there already. IP succeeded because there were so many important things that x.25, its original commercial competition could do. CLNP did not succeed for one fundamental reason. It was different but not different enough. I am convinced that IPv6 will suffer the same fate. It is simply not different enough, and there are other solutions to the problem that it is intended to solve.

***COOK Report***: So if we want to have a means other than route aggregation to solve the critical routing problems inherent in the growth of the Internet, we need something to replace the currently predominant destination-vector based protocol known as BGP. If this is the conclusion to which you have been leading us, who must recognize this as a critical problem? When and where? What needs to happen?

**Chiappa**: You must start by understanding at better aggregation is always going to be a part of the solution. We cannot build a routing technology that works without aggregation. But when you talk about a routing technology, do you want to build a technology with today's capabilities, or one substantially exceeding these capabilities?

***COOK Report***: So you are saying, among other things, that we could make a new routing technology that would give end users the ability to determine the routes that their packets will travel on from their ISP's host to their destination.

**Chiappa**: I sure am saying exactly this. Approach number one is more route aggregation. Approach number two is new routing technology. To the extent that all you get from a new routing technology is a certain amount of growth potential, you will discover a general consensus that the cost benefit ratio of this is lower than the cost benefit ratio of more aggregation. That is to say that more aggregation is probably a cheaper and better solution to the problem than new routing technology. Or to put it another way we might do new routing technology but only if we can derive additional benefits from it. Those benefits would probably include the ability for users to control the path of their traffic, but would almost certainly include other features as well"

***COOK Report***: Is anyone in IETF talking seriously about doing this?

## Technology Development or Deployment?

**Chiappa**: Oh, there has been new routing technology work underway for years. The technology is pretty much all done. The problem is getting consensus on what to deploy. In the mid eighties there was a group called the open routing working group. This group did a lot of useful discussion about what the path toward the future should be. I was a part of that working group and I came up with a set of ideas some of which were turned into a protocol called IDPR. That was actually field tested. However, it turned out that this was really not the right answer. A follow on called Nimrod, in which I am involved, was proposed.

Now there are some other approaches. These include a group of people around Deborah Estrin. What they are doing was called SDRP at one point, and then called Explicit Routing. There is sort of a general consensus on what future evolution in routing needs to look like - a consensus among routing wizards at least. I am not sure that the rest of the Internet has really bought into all this however.

Here is the problem. Let's talk a bit about what makes a protocol successful. It really helps for that protocol to have a property that I call self deployability. The concept is that if you have a new protocol that, when one percent of the people are using it makes their life a lot better, it will snowball. The web is a perfect example of this.

Let's imagine that one percent of the people in the network deployed a new routing technology. Big deal. Does it really make life that much better for them? The answer is probably not. Deployment in isolated areas of the Internet simply doesn't cut it. Until everyone is using such a new routing technology, it simply isn't of that much use. Therefore deploying a new routing technology is going to be terribly difficult.

***COOK Report***: Well, who is working on this? Who are the people or forces that could produce significant change in the next two or three years?

**Chiappa**: Nothing is going too happen on as short a time line as two or three years. It's simply impossible. I think what must happen first is that the users become educated enough to realize that these capabilities are possible and then start to demand them. We then need an understanding of the cost of doing it and a consensus that we should proceed.

***COOK Report***: So the problem is as much wound up with the sophistication of Internet consumers as it is with the creation of the technology.

**Chiappa**: Absolutely. The things that some people would like to do fundamentally cannot be done with the current Internet routing model. The current routing characteristics of the Internet architecture are two fold. First the computations on a system wide basis are destination - vector. Second routing decisions are taken on what they call a hop-by-hop basis. Each router makes an independent decision on where it will send the packets next. My belief, after studying the problem for many years, is that we have to change both of those in order to move forward. This would be a very radical step. People don't like radical things. We have a large group of engineers responsible for a very large part of the world's communications infrastructure. Given that situation there is a certain natural conservatism against taking radical steps.

***COOK Report***: when you look at what Bob Moskowitz and the automotive network exchange want to achieve, we wonder, if we could somehow install this new routing technology tomorrow, whether it would satisfy many of their demands?

**Chiappa**: If they are really that determined those people have a large enough chunk of money in their hands to go to someone and say: you guys develop this new stuff. They have enough money that they could make it happen. And maybe this is what it will take to generate the next generation of routing technology. It will require someone with a large chunk of the network under their control, be it the automotive guys or the Internet II guys, and enough cash to say let's develop this next generation routing stuff and use it on our portion of the network.

--------------

**Scott Bradner** asked us to let him read the interview text. On completion he commented: 'good stuff! About the only thing I might add is that there is new work going on in the IPv6 working group (Mike O'Dell's GSE addressing proposal) which may make IPv6 "enough different" on the routing side to help it succeed. Mike's proposal does address some of Noel's points but not all. Note also that IPv6's "source route header" is based on the ERP that Noel mentions but the ERP working group seems to have fizzeled out.'

The COOK Report Looks at Routing

# Why Aggregation is a Difficult Tool to Use in Solving Routing Scaling and Stability Issues

## Routing Expert Curtis Villamizar Speaks Out

From Internet Performance Measurement and Analysis (IPMA) http://www.merit.edu/ipma/press/curtis.html [**Note**: In January 1997 Curtis Villamizar provided this summary in response to a reporter's emailed questions. Surprisingly, none of the information was used in the article.]

BGP is an incremental protocol. That means that BGP does not periodically retransmit everything it knows as do protocols such as RIP. This is an important feature of BGP, given the 40,000 routes and upwards of 50 routing peers that many routers are faced with. It is not terribly surprising that across the globe there are always circuits or equipment misbehaving somewhere and therefore route change.

In principle distant routes should appear as aggregates which rarely change. An operational issue that is frequently discussed is the status of route aggregation in the network. Portions of the Internet topology are not well aggregated. This results in the 40,000 routes, a routing table larger than it should be. It also results in route changes to very small and sometimes very distant portions of the topology, even routes to individual networks, being propogated globally.

Doing an near perfect job of aggregation is a very difficult task. Aggregation of a site, such as a single building or campus has proven quite easy. Aggregation of multiple sites has proven somewhat more difficult since individual sites which are multihomed (connected to more than one provider for redundancy) or have moved to another provider require special accomodation in the form of a more specific prefix.

Aggregation across a very large provider has proven particularly difficult. Large providers are characterized as those whose customers include other Internet service providers or cover a very large geographic region, such as the the continental US. These providers peer directly with other large providers at most or all of the public interconnect points. These providers receive addresses directly from the NIC. Smaller providers are encouraged to obtain address allocations from the larger provider that provides transit for them to the public interconnects.

By now, most large providers have put into place internal procedures for route allocation and aggregation which allow aggregation of this scale to be accomplished and managed. Most large providers are also making some efforts to clean up prior allocations.

In conflict with the goals of aggregation is the small provider's fear that should they accept addresses within an aggregate of a larger provider, they will be asked to return the addresses should they decide to change to a different "upstream" provider. These fears are not without solid basis as with the IETF CIDRD (Classless Inter-Domain Routing Deployment) WG an internet draft referring to "address lending" gained support among many large providers and threatenned to become endorsed by the IETF as a BCP (Best Current Practices document). Some large providers still require addresses to be returned immediately upon terminating service or allow meager renumbering grace periods. Small providers interested in obtaining service from such a provider are put in a position of sacrificing their own interest in the interest of more perfect aggregation and the general good of the Internet.

In operational fora such as NANOG (North American Network Operations Group) and in various IETF WGs large providers are now being urged to offer very generous renumbering grace periods, extending such grace periods for renumbering hardships such deployment of poorly designed application software which is closely tied to numeric addresses and difficult to change, such as certain license server products.

On another front, a few providers still express interest in very large scale aggregation in which aggregates cover more than one large provider within a geographic region which can be as large as an entire continent. Aggregation of this scale, referred to as large strata aggregation, poses additional difficulties due to the need for coordination among providers and is particularly difficult if one or more providers in the geographic region is uncooperative.

The end result of all of this is that aggregation is highly imperfect and a consequence is that the size of global routing tables are large and there is constant change to the routing information. A prior issue which became critical in early 1994 had been the sheer size of the routing table, prompting the initial global deployment of the BGP4 routinng protocol and the use the use of CIDR. CIDR has clearly reduced growth in routing table size and rate of change. Routers today offer more memory, making the size far less of an immediate threat than in 1994. The amount of route change is causing high processor loads and interactions with certain cache architectures and small design weaknesses in deployed routers. While such products are imperfect, they represent the state of the art in available commercial routers.

These problems and potential solutions are discussed rather openly in certain fora. The vendor community should be commended for their openness in this regard. Over the past year studies examining routing protocol packet exchanges in detail have been openly discussed. In these studies it is clear that more information is being exchanged between routers than need be. In one case a design choice of a particular vendor, Cisco, was to announce withdrawl of any route to all peers without checking to see if the route had been previously announced to specific peers. This is a low overhead operation. The recipients are required to do some minimal checking and then discard the withdrawl request. This practice was determined to be benign but suboptimal and the vendor has agreed to change it. There was some initial thought, perhaps a year ago, that the effects may have been more serious than we now understand it to be. Though this change has yet to be deployed, further examination of the data indicates that additional, relatively suboptimal behaviors are likely to exist. While these have yet to be fully understood, it is believed these too are relatively benign.

During the NSFNET period we observed some of the same behavior that Merit has seen and discussed this nearly a year ago. Merit, with the help of others has explained at least one of the technical problems leading to the observed somewhat excessive localized exchange in routing information sufficiently that it can be corrected. We are hoping to gain an understanding of whatever problems remain so that they too can be corrected.

# Tony Bates Explains the CIDR Report

## Only about 25% of Advertised Routes Are CIDR Aggregates -- Report Helps Identify Leakage -- Dampening Seen as Way to Network Stability

**Editor's Note**: in this interview (December 10 at the San Jose IETF) Tony Bates, Consulting Engineer, Cisco describes the development and purpose of his CIDR report. This is a report that was recently resurrected after Internet engineers noted an unexpected bulge in the Internet's routing tables last summer. The following discussion explains the role that CIDR currently plays in keeping the net running. It also describes what is one of a number of tools likely to be increasingly used by those involved in developing a quality of service based Internet for use in mission critical industrial applications.

***COOK Report***: We have followed the CIDR issue closely and written a good deal about it. We have wanted to talk with you however because, after a year of stability, the number of announced routes has grown in 90 days by about 25%. We remember Paul Vixie saying that most of the growth had been in Class Cs which, in theory, shouldn't be announced at all. In short we don't understand the reasons for the growth. Faced with this growth, we note that you have restarted your weekly list of the worst offenders. We hope you can help us understand what is happening.

**Bates**: OK. Let's say that as I got back into this and looked at the new routes, I saw that indeed there was not much aggregation going on. This surprised me.

***COOK Report***: Does this have anything to do with what is called the "swamp"?

## A Need to Educate the Smaller Providers

**Bates**: I think the "swamp" is a bit of a red herring. I don't myself have a clear idea of where all the new routes came from. I believe that it is still a matter of educating the providers. I think the slow down in routing growth may have come mostly from route aggregation under CIDR on the part of the larger providers. Many of the smaller ones may well be continuing to advertise class Cs and those advertisements seem to gravitate up the food chain toward the defaultless backbones. For example if you look at 206 as a representative block, you will find a lot more 24/ prefixes being announced in the system than would be expected. The bigger providers generally are doing a good job of aggregation. However multi homing severely increases the total routes advertised.

Over all we have seen a large number of small advertisements being added to the routing tables in the last few months. I think this may be do to lack of education about how to do aggregation on the part of the smaller ISPs. Also there is a large amount of multi-homing going on. What happens with the tables and multi-homing really depends on where you get your address blocs from. Now big companies like Apple and Cisco have a lot of legacy address space. Therefore they function quite independently of their provider's space.

Now suppose you have a /22 prefix, aggregated out of your upstream's prefix 16, and you connect to someone else. Depending on the policy of your new upstream providers /16 prefix, they either will or will not accept routes that use the IP addresses from your /22 prefix.. Now using something called longest match routing, you could force the routing of the 22 to take a perhaps round about path, if your new upstream provider wasn't willing to route it directly.

Let's say that you connect to two providers. Sprint and UUNET as just a theortical example. You may decide to use Sprint as a primary connection and UUNET as backup. But the problem here is that unless you have a failure, there is no traffic on the UUNET line. So you ask what you can do with routing to even the load? The first thing is to send traffic to all customers from UUNET because they are easy to handle. So this works fine. But when you send to UUNET, what happens to the traffic that they want to send back? To do that, you have to send your routing information to them. Now let's say you got your numbers out of the Sprint block. In order to send those numbers to UUNET, you have to punch a hole into the global routing tables, because, although Sprint likely aggregates your / 22 prefix and does not announce it globally and separately, UUNET, in order to route its traffic to you, would have to do just this.

Now from a CIDR perspective, when you are multiply homed, you may sometimes have to actually inject fresh routes into the global routing system in order to get the load balancing you need to make most effective use of all your upstream providers. To create optimal load sharing, you wind up creating sub optimality in the routing table size.

***COOK Report***: Does anyone have any reliable data on how many people are multi-homed?

**Bates**: That's almost impossible to do. I'll explain why. My weekly report is a snapshot taken from one place in the internet. It currently runs off the Xaranet router at MAE East. Now this is a fairly well connected router at the biggest exchange point in the world. But all this gives me is probably Xaranet's view of the world. Now with BGP all you can do externally is to advertise your best route. You don't advertise secondary routes. So you lose all relative path information with respect to the source who is advertising to the receiver. Now I could do an analysis that would give me some indication of the presence of multi-homing in an autonomous system (AS), if I were inside the system. Not only are many ISPs multi-homed but many companies like Cisco also have connections to multiple providers. In addition to protecting them from service interruptions, multi-homing will give them routing flexibility.

But, whether or not I would see two routes from a provider that was multiply homed, would depend on whether the providers to which the multiply homed ISP was attached, each had separate routes to me. If the multiple homing took place at one level further down in the hierarchy, then I would likely have only a single route advertised by the time it got to my vantage point at MAE East. If I get two different paths to the same provider, then I know that provider is multi homed. Just beware that, depending on where in the internet topology the multi-homing takes place, you may or may not see multiple paths to the multiply homed entity.

Now there are techniques that you can use to lessen the impact of multi homing. You might advertise the routes to the second provider only when you were actually using it for backup. Another possibility is to advertise the routes all the time but to use Network Address Translation to make them fit into the new IP space. If you provide some hooks, network address translation could be fully coupled with routing. This however is something that Cisco has not yet implemented in its routing code. This idea was only presented for the first time by Yakov Rhekter at the October 1996 meeting of NANOG.

## Limitations of the Current System

***COOK Report***: Help us to understand what we had heard of as a limit of about 60,000 routes that could be handled by the hardware of the Cisco 7500 series routers. Has this changed?

**Bates**: I think this is a broken issue to be quite honest. It is a combination of the number of routes and number of paths that really starts to hurt.

***COOK Report***: OK. This is new territory. Help us to understand the difference between a route and a path.

**Bates**: The route is where I want to go, the path is how do I get there? X ASabc ASdef is a BGP route. "ASabc and ASdef" is the path attribute for the route "X". The route "X" originates in Asabc and goes through Asdef to get to "X". Now you may have 40,000 routes. But look at an average MAE East player who is connected to four big guys. For every route destination "X," you could have as many as four paths that you might use in order to get there.

Now when the routing table explosion began in 1993 we were talking about routers with 16 megs of memory. And CIDR implementation that began then did exactly what it was supposed to do. Namely slow down table growth. But the numbers of routes advertised have also been affected by other variables that range from the ability to learn how to do classless addressing and routing to the growth in multi-homing. To ask what total number of routes a Cisco can hold is asking the wrong question. Currently our 7500 series can use 128 megs of memory and, given the total number of routes out there now, we are nowhere near bumping into that as any kind of an upper limit. You can also use smaller boxes with less memory to route but, with these, you would be less able to optimize the routing involved in order to achieve your own policy goals.

## Who Pays the Penalty for Table Growth?

***COOK Report***: It sounds then like a situation where increases in the routing table size hurt the smaller players the most by either forcing them to upgrade their routers more frequently and thereby spend more money or by forcing to route less optimally if they choose not to upgrade. [**Editor**: as the interview with Noel Chiappa (that we also publish in this issue) shows, increasing the number of routers advertised at the Internet's defaultless core, has the negative effect of diminishing the operational stability of the Internet.]

**Bates**: That is correct and therefore we do what we can to hold hardware requirements down. But also, because it is sound engineering to care, the big guys don't want to see the totals of routes advertised grow without limits.

***COOK Report***: Because the tidier things are kept at the periphery the less they have to worry about over loading their own backbones?

**Bates**: Yes. If you want to control it from your perspective, while doing as little work as possible, one of the things that you can do is proxy aggregate. If your downstream customers dont aggregate well, you can fix it on the fly for them. If they announce four class Cs (/24 prefixes) which should be obviously made into an aggregate, you can do that for them.

This might be a good segway into what the CIDR report is about. When I worked at RIPE about four years ago CIDR was just emerging. Before CIDR everything was classful in the sense that you were talking about routing for entire class A, B or C networks. What would happen back in 91 or 92 is that we would get say eight class "C's that would be allocated without any nice class boundaries that would facilitate the forming of a CIDR route. As an examplke, you would likely get them contiguous, but it would likely be 204.70.1 to .9 which turns out not to be very useful - .0 to .7 would have been much easier to CIDRize.

***COOK Report***: In the sense of being in the same octet?

**Bates**: Right. Although it is not that they are in the same octect so much as on a Cidr alligned boundary. Now I looked at this and wanted to see what I could do with it. So the CIDR report worked from a copy of the full routing table. I said: nothing too fancy, nothing too complicated. I will just look at an AS and at what routes it originates. I will infer nothing about routing policy. I will just look at the list of routes that each AS propagates and then see how I could issue fewer routes by doing as much basic aggragation (asuming no holes in the aggregates) with the routes thatwere announced. When I did this I found that I could derive significant savings. Basically what was happening was that people just couldn't be bothered to form the aggregates they should have been forming. There were too many other things demanding their attention.

Now I kept sending this out week after week only to find that it really didn't make much difference. So I decided to highlight the top ten - those who, if they aggregated their routes, would make the most difference. Now I only look at classful nets. I look at what would happen if you if you aggregated these nets using CIDR. And I derive, from the comparison, the unnecessary gain in routing announcements.

Now despite the fact of nearly four years of CIDRization, of the 40,000 routes being announced, nearly 30,000 are classful - that is to say they are in the old classful style of class As, Bs and Cs. And as Paul Vixie pointed out most are quite recent. At one level CIDR is being enforced as far as getting address blocks from the upstream providers. But, at another level, the way that the down stream customers are announcing their addresses, shows more work is needed in aggregating announcements. Now an ISP might give a customer a class C, but to see this ISP advertising individual class Cs all the way up the food chain makes no sense. What I have just described is known as the AS aggregation part of the report.

Let me give you some history. I started this report when I was working at the RIPE NCC. I did it while CIDRD was in its heyday. It worked in that some people were embarrassed to appear on this list and did something about it while others did indeed ignore it.

***COOK Report***: Does your report segregate old from new allocations?

**Bates**: No. That would really be difficult to do.

## The Swamp and Route Leakage

***COOK Report***: Tell us about the Swamp.

**Bates**: The way to think of the Swamp is just like you would think of any legacy. There is really nothing special about the swamp as to its routability or lack thereof. The Swamp is a bunch of classful addresses which could be aggregated. The CIDR allocations really began with the 193 bloc and 192 therefore was the stuff that had been handed out in the many years before CIDR as the net really began to grow. I think people with numbers in the swamp thought they either shouldn't or didn't need to do any aggregation. But that is not really true. You can do effective aggregation almost anywhere. Now we should be careful because a lot of people have aggregated and a lot have even returned addresses including class Bs. There are a lot of good players out there but there is also a lot more work that could definitely be done. Also pre the 192 space we didn't even have the regional registries.

***COOK Report***: Sometimes your reports mention a leakage of several hundred or even a thousand routes. What causes this?

**Bates**: I think in general it is caused by operator error. With a snapshot taken once a day, you will see with 41,000 routes a delta there of about a thousand routes. It turns out that even though the entire Internet system is very dynamic, a thousand routes is still quite a large change. You could understand five hundred guys adding two routes each to the system. But in reality it is more like two guys adding five hundred routes each to the system. Since no reasonable increase in usage could account for such sudden and concentrated growth, I assume that there must have been route leakage.

So the report starts with a history piece. Then follows the AS aggregation part. Then there is the Delta piece. The Delta follows 7 days behind and looks at the routing table then and the one I have now. I am still making this big distinction between classful and classless routes. When three quarters of the network are still composed of classful fashioned routes I figure I better do something to call attention to things. The whole point of the report was to get people thinking about what they can do with classless routing.

With a snapshot taken once a day, you will see with 41,000 routes a delta there of about a thousand routes. It turns out that even though the entire Internet system is very dynamic, a thousand routes is still quite a large change. You could understand five hundred guys adding two routes each to the system. But in reality it is more like two guys adding five hundred routes each to the system. Since no reasonable increase in usage could account for such sudden and concentrated growth, I assume that there must have been route leakage.

## Classless and Classful Routes

***COOK Report***: A classless route comes by definition from a CIDR block?

**Bates**: By definition everything is a classless route. But if you start looking at classful routes in the A, B and C space, they will always begin with the /8, /16, or /24 prefix. These could be allocated as CIDR blocks but what is happening is that they are being advertised in the old classful fashion and to be quite honest, it is rather hard to see why you need to see all these class Cs advertised out of new allocations.

***COOK Report***: Why? Because their upstream providers should aggregate them into their routing advertisements?

**Bates**: Right. ISPs are not getting new IP numbers in dribbles of one /24 prefix at a time. They are getting somewhat larger blocs than that and they are supposed to do the appropriate thing with it. Clearly, because we still have so many classful old style routes, they are not always doing the appropriate thing.

***COOK Report***: So a classful route is one with a prefix of 24,16, or eight?

**Bates**: Right. Now from a CIDR point of view, it doesn't matter what the prefix is just so long as the routing table is kept small. The way you keep the table small is with aggregation.

Now there is another part of my report called "interesting aggregates." You might ask what if you went the other way and got a piece of the B space and advertised it as a /23. There is nothing wrong with that except that today we generally do not allocate out of B space less than B sized chunks if at all. Now let's look at the C space where the worst case scenario is a /32 or one machine. Now the algorithm that makes the report labels that " interesting" because it is denoted by things that are going in a direction contrary to what we might expect. I am just trying to highlight where they may have made a mistake in what they are advertising. This highlights very interesting things like host routes. And when you see a /29, this indicates another interesting form of leakage. The report is a diagnostic tool which can highlight a lot of configuration problems.

## What is Scary

But what is scary about these advertisements is that, because they are not classful, they have to be formed some how. Someone had to form a subnet from classful addresses. Having been formed they leak. The CIDR report has become very useful as a diagnostic to check for leaks. For example there was one a few weeks ago when LanLink (AS719) leaked something like 1500 classful routes in a single day. I sent them an advisory note because it certainly looked like a mistake. They said thanks very much and quickly corrected it. I think I saw one from BBN today. The scary thing is that they are way too common.

***COOK Report***: We had seen SURAnet. Should we have been thinking BBN?

**Bates**: Yes. Let me tell you more about how the report is done. Remember the information that I have is the route, the net, the prefix and the AS number to work with. Now with each AS number I use whois and the Internet Routing Registry. Using this to look up an AS number will return whatever name the AS number is currently registered under - in this case, although BBN owns SURAnet, the AS number is currently registered as SURAnet.

We are seeing too many interesting aggregates and way too many deltas. These random leakages of 1000 to 1500 routes make it hard for the guys at the edge of the network to plan their own infrastructure growth because they keep things way too much of a moving target.

One of the things that I have found since I restarted the report is that there are no trends here whatsoever. However if we were all more careful we could easily reduce the 40,000 routes by 10,000. Multi homing is not adding thousands of new routes. It is in reality responsible for only a small increment. The key problem is likely that people are so busy at fire fighting and in building these networks that they are not looking at good address allocation to start with. But one of the reasons that good address allocation is important is that, if you don't plan your network right from the very beginning, you will put yourself in a position where you cannot get maximal aggregation. In other words, when you start handing out routes on your borders to your customers, if you don't hand them out in the right chunks, you will create problems for yourself. To do this well you have to think seriously about network design and engineering. With hardware that will soon be coming into the market place, people are beginning to talk about being able to handle 250,000 routes or more.

***COOK Report***: Does doing all this properly have an effect on route dampening and the ability to handle route flaps?

**Bates**: Of course. Because, when you aggregate, you effectively withdraw potential routes. If you do your aggregation correctly and have a router flap, the chances are that you won't be able to even see it from the realm of the outside world. Another component to help in this is route flap dampening. Whilst this doesn't not help increase the amount of aggregation it will help in the amount of route flap going on in the Internet.

***COOK Report***: By dampening you mean if you have a flapping route and I don't like what you have been doing I can basically do something with my router code that says I won't recognize requests from you for a period of time.

**Bates**: Yes. Route dampening basically says that I have oscillating going on at a certain rate and that, if this goes over a certain tolerance, I will basically not listen to it anymore. With better network monitoring, and address allocation coupled with dampening we can stop the ill effects of flaps. Flapping is bad but at this point we have the tools to control it and to see that it doesn't bring down the net.

With hardware that will soon be coming into the market place, people are beginning to talk about being able to handle 250,000 routes or more.

Route dampening basically says that I have oscillating going on at a certain rate and that, if this goes over a certain tolerance, I will basically not listen to it anymore. With better network monitoring, and address allocation coupled with dampening we can stop the ill effects of flaps. Flapping is bad but at this point we have the tools to control it and to see that it doesn't bring down the net.

## Filtering at the Edges

But related to this is that one of the things that doesn't happen enough is filtering at the edges of people's networks. Let's suppose for a minute that one's customers are doing crazy things. They are given a prefix 23 but they don't understand what this means. So they route the two class Cs that make up the /23 as two /24s. They then advertise them to their upstream provider which does not do any checking so the advertisement slips through. You can argue that this is as much a provider problem as it is a customer problem.

Now one thing that some ISPs (MCI for expample) do is to use the routing registry. A routing registry is a repository of information. The Routing arbiter runs one such. MCI runs another. Ripe and ANS each run their own. CA*NET runs one. What you put into a registry is a route, the prefix length and the AS it belongs to. Now lets say that a customer registers, if I'm in a different domain I can use the registry to derive the routes I would expect to see from this customer. So that if I could write a tool that extracted the information from there based on a filter that would stop immediately the problem of that provider advertising two /24s when it should have advertised one /23. Now MCI, for example, does this. It has and uses such tools. Two or three other providers also do it. But this does not appear to be the norm currently.

But think about this for a moment. If everyone did this at their customer boundaries, on all these edge connections, we'd have a much more stabile routing system. Today, if you don't filter what comes to you from another network, anything can be injected into the routing system and will percolate through to the default free core. There is a classic example of someone at MIT who went on vacation to Florida and announced his local MIT route through his Florida dialup provider and, all of a sudden all traffic for MIT was flowing down to Florida. You're not going to prevent people from occasionally announcing routes that don't belong to them but when this does happen we should do a much better job of catching it and quickly nipping it in the bud. Why don't people use the registries and do more filtering?

The registries need to be accurate and consistent for sure. But with a situation like MCI you don't have a problem with consistency. If you are a customer of MCI, you have to use its registry if you want to be routed. The reason that I brought this up is that filtering in combination with routing registries is yet another means of promoting route aggregation, good engineering and good address allocation.

By the way it is hard to do the AS stuff. When you get an AS from a registry, you don't now have much incentive to keep the registration up to date. It's really nothing more than contact information. The registry goes a step further in giving you the ability to register policy information. But I use nothing more than the name translation. Now Now on my web page (http://www.employees.org/~tbates/cidr-report.html). I recently added a search function because people are saying to me that it is hard for them to get a good understanding of what routes they advertise to the outside world from just the single snapshot viewpoint of MAE East. So I added the ability for them to see how many routes they advertise, how many classful, how many classless, what their current aggregation values are and their interesting aggregates. I hope providers find it useful and will help make it easier to aggregate where they can.

# Jack Buchanan Offers Insights into Needs of Cost Effective Grass Roots Telecom

**Editor's Note**: For about a year we have seen some excellent material from Jack Buchanan of the University of Tennessee Memphis. Jack has some important insights into the responsible use of this technology.

On the telecom reg mail list on Feb 13 Buchanan quoted **Eric Rabe** of (Bell Atlantic): As is well known, Bell Atlantic Internet Solutions offers an ISP service, bellatlantic.net. But rather than put our own subscribers' traffic on the voice network, Bell Atlantic uses an alternative data network to handle data traffic. It is, of course, available to any ISP who wants to use it, but it requires ISPs to move away from their modem investment and instead receive traffic on digital circuits (SMDS).

**Buchanan**: I have questions about telco pricing in this environment. This is not intended to be confrontational, but rather to design an ISP like system which DOES NOT have a present heavy investment in modems. I would really like to know the technical issues involved so I can be an informed consumer in this environment.

Two case studies. **Case Study 1**: In an effort to provide dial-up access to our municipal backbone, which is connected to the Internet, for community centers, non profit agencies and a few individuals, we are considering boxes which have a single (or dual) connection to the telco network via Primary Rate ISDN or channelized T1 with connections on the other side (usually ethernet based) for our data network. These boxes include analog modems so they can very flexibly serve analog and ISDN users. A prototype would be an Ascend MAX (4000??) box. We approach both a competitive access provider who is serving us well for our backbone stuff as well as Bell South. Here is what we find out.

a.) Competitive Access Provider:
Cost for Primary Rate ISDN
(1PRI) $900-$1000/mo
Cost for "channelized T1"
$250-$350 /mo

Of course, the T1 solution gives only 56kbps channels rather than the 64k for ISDN and requires additional software for our box. When asked about the significant price differential we are told "It takes a lot more to provision a PRI than a T1".

Question: Is this true and why? I would think it would be to everyone's advantage to have this traffic on a PRI. Why the tremendous incentive NOT to go with more modern technology?

b.) Bell South as Provider:
Cost for 1 PRI $1000-!200/mo
Cost for 12 (2B+D) Basic Rate Channels ($36/mox12) $430/mo

The price break for the BRI channels is because Bell allows residential rate for educational and not-profit use of ISDN but *will not give that rate if it is aggregated into a PRI*. Why? It appears to be to no ones advantage to bring in a PRI, break it out into individual B channels with attendant punch-down blocks, etc and then buy the rather expensive and much less dense B channel interfaces into our "modem" box. Again we are forced into unreliable "old technology" (punch down blocks and copper pairs). I would also expect that the out of band signalling requirements of a PRI are less demanding than the 12 D channels for the BRI's but am out of my league on that. (i.e. the required extra demultiplexing seems to make no sense to me).

c.) Nobody has tried to sell us an SMDS solution. Why?

**Case Study 2**: Rust College, a small historically black college in north Mississippi, about 30 miles from here has an NSF "Connections" grant to get on the Internet. For programmatic and logistical reasons, access through us is obvious. They have about $30,000 from NSF to spend up front and have committed to maintain the connection after that with their own resources.

a.) Competitive Access Provider: Crosses a LATA boundary; Not interested.

b.) Bell South: No ISDN in this part of Mississippi. Propose a 56 KB frame relay connection via Tupelo (about120 miles east, i.e. the other way) then back to Memphis. Fixed costs will be $900-$1200/mo. They don't know what the usage charges are likely to be. How, with a straight face, can they ask this small poor college to pay $1000/mo plus usage charges plus Internet access fees for a *56Kbps* connection. I know a LATA boundary is involved (with Bell South as the dominant carrier on both sides) but really.

c.) Nobody mentioned SMDS here, either.

d.) We may go to a dedicated wireless link here. If we like it, may turn back other T1's and ISDN links and go back to RF school so we can maintain them.

Maybe this is all too specific. But, these are real situations and excerpts of real conversations. I submit that the ISP's aren't the only ones forcing the use of inefficient and inappropriate technology. We go (in the long run) where the Telco price practices force us. Why is it so hard to BUY state of the art for affordable prices?

**Editor**: White Jack received no answers from Bell Atlantic, on February 24 he wrote on the LII list:

Maybe a small point, but I am a little surprised at the emphasis on "isolated *rural* poor information have nots". I come from a rural background (at least if you consider a town of under 10,000 as rural). Nevertheless, the most *isolated, poor information have nots* I know about are in the *inner city*. We must not battle each other on this point. Concentration of *only* schools, libraries hospitals (with an emphasis on the rural versions of each) may be the fatal flaw of the Telecom act as it relates to LII.

Putting on another hat, the Telecom act promised Telemedicine support for *rural* hospitals. Since hospitals are very labor intensive, the cheapest hospitals to operate are in the *rural* areas where labor costs are generally lower. Why do you think HCA and the like are buying them all up? Why not equal Telemedicine access to the *inner city* which has traditionally had the least access to health care????

I can't decide whether the Act's playing populations off against each other is misguided do-goodism or deliberate fragmentation and cream-skimming in advance by the telecom industry, assisted by a coalition of librarians, teachers and rural hospital administrators.

And another small point:

If the schools and libraries are going to be the point organizations in a Local Information Infrastructure, Nevertheless, the most *isolated, poor information have nots* I know about are in the *inner city.* We must not battle each other on this point. Concentration of *only* schools, libraries hospitals (with an emphasis on the rural versions of each) may be the fatal flaw of the Telecom act as it relates to LII.

1. The schools are going to have to quit closing for half the year and at 3:30 p.m. for the half year they are open at all,

2. The security guards are going to have to let the public into the schools, including those without students there

3. Inner city branch libraries need to be opened (in my town with 21 branch public libraries, zero or one (depending on how you count) are in the inner city). More rural libraries are also needed for the same reason.

I say this with constructive intent and as a recipient (as PI) of NSF funds targeted toward education and as an active member of the Memphis Urban Systemic Initiative (NSF) Citizens Advisory Council and public schools advocate with a 6th grade son in an arguably inner city Memphis Public School. I also have talked umpteen dozen times to the local library folks who always attend meetings, appear to listen politely, and then do precisely what they want(mostly VMS and proprietary online catalog stuff), without really listening very much to what anybody else thinks (They make great Rotary Club speakers, though).

And on February 26 Jack wrote: I was undecided whether a discussion of Telemedicine was appropriate here but rural telemedicine is one of three areas targeted for legislated discounts in the Telecom Act, and an area where I, at least, have grave reservations about the approach being taken by some. I speak having been involved in PACS (Picture Archiving and Communications Systems) and telemedicine research in the 1980's in North Carolina (UNC-Chapel Hill, Bell South, Fujitsu, Microelectronics Center of North Carolina, etc) and with a present fairly heavy regional involvement in telemedicine with the US Department of Veterans Affairs, one of my employers.

First, medical EDUCATION and delivery of medical services are NOT the same thing. This is true whether the target of the education is the consumer, medical students, continuing education for nurses, etc. Even the NIH gets this wrong, essentially putting all telemedicine in the National LIBRARY of Medicine rather than in its disease specific branches or the Institute of General Medical Sciences, where it probably belongs. Since when did you see the LIBRARY take care of patients. To recap point one, medical education, consumer information about medicine and health, and practice of medicine are related to each other, but they are in no way synonomous.

Second point: Though it may suprise those who know of my vested interests in medicine, telecommunications and high performance computing, I believe the discounts for "rural telemedicine" legislated by the Telecom Act are probably not a good idea and will probably result in wastage of lots of money. Why? Because the underlying medical systems are not ready for such widespread deployment. Clinical records for the most part are still PAPER BASED. A whole mini-industry and academic discipline is working toward a Computer-Based Patient Record System (CPRS-the current buzzword). Until we have a computer based patient record, what will the telecommunications channels communicate? Do we just want high speed fax machines?? I think not.

I strongly support RESEARCH into telemedicine. In a RESEARCH environment projects which provide niche information (or partial information) like the patient's ECG, the patient's chest xray, pictures of the patient's skin lesions, etc, are important in defining what the ultimate CBPR will be like. In the CLINICAL, SERVICE DELIVERY environment, rural or urban, such isolated systems are more trouble than they are worth. Why have an "electronic ECG" if the other 20 things the practitioner needs are paper based? Systems to be deployed need to be comprehensive and ubiquitous. Standards for data interchange, common data dictionaries, etc have to be refined. Here, probably surprising to the capitalists, government institutions such as the Defense Department and the VA, since they have restricted patient populations, quantifiable service delivery locations and extreme pressures to be cheaper than the private institutions who are trying to put them out of business, have a head start. Even there though significant portions of the patient record are still paper based. To recapitulate point 2. We can deploy the "Tele" part of Telemedicine rather easily, it is the "medicine" part of Telemedicine which is not ready for prime time. I suspect the Telecom Act writers hadn't a clue about any of this.

Caveat: Much of this is obviously controversial. My points are based on the thesis that the people to give the information to are the people actually taking care of the patient (i.e. primary care physicians and "physician extenders"). Often, radiologists and other specialists will believe they need the information design systems differently and often won't understand the above points.

I did not mean to imply that there is anything wrong with health informationon the web or that there is anything wrong with an informed health consumer--just that that itself is not telemedicine. In addition, a lot of companies and academics are trying very hard to bring modern information technology to the medical environment. In the areas of inventory tracking, billing (individuals, insurance companies, and the government) a lot of strides have been made. Less has been made with regard to the actual medical record. I contend that the purpose of the whole system should not be limited to making sure Blue Cross/Blue Shield gets paid. The purpose of the system is to deliver health care services. Part of that process happens to be making sure the bills are paid and every Aspirin tablet appears on the bill, but that shouldn't be the whole process.

If you become unconscious in a strange city, or usually even in your own city the Emergency Department will know the balance on your VISA card long before it knows anything about your medical history. Your written medical record will probably be so hard to get that they won't bother. This may be getting off topic, but is part of why I think the discounts for "rural telemedicine" will not be helpful until the reengineering of the health delivery system gets farther along. Just as discussion of this subject may not belong here, attempts to reform the medical system probably didn't belong in the Telecom Act.

Jack Buchanan, MSEE, MD Associate Professor of Biomedical Engineering and Medicine University of Tennessee, Memphis; Adjunct Professor, Herff College of Engineering, University of Memphis

# Executive Summary

## DNS Under Stress, IAHC Position Weak pp. 1- 6, 24

With the completion of the International Internet Ad Hoc Committee final report calling for a Council of Registrars and the creation of a shared database system for up to 28 Registrars of Domain Names the 18 month old dispute kicked off by Network Solutions’ beginning to charge for domain names in September 1995 is coming to a head.

We present a detailed history of the solicitation leading to the InterNic award. We examine the role of NSF from the start of the award in April 1993 to the present and conclude that its actions were reasonable and proper. Using a network of sources we describe the current positions of NSI, IAHC and the Alternic “glitterati” - the last of which we conclude are not serious players. While we are not sympathetic to NSI’s plans to capitalize on their good fortune with an IPO, we find no reason to condemn their stewardship over the .com, .org., and .edu domains.

While we think the goals set by the IAHC are worthy of support, we believe that they face very difficult odds that may scuttle their effort. While the seven new top level domains that they are supporting will increase the range of choices for Internet users, they face a very difficult task.

In evaluating at IAHC's chances of success one must ask two things. Where are we to assume that it will get the considerable sum of money required to build its infrastructure and make it work in a very short period of time? Second who will register what domains and how quickly will the data from the new Registrar's IAHC sponsored seven top level domains start being added to the data bases of the root DNS servers? The IAHC side has indicated that it would like NSI to join CORE and add .com to the domains registered by the CORE Registrars. As soon as the IAHC CORE system has an operational shared database, we'd like to see NSI join and do just this. However when one looks at the stark reality that NSI would have to give up control over its pre-eminent cash cow in order to do so, we think it very unlikely that this will occur in the short term. After all the shared database system has to be built first and it is not clear where the sponsors are going to find the money necessary to do so. If they do succeed in building a working system that gains acceptance, IANA and community pressure could force some changes. The biggest immediate question is how the entire system will weather a legal challenge.

For IAHC will be lucky to escape being sued and if it is sued, there is one very weak link in the whole process: The authority chain. In other words the IANA - Jon Postel. The IANA's authority endorsing it is what gives NSI's registry value. The operators of the root DNS servers are willing to accept Postel's recommendations as authoritative. They carry the registry database (s) that Postel asks them to.

But what if a court were ever effectively to say to NSI: the authority of the IANA is in dispute and/ or no longer valid? Therefore you may no longer use the IANA as your authority in asking the root servers to carry your database. Or perhaps more likely were a court to say to the operators of the root servers: you may no longer restrain trade by accepting only data that the IANA deems authoritative?

In a worst case scenario, you then might then find multiple groups with databases of questionable quality insisting that they be added to the root servers. If the databases conflict and the system falls apart, too bad. This probably won’t happen. But it could. Being fully aware of the dangers may be the best way for all parties to avoid disaster.

## Noel Chiappa on the Scaling Problems of Current Routing Technology, pp. 7-15

In a long interview Noel analyzes the problems of the current technology with its premiums on routing gurus. The more routers there are in the defaultless core of the net the longer a flap takes to stabilize. However the growing number of routers increases the likelihood that some router somewhere will flap more and more often. This problem presents curves that, undisturbed, will intersect. Of course they cannot be allowed to do so because at such a point the net would crash. To ensure that they do not intersect, Noel believes that either aggregation must be more rigorously applied or new routing technology developed to replace BGP4. He explains why the new technology approach will be very time consuming and therefore difficult. (As we prepared to publish the discussion, it seems to us that route dampening might be a third option.)

In any case the discussion for those of us who have never had to use a router is an extremely informative guide to the ways that routing technology may be applied to today’s network.

## Curtis Villamizar on the Difficulty of Aggregation, p.16

In a response to a journalist’s query about Cisco withdrawls of routing information, Curtis explains the difficulty involved in large scale aggregation and efforts to deal most effectively with route flaps.

## Tony Bates Explains the CIDR Report pp. 17 -20

After an interruption of several months, Tony Bates restarted his weekly CIDR report last fall. The report serves as a tool that is useful in showing which ISPs are are implementing CIDR aggregation extensively and which are not. He notes that after four years of “CIDRization” of 40,000 routes announced only 10,000 are CIDR aggregates. He suggests that with reasonable effort 10,000 routes could be cut from the total announced to the defaultless core. He describes how the report serves as a useful tool to identify inadvertent leakage of routes. Such leaks are significant. When they occur, they often are in the 500 to 1500 route range. He suggests that routers will soon be capable of handling up to 250,000 routes. Implied is that for the routes of the defualtless core to continue to grow significantly, routers in the core will continue to grow in number and dampening will have to be increasingly used to avoid having the net taken down by ever more frequent router flaps.

## LEC Charging Policy Questioned pp. 21 - 22

Jack Buchanan of U. of Tenn. Memphis questions local loop pricing policies that would appear to make digital circuits that are less costly to provision and have less impact on local Central Offices than analog POTs equipment. Jack also offers useful insight into the fallacies of the universal service provisions in the 1996 telecom reform act.

**con't from p. 6**

to pay you to get in your database as well as pay to get in the ISOC/IAHC database? If you want your database to be a part of the process, then apply to IAHC to become one of the *registries* contributing to its distributed database.

As far as I can tell you are not even on the radar screen of the technical discussions going on. Have you been participating in the newdom mail list? If you have I am surprised that I have never seen anything about you from the technical spill over of that list. Again, anyone who invests in your attempts to create an alternative to the alternative that has now been officially sanctioned by the Internet community and expects the investment to *pay off* with something technically usable is being quite foolish.

Oh.....by the way..... You said the case begins this month.....what case? I read 50 or 60 technical and non technical lists..... I have never seen your name before and I have read megabytes of blather over the dns issue. Tell me if you have a court date for the next two week how is it possible that the entire on line media has missed it?

In your post in answer to the *Economist* you say: Name.Space is developing an enhancement to the current domain name software enabling a dynamically updated, shared database in which independent, competing registries can share all top level domains without conflict.

Sorry I am not impressed. You are proposing to do what the Internet community has already sanctioned IAHC to do. Please tell me how a database can have more than one set of authoritative records? Either yours or theirs. It is *not* a case where the courts can allow *everyone to play*. Lets assume you are right and to keep you happy the Internet waits until it gets an international treaty. What makes you think the treaty will adopt YOUR solution? And again with database records at the level of *root* servers there can be only one authoritative set. [Editor: end of Feb. 16 response to Paul Garrin.]

**Coming in the May *COOK Report***: A report on SprintLink An Interview with Brad Hokamp and Benham Malcom of SprintLink. (We decided to hold this interview and publish it alongside one with Hank Kilmer, Together the two are very informative.)

We will be in Russia from May 14 until June 9. Will publish June issue by May 1 or sooner.

## *1997 COOK Report* Subscription Rates

1. Individual - now in ascii form only - $100

2. Individual - hardcopy via first class US Postal Service - $175

3. Academic Libraries and Higher Education - $200

4. Small business, small government agency (hardcopy) - $200

5. Large government agency or business over $10 million a year gross revenues (hardcopy) - $350

6. Corporate or government agency site license (same as 5) except subscribing business receives desktop published hardcopy and electronic copy and permission to redistribute either or both to as many of their employees as they will. We will email directly to an adressee of your choosing or an internal mail reflector. We also encourage you to reproduce the hardcopy and route internally to as many people as you wish. - $650

7. Deluxe Corp Site License - same as 6 with added right to place electronic text on employee only corporate wide web server - $1,000. see **http://pobox.com/cook/**

**Internet: cook@cookreport.com**

Gordon Cook, President
COOK Network Consultants
431 Greenway Ave
Ewing, NJ 08618, USA
Telephone & fax (609) 882-2572

The COOK Report on Internet
COOK Network Consultants
431 Greenway Ave.
Ewing, NJ 08618