The future of ARQ in congested HF bands

by Peter Martinez G3PLX
High Blakebank Farm, Underbarrow, Kendal, Cumbria, LA8 8HP, United Kingdom.
Email: Peter.Martinez@btinternet.com
Phone: +44 1539 568740

for ARRL/TAPR Digital Communications Conference September 2003

Summary
It is almost exactly 25 years to the day since the writer conducted the first experiments in the use of Automatic Request (ARQ) techniques, introducing for the first time the concept of error-free delivery of messages over the error-prone HF amateur bands, following this with the first experiments with automated store-and-forward techniques using ARQ links. This paper looks at the way in which ARQ techniques have been used within amateur radio since that time and attempts to extrapolate into the future. The conclusions reached are that current ARQ techniques have major flaws within the context of today's amateur radio community, and will have to undergo changes if the demand for error-free message-handling is to continue as part of amateur radio into the next 25 years.

In the beginning...
Error-correction did not start with AMTOR in 1978, it had been around since the dawn of radio, but was never part of the radio system - if you didn't copy what the other operator sent, you simply asked him to repeat it and thought no more about it. The difference with the introduction of AMTOR was to automate this process and remove the human element from the error-correction loop. In those days the volume of traffic to be transmitted was much smaller than the channel capacity, so the fact that it took longer to transmit a message when there were repeats was not a limiting factor and never worried us. Another thing that never worried us in those days was the fact that some of the repeats were needed to overcome co-channel interference. Since this interference was light and sporadic, it could be treated in the same way as fading and noise and merely slowed-down the transmission to a tolerable extent. There was no thought in those days of what problems might result from the use of ARQ in overcrowded bands.

The effect of congestion
Before considering the effects of congestion on ARQ activity, it's useful to study how congestion affects other kinds of activity, and in particular it's useful to see how the build-up of congestion on today's HF bands has influenced the way that radio-amateurs operate on the air. We can use a bit of psychology to study how radio operators have - mostly without even thinking about it - adapted their operating patterns to mitigate the effects of congestion. I would like to suggest that the most significant technique used by amateur radio operators to counter the effects of congestion on the bands is simply to avoid going on the air when it's busy. This might seem a trivial self-evident truth, but it *is* a vital factor. Amateurs go on the air for enjoyment, unlike commercial users who have schedules to keep and traffic to pass, the volumes of which are dictated by outside factors. Only amateurs have the option of choosing not to go on the air if it would be unpleasant because the band is too busy, and this provides a beautifully simple feedback mechanism to control the level of congestion. I would like to suggest that this one simple cause-effect link is the most important single factor which makes life on today's amateur bands tolerable, and does so without the need for any rules and regulations. Unless you are in the CW contest, you stay away from the bottom end of 20m on contest weekends. In Europe at least, you keep away from 40m SSB on a Sunday morning unless its to call into a net. The essential point is that although we may not think of it as such, the innate predisposition of ordinary amateur radio

operators to keep away from congested bands, is a very powerful control mechanism. Without this mechanism, we would either need to impose very strict rules to limit activity at busy times, or there would be chaos.

Before we get back to talking about ARQ itself, lets think about the kind of activity for which ARQ is the main application, namely that of passing third-party messages. Here I am defining this term as being the passing of messages which are not between the two amateurs sitting in front of their radios. How does this kind of activity differ from "conventional" amateur activity, and how does it respond to congestion? The first thing that we can say is that the operators of such message-handling systems do not "go on the air" for recreation, but to clear the traffic. This removes the inherent control mechanism that I have shown is an essential feature of conventional one-to-one amateur activity. If the message-handler does *not* choose to stay off the air when it's busy, but needs to go on the air when he has traffic and stay on the air until he has cleared that traffic, there will be a greater tendency for bottle-necks to occur than if the activity is only by ordinary folk who go on the bands for fun. In the extreme case of a scenario where *all* the stations on the band were passing messages which had originated elsewhere, there is a real risk that the band could be saturated at busy times unless some external controls were applied.

The grid-lock problem

It gets worse. Not only does the nature of message-handling itself tend to make congestion-control more difficult than in a recreational activity environment, but the use of ARQ for message handling makes it worse still. After putting forward the hypothesis that congestion is made tolerable in a recreational scenario by the predisposition of live operators to avoid it, we can generalise and say that we should be able to control congestion in other kinds of activity if we can introduce a similar mechanism, namely one in which the presence of congestion triggers by some means a reduction in the amount of activity by each participant. If there is a shortage of a resource (the time-bandwidth product in this case), we need to ration it in order that everyone gets a fair share. This sounds like another trivial self-evident truth, but it takes on a frightening significance when we realise that not only does the message-handling environment provide *no* inherent mechanism to limit congestion, but the ARQ process actually behaves to make any congestion it encounters *worse*. This arises from the fact that co-channel interference on an ARQ link (resulting from congestion), causes repeats in the traffic stream, resulting in the message taking longer to complete and thus *increasing* the usage of the scarce time-bandwidth resource rather than reducing it. This reasoning shows that the widespread use of ARQ on the amateur bands could not only make congestion-control very difficult, but could actually make it unstable and prone to grid-lock.

It may seem inconceivable that we could have the same kind of gridlock on the amateur bands that we see on congested urban roads. This frightening possibility will indeed never occur if there remains a significant amount of conventional recreational activity. If there are enough live operators who can choose (or be persuaded) to go off the air if the bands are busy, then the peaks and troughs in traffic volume among ARQ-based message-handling stations can be accommodated without grid-lock. But if ARQ systems become the dominant activity in any congested finite-width band, there will be a risk of a complete lock-up or at least a considerable reduction in the traffic capacity of the band in question, in the same way that we see on congested urban freeways.

I have described the "run-away" behaviour of the congestion control mechanism for the case of ARQ links suffering interference from other users in a general context, but the problem is particularly severe for the case of ARQ links suffering interference from each other. Two such links attempting to use the same channel will, if nothing is done externally to prevent it, request repeats for ever, thus reducing the total channel capacity to zero until one link aborts. It's possible to imagine a sub-band entirely populated with ARQ links which, in the presence of congestion during peak traffic hours, becomes completely blocked with no traffic passed at all.

Again, this might seem an unlikely scenario, and has probably never actually happened yet, but it is worth looking at how the amateur ARQ message-handling networks have avoided this problem so far. It's a feature of the way these networks have been configured as a star-shape, with a central node (a "mailbox"), allocated a unique frequency or frequencies, each one servicing a set of users. All contacts are constrained to be between the mailbox and the users on one of the mailbox's dedicated channels, never between users on non-mailbox channels. This neatly avoids the possibility of collisions between links. By spreading the mailbox channels out over the band so that no two mailboxes compete with each other for a channel, the potential lock-up behaviour of the network is avoided.

The way forward

Before suggesting some possible solutions to the problems of run-away congestion, it's worth noting that the problem described is already beginning to show itself, manifested as an on-going conflict between live-contact digital-mode operators and the operators and users of mailboxes using the Pactor ARQ system. If the mailboxes migrate from narrow-band ARQ techniques to the recently-introduced voice-band ARQ systems such as Pactor-3, this conflict will probably migrate into one between mailbox users and SSB operators. In the light of the above discussion, it's clear that some of the solutions already proposed for resolving this conflict are unlikely to succeed. For example, the suggestion that mailbox operators should be confined to modest-sized sub-bands from which other band users will be excluded, is unlikely to be acceptable to the mailbox fraternity, because such a configuration would be unstable in the presence of traffic congestion. The above reasoning shows that mailboxes would achieve their objectives best by light usage of a large number of channels than by intensive use of the minimum necessary number. The problem is that in a congested, shared band, with no externally-applied regulatory structure, we cannot justify, nor do we have the right, to leave *any* channel idle for any reason. This unfortunate "sparse" network structure thus forces channels to be shared between network and other activity. This "sharing" is perceived by the live-contact operators as unfair, simply because the ARQ links are not predisposed to reduce congestion conflict in the way that human operators do. I am not suggesting for one moment that the mailbox community is deliberately inflicting this scenario onto the live-contact community. I am suggesting instead that it's a fundamental flaw in the ARQ process itself that is leading the mailbox operators to configure their networks so that their activity on each channel is light. The flaw is effectively that the ARQ protocol cannot cope with co-channel interference *from another ARQ link.* ARQ links cannot tolerate being crowded together but they don't mind competition from other non-ARQ activity. It's not surprising that the non-ARQ fans see this as unfair.

Quite apart from this argument against dedicated compulsory sub-bands for mailboxes, it's very difficult to see how *any* kind of regulatory framework to control the conflict could succeed in the present climate, since the licensing authorities are reluctant to impose or enforce such rules. Voluntary agreements can only work when both sides benefit, not when one side loses.

So what *is* the way forward. It's not my intention in this paper to propose a solution. I am simply drawing attention to the problem in order to promote debate, but I can set the discussion rolling with a couple of ideas, based on the reasoning I have set out in this paper. I have shown how traditional amateur radio activity has to a great extent been self-regulating in the presence of congested bands, because of the innate behaviour of the operators, specifically in staying off the air when it's busy, but in general by adapting their behaviour to avoid congestion, and I have shown that the large-scale use of ARQ message-handling networks could have disastrous consequences because they either do not behave in this way, or in the specific case of ARQ links, they actually make congestion worse when they experience it. I would like to suggest that if we could somehow design such networks to behave in the same way as human users, then we could restore the balance. We need to somehow come up with an ARQ protocol that will reduce the amount of time or bandwidth that they consume when they experience congestion.

I will now stick my neck out and suggest out that there *is* an existing protocol which has the required properties. I am thinking of AX25. I know it is universally condemned on HF bands because of it's abysmal performance, but this is entirely the result of the inappropriate choice of 300 baud FSK as the standard modulation system. The message-handling protocol, and the behaviour under congestion, are ideally suited to today's congested HF bands. In the presence of interference it "backs-off", leaving the channel unoccupied in such a way as to permit other traffic to clear before resuming, thus reducing it's own consumption of the shared time-bandwidth resource in an fair way. A network of such AX25 mailboxes on HF could survive being constrained to operate within a set of channels configured to take the total planned traffic load without suffering the gridlock effect inherent in a fully-loaded set of Pactor channels. There should be no difficulty in developing modern modulation methods to carry the AX25 data - the frequency-division multiplexed PSK modulation techniques used by Pactor itself are ideal. I mention AX25 only as one possible example of a solution, but the essence of this approach is that we *must* design the protocols for the links in a message-handling network in such a way as to encompass the behaviour we require of the network as a whole, within the context of shared spectrum, and not as at present where each network link regards all co-channel activity (including other ARQ links) as unwanted noise which it attempts to overcome by force of repetition. The ideal network should conform to the rules by which the whole of amateur radio operates if it is to have any chance of wide acceptance. We need ARQ links which not only listen on a channel before they seize it, but will recognise that "noise" which is unwanted by this link could well be "traffic" being passed by another link (whether ARQ or not) and which has an equal right to the channel. We need ARQ with a social conscience.

There is another more philosophical question which could be discussed, although perhaps not in the context of this conference, since it affects amateur radio as a whole. Given that recreational live-contact activity between human operators *does* have the required self-limiting property in the face of congestion, and given that there are other kinds of activity which do not have this property, should we consider carefully whether such activities should be permitted within amateur radio at all? If we have no wish or no power to prohibit such activities, how do we prevent our bands from overcrowding if such activities become too popular?

Conclusion

ARQ-based message-handling links were fine in the old days when there were only a few of them. I have shown in this paper that with the higher levels of spectrum congestion in today's bands, there are problems. ARQ links respond adversely to congestion in ways which are not acceptable to the amateur radio community as a whole, and which prevent networks of such links from being configured efficiently. We need to re-invent ARQ protocols to take account of these problems, which are unique to amateur radio because we have no way to remove congestion by the imposition of rules. We need to devise ARQ protocols which intrinsically respond to congestion in ways which are socially acceptable and which are efficient in their use of time and spectrum. We could take a lead from the structure of the AX25 protocol, and even from study of the way human operators have unconsciously adapted their operating techniques to mitigate the effects of crowded bands.