Figures of Merit for the Capacity of Wireless Messaging Networks

Allan Angus, PE
WebLink Wireless, Inc.
3333 Lee Parkway, Suite 100
Dallas TX 75219

***Abstract*-We introduce a figure of merit for the capacity of wireless data messaging networks. Using this metric, we compare and contrast GSM SMS, GPRS, and ReFLEX v2.7 networks.**

I. Introduction

In the following, we compare and contrast the air interface and economic efficiencies of message delivery over Phase 2+ GSM networks, over evolving GPRS networks, and over ReFLEX/WMtp networks. The analysis is a jumping off point for some comments about potential future evolutions of both Narrowband and Broadband PCS messaging.

II. Messaging

Messaging is a unique, and yet broad, service category. It encompasses the notions both of short content and of short delivery time. In the context of alphanumeric messaging, the category includes traditional paging ("1way"), Short Message Service (SMS), mobile peer-to-peer messaging ("2way"), Instant Messaging (IM), and some uses of email. In a broader context of binary data transport, the category also encompasses a variety of telemetry and telematics applications, process monitoring and alerting, and process control and inter-operation. As well, "analog" messaging, typically for the transport of short voice content is yet another element of the category.

Wireless messaging is an application that appears to have established itself. In Europe, there are about 10 billion messages sent per month on the GSM systems there, and about 10% of operator revenue is from messaging. In the Philippines, about 70% of all cell phone usage is from messaging. In the US, more "2way" devices of the T900 and Blackberry class were added last year (2000) than in all previous history. The 1.5 million ReFLEX units in service in the US (as at June, 2001) send and receive on the order of 300 million messages per month.

By contrast, messaging is neither real-time nor stream-based. That is, short and variable delays in transport latency can be perfectly acceptable. As well, message content does not accumulate as some strict function of connection time. As well, the expected accuracy for message transport approaches a "zero defect" limit; i.e., a non-measurable bit error rate.

In short, wireless messaging is not wireless telephony, as a more complete expansion of its Quality of Service (QoS) requirements would show. Rather than follow that avenue of exploration, this paper aims at an appropriate figure of merit for the capacity of wireless messaging networks, given the distinct QoS attributes of messaging.

III. Wireless Network Capacity

A. *Figure of Merit with Erlang B telephony load*

One simple metric for the capacity of a telephony-based cellular system is given in (1),

$$L = W/(N \cdot B), \tag{1}$$

where L is the number of trunks per cell, W is the total available bandwidth in Hertz, N is the cellular reuse factor, and B is the bandwidth consumed per trunk. Since total available bandwidth tends to be constrained, the product $N \cdot B$ has become a suitable figure of merit for the capacity of cellular networks. Naturally, it is desired to have $N \cdot B$ as small as possible to maximize network capacity. The "text-book" capacity analysis proceeds by way of an assumption of Erlang-B statistics [1] for offered load, and generates a measure of Erlangs per cell at some acceptable blocking probability.

B. *Figure of Merit with queued data load*

As a figure of merit, $N \cdot B$ is unsuitable for data-oriented networks, which do not support traffic that accords with the usual Erlang-B call model. Data models usually envolve some form of queuing statistics, which are often described in the form *AD/SD/#s/#w/P*, where *AD* is the arrival distribution, *SD* is the server distribution, *#s* is the number of servers, *#w* is the number waiting, and *P* is the population. The Erlang-B model is consistent with *M/G/s/s* queuing, that is Markov (or Poisson) arrivals, Gaussian servers, and the number of servers equal to the number waiting, so that blocked calls are dropped. Erlang-C (blocked calls held) is consistent with *M/G/s/k* queuing, given that $k>s$.

A more natural set of queuing models for data are based on *M/M/1* paradigm, in which traffic arrivals and server delays are both Markov, there is one server only, infinite waiting (messages are never dropped), and the population is infinite. In this case, (1) might easily be modified to read

$$C = W/(N \cdot B \cdot T), \tag{2}$$

where C is the capacity in bit/s/cell, and T is the observation time, and where a reference message of some specified length is being sent.

Given that the value of C is in excess of the mean value offered load per cell, and neglecting for a moment any transmission errors, the expectation might be that the network could operate with no loss of data. Given that the offered load is a Markov random process, this simple assumption would not be strictly true for two inter-related reasons.

First, any given observation interval might see an offered load in excess of the buffering capability of the queue managed by the cell server. Given that the offered load is Markovian with some specified statistics, the likelihood of such an event can be defined, and a corresponding grade of service given in a manner similar to, say, an Erlang-B call blocking probability.

Second, because the present state of any queue is the net summation of inflow and outflow processes, and because in the *M/M/1* model as defined these processes tend to have Gaussian second order statistics, queue depths tend to execute a "random walk" [6]. That is, queue depth will "diffuse" away from its initial value at a rate proportional to the square root of time, assuming that the mean difference in inflow and outflow is zero. This implies a requirement, from time to time, to accelerate the outflow from the queue in order to avoid lost data due to overflow as queue depth diffuses toward some "high water mark." If this accelerated flow cannot actually be delivered through the queue server, then data is lost.

C. *Queuing with Paretian offered load*

Frankly, a more realistic set of data models might be referred to as *F/M/1* or *P/M/1* for either a fractal or power-tail distribution of offered load ([2], [3], [4], [5]). Unlike the *M/M/1* case, the *F* or *P* load distributions do not have variances that converge. In these models, the diffusion or random walk of accumulated flow is more highly correlated than for a Markov process. As a consequence, it is not feasible to engineer buffers of finite depth that assure any finite probability of buffer over-load. A necessary implication of the *M/M/1*, *F/M/1*, and *P/M/1* models is that accumulated data arrivals will scale as some function of time. In the case of the *M/M/1* model, the diffusion of the queue depth varies as the square root of time.

Besides considering the scaling of peak to mean ratios as the amount of data is increased, another way of looking at the difficulties in managing real data traffic is in terms of the ratio of the range, R, to the sample standard deviation, S, of the traffic where

$$R(n) = [max(0, W_1, W_2, ..., W_n)] - [min(0, W_1, W_2, ..., W_n)] \quad (3)$$

and where

$$W_k = (X_1 + X_2 + ... + X_k) - k(X_1 + X_2 + ... X_n)/n. \quad (4)$$

This statistic, $R(n)/S(n)$, has been called the "rescaled range" following the usage of the civil engineer, Hurst, who first employed it in his study of the Nile River flows during the design of the Aswan High Dam [10], [11], [12]. Hurst's problem was to estimate how high the dam had to be in order to avoid an over-flow and down-stream flood, given the fluctuations in incoming river's flow. His problem is structurally identical to the issue of estimating how much bandwidth must be available, or how deep a buffer must be, in order to avoid losing data in a channel; hence the interest in this statistic among communications engineers.

It is straightforward to show that the rescaled range follows a power law as n becomes large; namely,

$$E[R(n)/S(n)] \rightarrow const \bullet n^H \text{ as } n \rightarrow \infty, \quad (5)$$

where H, is called the Hurst exponent, $0 \leq H \leq 1$. For Gaussian and Poisson statistics, $H = 0.5$. For a broad variety of "real world" phenomena, including most data flows, $H \approx 0.7$.

Generally, the behavior of a flow for which H is greater than 0.5 shows a strong correlation. In other words, "it never rains but it pours data." And these "data floods" and "data droughts" occur at all time scales, both short and long. That is to say, a large burst of data will occur at random "encapsulated" within a period of low traffic flow, and a period of low data will occur within a burst. The last statement is true independent of the period of observation. And a natural corollary of the Hurst exponent being larger than 0.5 is that the observed extremes in requirements for data buffering or transport are greater than what would be expected for data flows that followed Gaussian or Poisson statistics.

It may seem counter-intuitive to describe a set of conditions under which the average rate of change of a random variable is zero, the statistics of the rate of change of the variable are second order Gaussian, and yet the expected state of the variable involves a diffusion away from its starting point. Yet, this is precisely the same picture as given for the Brownian thermal motion, or random walk, of the molecules of a diffusing fluid. Given that there is no statistical correlation in the random variable driving the process, one expects the rate of diffusion to be proportional to the square root of time, as given by Einstein's diffusion equation. If the increments of the random process are perfectly correlated, then the Hurst exponent goes to 1, and the diffusion is directly proportional to time. In contrast, if the increments are perfectly anti-correlated, then the Hurst exponent is 0, and there is no diffusion in the accumulated random process. The naive assumption seems to be consistent with this last case; that is, that accumulated mean zero random processes would show no diffusion.

A related problem is the "gambler's ruin" which involves the probability of the complete loss of a gambler's capital playing in a game of chance against a bank. Let the bet be one dollar on each play, let the probability of a win be p and of a loss be $q = (1 - p)$. If the gambler's capital is K, and that of the bank is B, and given $0 < K < B$, then Schroeder [7] shows that the probability of ultimate ruin, q_K, is

$$q_K = [(q/p)^B - (q/p)^K]/[(q/p)^B - 1], \quad p \neq q. \quad (6)$$

In a fair game, $p = q = 0.5$, and (6) becomes

$$q_K = 1 - K/B, \quad p=q. \tag{7}$$

In the case of the gambler, ruin is equivalent to emptying his finite "buffer." In the case of the data service provider, ruin is equivalent to overflowing his finite buffer, and losing subscriber data. In the case of the civil engineer's dam, either overflow or underflow is ruin. The symmetry of the first two cases should be clear, first by trading increments of money for increments of data, and second, by running the game backwards in time; that is, by starting with the "empty buffer" and filling it from the bank to a level, K. By assumption, the equivalent of the bank, B, in the case of the service provider, is the indefinitely large volume of data that can be offered by subscribers. And K is his finite network resources. Clearly, the probability of ultimate ruin becomes certain as $B \to \infty$ for any finite K.

Schroeder shows [8] that the scaling factor of capital goes as the square root of any scaling of time, in the execution of a fair game, where $p=q$, just as for Brownian motion. Likewise, the cumulative probability distribution of the ruin-free distances, $P(z) = \Pr(Z \geq z)$ must be Paretian; namely,

$$P(z) = \text{const} \cdot z^{-\alpha}, \quad 0<\alpha<1. \tag{8}$$

In particular, for the fair game, $\alpha=0.5$. This Paretian probability, and the dynamics of the ruin-free intervals that it yields, is identical in structure to the probability functions introduced by Mandelbrot in [3] and [9] for the error-free intervals in communications systems. For the German Federal telephone system studied in [3], the exponent, α, was about 0.3. More generally, it can be shown [8] that the ruin-free intervals of time corresponding to (8) constitute a set with Hausdorff (fractal) dimension, D, equal to α [13]. Again, from Schroeder [14], we have that the Hausdorff fractal dimension of a set is related to the Hurst exponent as

$$D = E + 1 - H \tag{9}$$

where E is the topological dimension of the set. In the case of the ruin-free or error-free intervals, $E = 0$ (we are considering points on a time-line). So Mandelbrot's German Federal telephone data are consistent with a process having a Hurst exponent of about $0 + 1 - 0.3 = 0.7$.

To introduce some of the other characteristics of this class of accumulated data flows:

1. The variance of the sample mean decreases more slowly than the inverse of sample size, in particular, it decreases as $n^{-\beta}$.
2. The autocorrelation decays hyperbolically instead of exponentially, in particular, it decreases as $k^{-\beta}$, for large k.
3. The power spectral density approximates a "*1/f* noise" near the origin, in particular, it varies as f^{γ}, where $\gamma=2H+1=1-\beta$, as $f \to 0$.

If we take it as given that the overflows of any real channel server queue, left to its own devices, will have the fractal structure just described, then it stands as a natural consequence that the attainment of any given grade of service, expressed in terms of the probability of data loss per unit time, or per unit of data transfer, cannot be assumed to be a strong function of the ratio of offered load to channel or server capacity. Rather, since data loss is virtually guaranteed, the issue instead would appear to be the development of strategies to deal with this simple fact.

D. Strategies for managing Paretian offered load

On this count, there appear to be two extreme paradigmatic cases worth consideration. In the first, a buffer of extreme depth would be created and managed in such a way that the channel would rarely, if ever, be empty. This, of course, requires that data would be held for an indefinite period in order to utilize every possible opportunity for data delivery. While an infinite buffer may not be practical, buffers capable of storing several busy-hours of data may can be created in many cases. Whether such delays would constitute a generally acceptable grade of service is a different matter; that is, what magnitude of delay is perceived by subscribers as message loss? (Practical experience [18] in paging suggests this delay value is around 6 minutes. Delays greater than this value in a uni-directional paging network virtually guarantee an avalanche of message retransmissions by the originators.)

In the second strategy, network elements and channels contain no appreciable buffers at all. Rather, elements and channels are designed to withstand a specified peak flood of data relative to an assumed busy-hour average, and any loss due to server overflow is left to end-to-end transport layer functions between communicatings hosts to manage. This end-to-end argument is itself embodied in the Internet and represents a significant and powerful design philosophy [15]. This is not to say that the end-to-end argument is without its detractors, especially in the context of wireless data transfer. Pentikousis in [16] presents a number of powerful arguments for the failure of current versions of TCP to adapt successfully to the conditions presented in an Internet that merges both wired and wireless LANs and WANs.

From a very broad perspective, message transfer from sending to receiving host is intended to be perfectly accurate. Messages that are not correctly received must be resent. In the end-to-end model, accurate transmission is left to TCP operating in the two communicating hosts; and the retransmission point is the originating host. Calling the first approach the "proxy" model, the information that the other end host is a wireless device is hidden from the wired host. Agents in the wireless network accept data from the Internet using common protocols and then relay that information using protocols that compensate for the wireless network's vagaries, without

reflecting their impact to the wired host. The retransmission point is a wireless network element.

Commensurate with the fractal arrival structure of retransmissions due to queue floods and uncorrectable message errors, the throughput of a pure TCP/IP network will have a significant peak to mean ratios, from 2:1 to 5:1 or higher. This is entirely consistent with the relatively high fractal dimension of the summation of the independently fractal sets for queue overflow and error rate. In contrast, the proxy model typical of paging and ReFLEX networks, throughput during busy-hour offered load conditions tends to have peak to mean ratios of 1.5:1 or less. This is consistent with the smoothing effect of distributed buffering together with the reduced raw error rates of this air interface. In other words, these networks build in elements to reduce the fractal dimension of raw offered load as it is mapped onto throughput.

E. Corrected Capacity Figure of Merit

In [17], Angus has computed the time bandwidth products needed to transfer 100 octet messages over a variety of wireless messaging networks. In that paper, these data were used to show network efficiencies for short message transfer in bits/s/Hz before and after the impacts of message retransmission and queuing were estimated. In Table I, the estimates of [17] have been used to compute our dimensionless capacity figure of merit ($N{\cdot}B{\cdot}T$) for 100 octet message transfer. The figures in the table assume that message retransmissions due to queue overflows and datagram errors are fully accounted for, that the reuse in GSM/GPRS is N=3, that the reuse in ReFLEX v2.7 is N=4, and that the peak to mean ratio for GPRS is 5:1. Other potential IP network "costs of operation" have not been allocated into the GPRS figure of merit. These would include access to Domain Name Servers (DNS) , IP address assignment via Dynamic Host Configuration Protocol (DHCP) or the equivalent, header overheads for Simple Mail Transfer Protocol (SMTP) where it might be used, and the like. It is difficult to estimate these overheads.

Our corrected capacity figure of merit for message transfer does something that the original, $N{\cdot}B$, did not have to do in the telephony case; namely, it adds back into the picture the reused bandwidth costs of retransmission needed to support a 100% accurate and reliable message transfer grade of service.

TABLE I

CAPACITY FIGURE OF MERIT FOR 100 OCTET MESSAGE

Protocol	$N{\cdot}B{\cdot}T$
GSM SMS	64,500
GPRS CS-1	33,333
ReFLEX v2.7	22,668

IV. CONCLUSION

Reference to Table I reveals that ReFLEX v2.7 stands out as an efficient air interface for the accurate transmission of short messages relative to other representative wireless data protocols. While this result is not without its costs in mean latency and simulcast infrastructure, it is of some interest that data transfer efficiency can be achieved in subtle ways, by the appropriate management of the structure of throughput, rather than through the more obvious tactic of increasing raw channel capacity.

REFERENCES

[1] D. Spohn, *Data Network Design*, McGraw-Hill: New York, 1993, pp.514-517.

[2] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)", *IEEE/ACM Trans on Networking*, vol.2, pp.1-15, 1994.

[3] B.B. Mandelbrot, "Self-Similar Error Clusters in Communication Systems and the Concept of Conditional Stationarity," *IEEE Trans Communications Technology,* vol.13, pp.71-90, 1965.

[4] M. Schroeder, *Fractals, Chaos, Power Laws*, W.H. Freeman, NY, 1991, p.167, 209, 321.

[5] B.B. Mandelbrot, *The Fractal Geometry of Nature,* W.H. Freeman, NY, 1983, Ch.8, p.79 and Ch.31, p.280.

[6] Schroeder, *ibid.*, Ch.5.

[7] *ibid.*, pp.145-7.

[8] *ibid.*, pp.152-153.

[9] B.B. Mandelbrot, *ibid*., p.280-4.

[10] H.E. Hurst, "Long term storage capacity of reservoirs," *Trans. of the American Soc. of Civil Engineers*, vol. 116, pp.770-808, 1951.

[11] H.E. Hurst, "Methods of using long-term storage in reservoirs," *Pr. of the Institution of Civil Engineers*, Part I, pp. 519-577, 1955.

[12] H.E. Hurst, R.P. Black, and Y.M. Simaika, *Long-term storage, and experimental study*, Constable, London, 1965.

[13] B.B. Mandelbrot, *ibid.,* pp.363-364.

[14] Schroeder, ibid., p.137.

[15] J.H. Saltzer, D.P. Reed and D.D. Clark, "End-to-End Arguments in System Design," *ACM Transactions in Computer Systems 2*, vol.4, pp 277-288, November, 1984.

[16] Pentikousis, K., "TCP in wired-cum-wireless environments", http://citeseer.nj.nec.com/436764.html.

[17] A. Angus, "Short Message Transfer in Narrowband and Broadband PCS," unpublished, available from the author or at http://www.braddye.com.

[18] John Davis, personal communication.