

Short Message Transfer in Narrowband and Broadband PCS



Why NPCS is better

Introduction

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

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 than in all previous history. Consumers are demanding mobile messaging.

The paper brings to light the technical reasons why narrowband wireless data systems deliver more reliable messaging with superior mobile device performance than with broadband systems. First and foremost, the longer bit times and narrower noise bandwidths of NPCS give it about three orders of magnitude better E_b/N_o performance relative to GSM Short Message Service (SMS). Second, the inherently diverse forward and reverse channels used in NPCS will typically add another three orders of magnitude improvement in E_b/N_o . Third, these significant radio link factors allow an overall network approach of assured message delivery in the NPCS case with reasonable penalties in air interface bandwidth consumed. Fourth, these gains in link performance are achieved without driving up the power levels of the Mobile Stations (MS). Fifth, unlike almost any BPCS protocol, NPCS does not require any sophisticated management of time alignment, power levels, logical channel

allocations, or other attributes of the radio link, implying that there is no distinction required between random and scheduled access to the air interface. Finally, and oddly enough, NPCS makes more efficient use of spectrum for short message transport than does broadband, implying that it is more cost effective.

For a wide variety of short mobile data transport applications, NPCS better satisfies consumer wireless messaging requirements at current implementation levels.

"Narrowband" & "Broadband"

There have been twin drivers in the move away from narrowband communications. First, increasing the number of point to point communication streams that can be supported on a single transmitter or receiver has been viewed as an improvement in capital efficiency by service providers. Second, increasing the data rate of any individual stream allows for more advanced communication services to be offered by service providers. The two drivers are not necessarily independent: one way to increase the data rate of any individual stream on a carrier is to dynamically allocate more of the bandwidth of a multiplexed channel to a single source for some period of time. We shall see this approach later in our discussion of the General Packet Radio Service (GPRS.) Both of these drivers yield a trend to higher bandwidths per modulated carrier.

But in the drive towards higher bandwidth over cellular and PCS networks, it is easy to overlook some basic communications physics. This is of fundamental importance in contrasting Narrowband and Broadband PCS offerings.

First, as the names suggest, NPCS uses channels of lower bandwidth than BPCS, 12.5 kHz versus 200 kHz or more. This is associated, obviously, with a lower bit rate per radio carrier in the NPCS case. Conversely,

NPCS has the longer bit times, T_b .

In binary communications, the receiver detects energy in each bit time and analyzes this energy in some way in order to select one of the two choices; say ‘1’ or ‘0’ for the received information. In doing so, the receiver must discriminate as well as it can against noise energy and other forms of symbol distortion and interference. As a broad generality, the energy per bit, E_b , increases both as the power received and as the bit time:

$$E_b = P_r T_b$$

Likewise, the noise per bit, N_o , increases as the bandwidth, W_b , required to detect the energy per bit,

$$N_o = \sigma W_b = \sigma / T_b$$

where σ is the noise spectral density [1], about 5.85×10^{-21} J·s at normal temperatures. The so-called contrast ratio, $\gamma = E_b/N_o$ is of central importance in the estimation of receiver performance; and

$$\gamma = P_r T_b^2 / \sigma$$

Note that the contrast ratio scales as the square of the bit time at constant received power. This is central in what is to follow. [As an aside, the unit, Joule-seconds, is the unit of *action*. This equation can also be seen as the ratio of *action* per bit to noise *action*. Since most communications engineers take contrast ratio to mean a ratio of bit energy to noise energy, it is often expressed in *decibels*.]

The GSM and SMS

The GSM allocates carriers on 200 kHz channel centers, and transmits at a raw rate of 270.833 kbit/s. Transmissions are framed at a repeating period of 4.62 ms, and each frame comprises 8 slots of 576.9 μ s or 156.25 bit times.

Slots can contain bursts of transmission in either direction [2]. Several kinds of transmission burst are defined, of which only two carry user data. These are the normal burst and the access burst.

A normal burst has a data payload of 116 bits; the rest of the 156.25 bit times being allocated for synchronization, guard time, training, and so on. The normal burst, as the name suggests, is the one intended to be used for

almost all logical channels at the air interface.

An access burst has a data payload of 36 bits; the bulk of the 156.25 bit times being idle guard time. An access burst is used only for random access to the air interface, and is designed to solve a particular problem of transmission physics, as it were. The issue is the unknown pure time delay from the mobile station to the base station receiving antenna at the time of random access.

Because of the finite speed of radio transmission (about $300,000 \text{ km}\cdot\text{s}^{-1}$) for cells of radius over 30 km, there may be 100 μ s of more of difference between the arrival times of random access bursts launched by two MSs, one near a base station (BS) and another at the cell boundary. The design solution embodied in the access burst, used on the Random Access Channel (RACH), is to shorten the active transmission interval during the access burst so that heavily delayed transmissions from MSs near the cell boundary cannot “spill over” into adjacent time slots.

In practice, the 116 bit payload of the normal burst and the 36 bit payload of the access burst are further reduced by coding for the logical channels that they carry. For example, in the specific case of a Broadcast Control Channel (BCCH) operating over normal bursts, 228 bits of data are run through a rate 1/2 convolutional encoder, grossed up to 456 bits, and transmitted in an interleaved fashion in 4 normal bursts. These 228 bits are comprised of 184 bits (23 octets) of BCCH data, 40 bits of CRC, and 4 tail-bits for the convolutional code. The coding model for the logical channels of the Paging Channel, Access Grant Channel, and Stand-Alone Dedicated Control Channel (PCH, AGCH, and SDCCH) is similar. On the RACH, using access bursts, the 36 bits are used with a rate 1/2 convolutional encoder to transport 18 bits of data, comprised of 8 bits (1 octet) of RACH data, 6 bits of CRC, and 4 tail-bits for the convolutional code.

Consider this information in the context of channel efficiency for a moment. First, let us calculate the time-bandwidth product used for message transport. In the case of the SDCCH, we have a 200 kHz channel consumed for four time slots of 576.9 μ s, for a time-band-

width product [3] (BT) of 461.5. This time-bandwidth unit carries 184 bits with a net efficiency of $184/461.5 = 0.399$ bit/s/Hz [4].

For the RACH, we have 8 bits in a single burst. The time-bandwidth unit is 115.3, and the efficiency is 0.07 bit/s/Hz. There is another effect that we need to account for here, however. The RACH is a slotted Aloha contention access channel, for which the theoretical maximum allocation efficiency is about 36% at infinite access delay. More practically, a single server slotted Aloha link will run at about 20% to 25% channel allocation efficiency in order to minimize access latency due to collisions and retries. This is in distinct contrast to the allocation of traffic channels for voice streams in a cell using the typical Erlang B (blocked calls are dropped) call model. In modern digital cellular, with more than 60 available trunks/channels per cell, trunking efficiencies approach 100%. Adjusting the RACH to account for contention access efficiency will increase effective

efficiency for BT reuse. In the original AMPS scheme [5], $N = 7$. In the GSM, $N = 3$. For a service provider with access to a total bandwidth of, say 10 MHz, the time-bandwidth product available per busy-hour per cell would be $3600 \text{ s} * 10 \text{ MHz} / 7 = 5.14 \times 10^9$ in the AMPS case, and 1.2×10^{10} in the GSM case.

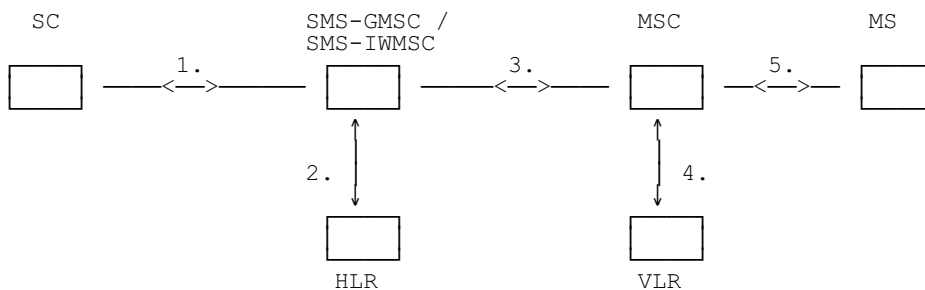
SMS call patterns

The reference model for short message service delivery is shown in Exhibit 1 [6]. In the forward direction, messages arrive at one or more Service Centers (SC) for Mobile Stations (MS).

These messages are relayed to the MS through the Gateway, Home switch (MSC-HLR), and Visited switch (MSC-VLR). The SMS is delivered over a Stand-alone Dedicated Control Channel (SDCCH). Idle devices do not operate on an SDCCH; and so, a call setup procedure for channel assignment to an SDCCH must

be executed at the Visited switch upon the arrival of the SMS. As for any Mobile Terminated (MT) voice call setup, this will typically involve broadcast of a page to the MS on a PCH, MS response on a RACH, notification of channel assignment on an SDCCH via an AGCH, and then message delivery.

Exhibit 1. Reference Model for GSM SMS



BT to around 500 and decrease the efficiency to about 0.016 bit/s/Hz. This 25-fold reduction in air interface efficiency constitutes a significant protocol design issue for GSM.

The drafters of the GSM have drawn a distinction between purely random access to the air interface and managed access. Purely random access is, by design, a pathway to managed access; and the duration of time for which random access is supported during any transaction is, reasonably, limited.

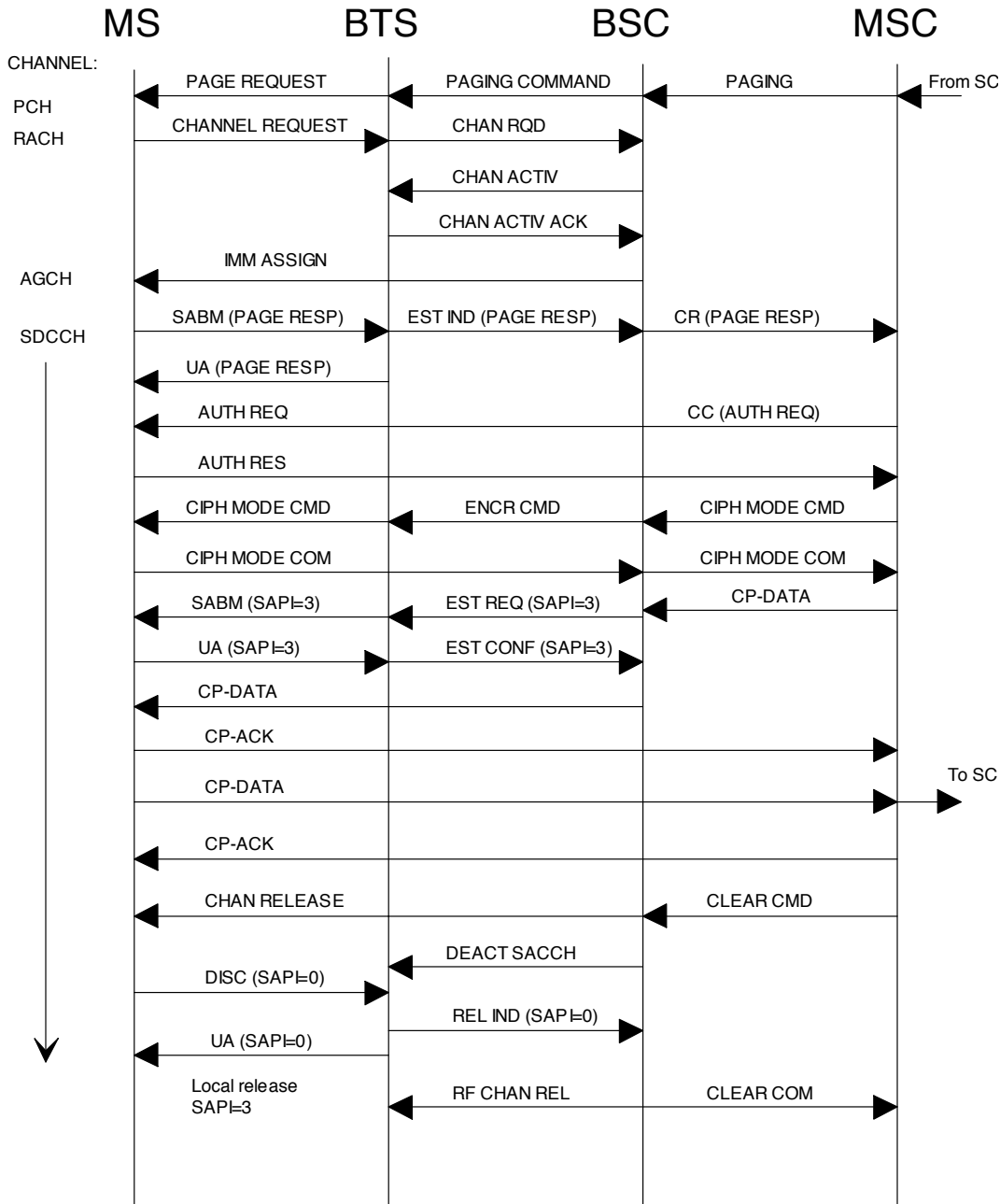
In a cellular system like the GSM, the fundamental air interface resource of time-bandwidth product can be reused in an orderly manner as defined by the reuse number, N . The lower is N , the greater is the opportu-

With success, the message is finally acknowledged. This pattern is shown in the ladder diagram of Exhibit 2 [7].

Mobile originated (MO) or inbound message transfer is similar, save that assignment to the SDCCH is done based on a MS request, initiated on the RACH, just as for voice call setup on a Mobile Terminated (MT) call. The system replies on an AGCH with a re-direct to an SDCCH, and the message delivery proceeds.

We can note that in order to complete an entire SMS delivery, air interface time-bandwidth resource is consumed at different rates at different stages of the process in order to accomplish the overall message delivery transaction. For example, the geographic scope of the

Exhibit 2. Ladder diagram for Mobile Terminated (MT) GSM SMS



outbound page to indicate a pending message must include at least all cells within a Virtual Location Area (VLA). Depending upon network configuration, failure of response in a VLA may lead to escalation of the page to a larger geography. Of course, such considerations are beyond the scope of the GSM standards as such; but this does not constrain their pragmatic value to the service provider attempting to optimize

the balance between air time consumed and message delivery success.

Given that the SMS user is actually interested in the successful transmission or reception of the content of his or her message, then all of the other “payloads” previously described here are immaterial to them. An accurate measure of transaction efficiency would constitute an accounting of all air interface time-bandwidth resource consumed in order to transport a successful message. This accounting should include any reuse efficiencies, or lack thereof. It should account for expected message retransmission probabilities due to errors or link failure. It should also include the effect of queuing, trunking, batching, and multiple access efficiencies. Using the ladder diagrams, one could proceed with a detailed analysis of air interface resource consumption on a transaction fragment by fragment basis.

However, such an analysis is largely pointless in the face of a great simplifier; namely, once the MS is assigned to the SDCCH, it is consuming 1/64th of a full 200 kHz channel for the duration of time that the overall message transaction takes to complete. This time duration will typically not be dominated by the time to transfer message content, but instead will be controlled by delay times for messaging between network entities at differ-

ent physical locations; for example, between MSC and BSC.

Whether the SMS transaction is MO or MT, once the MS is active on the SDCCH, a bandwidth of $200 \text{ kHz}/64 = 3.125 \text{ kHz}$ is fully dedicated to it. The pure transport delay time to move, say 100 octets of user data, at a rate of 23 octets per 4 bursts is about 20 ms. However, this must be adjusted to account for the fact that any user has only $1/64^{\text{th}}$ of the channel. The delay time becomes 1.28 s accounting for this fact. This pure transport delay will easily be doubled by a round trip delay of several 100 ms for each MSC-BSC transaction. Depending on how one counts, there are roughly four of these bi-directional MSC-BSC transactions: call setup, authentication, cipher mode set, and call release. Let us go ahead and assume that these transactions will take about 2 s to complete, purely as an order of magnitude guess. In other words, this doesn't all happen in 100 ms or in 10 s.

This simplifying assumption yields an estimate of $BT = 2 \times 200 \times 10^3 \times 2/64 = 12,500$ for the part of the transaction executed on the SDCCH, whether or not it is MO or MT. [Note that we have doubled the bandwidth consumed since both the forward and reverse channels of the SDCCH are allocated even though message flow is dominantly unidirectional.] Let us assume that the average user payload being delivered is about 100 octets.

As far as the distinction between MO and MT SMS is concerned, the key differences arise in use of RACH and AGCH for MO and PCH, RACH, and AGCH for MT. In the MO case, there is a net consumption of one RACH burst for call setup, ($BT = 500$) and of one block on the PCH ($BT = 460$). The net marginal BT for setup is 960 in the MO case.

In the MT case, this is almost identical, save for a scaling factor on the PCH, which must be transmitted across a VLA. Let us assume, for the sake of argument, that a reference VLA comprises 15 cells. We then have a effective BT for MT setup of about 8,360. Assuming that traffic is roughly bi-directional, the expected BT consumption per message is $(0.5 \times 960 + 0.5 \times 8,360) + 12,500 \approx 17,200$.

Since we have assumed that we move 100 octets, our aggregate efficiency in transport is about $0.047 \text{ bit/s/Hz/cell}$, referenced to a successful message delivery.

The NPCS case

The NPCS case is much simpler. The typical NPCS network uses 12.5 kHz channels at 6400 bit/s per stream. In the forward direction, the delivery unit is a code-word (CW) which transports 21 bits of data in a 32 bit time interval. A frame time of 1.875 s is established, and this is capable of moving 352 CWs in 11 blocks of 32 CWs.

The time per CW is 5.362 ms, and the raw payload is 21 bits. A typical message requires 4 CWs of address, 2 CWs of vector, and a number of data CWs that can be estimated as $(2 + \text{characters})/3$. For a 100 character message, this is simply 40 CWs. The BT is 2,680. The acknowledgement of the message requires a single inbound slot. See Exhibit 3. The inbound or reverse channel is divided into frames synchronous with the outbound stream. However, the slot structure is different, and at 6400 bit/s, there are 77 slots/frame. It is typical to allocate about 15% of these slots in each frame for slotted Aloha random access, leaving about 65 slots/frame for scheduled access. In the present analysis, we shall allocate the cost of random access uniformly across the 65 scheduled slots, and count the BT of the scheduled message acknowledgement as 360.

The total BT is 3040. Since characters are 7 bit ASCII encoded, the channel efficiency here is $0.23 \text{ bit/s/Hz/stream}$, referenced to a successful message delivery. [Note that the allocation of bandwidth in the forward and reverse directions is completely uncoupled in NPCS.]

In the inbound direction, the transport process is started with a single slot random access, but this cost of doing business is already allocated in our stream model. Messages are scheduled by a forward channel command (Schedule Inbound Message Command, SIM) of 11 CWs, ($BT = 732$). The inbound message structure includes 1 slot of message Start Address Unit (SAU) and a sequence of 3 slot Data Units (DUs) which have a user payload of 332 bits. This sequence is acknowledged

by a forward command of 7 CWs, ($BT = 466$) and completed with a scheduled acknowledgement from the mobile station ($BT = 360$).

For a 100 octet inbound message, the encoded content is typically 1,000 bits in size, because of FLEXsuite encoding rules. This implies the need for 3 Data Units, or 9 slots. The total scheduled inbound slot usage for the entire transaction is thus 11 slots ($BT = 3966$). The grand total BT is 5524. The net efficiency is $800/5524 = 0.14$ bit/s/Hz/stream, referenced to a successful message delivery.

If we assume an average across both messaging directions, this nets out to about 0.185 bit/s/Hz/stream, referenced to successful message delivery. And this is about 4 times better than the GSM SMS.

Message delivery failure

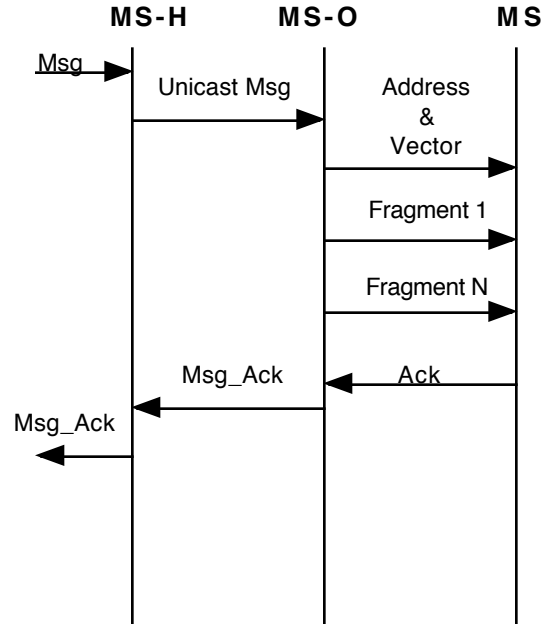
The observant reader will have noted that we have used a different reference point for the efficiency measures in the GSM and NPCS cases; namely, a *cell* and a *stream*, respectively. The physical radio resource associated with a cell is well-known and doesn't bear repeating here. The same is not true for the notion of a stream.

The physical resource associated with a stream will consist of a distribution of transmitter and receiver sites, typically, but not necessarily, collocated. Streams are attributes of a logical entity called a subzone, and subzones are elements of zones. A zone is usually associated with a serving area; for example, greater New York City. A market/zone like NYC may have on the order of 100 NPCS sites, shared between a collection of subzones, each of which will support one or more streams.

Reasonable numbers of sites per stream are of the order of 5 to 10 in a high density system. However, there is nothing to prevent the implementation of a system with 1 site or 100 sites per stream. Both extremes can be practical under appropriate conditions. On the forward channel, simulcast transmissions are synchronized by GPS to within $\pm 10 \mu s$ [8].

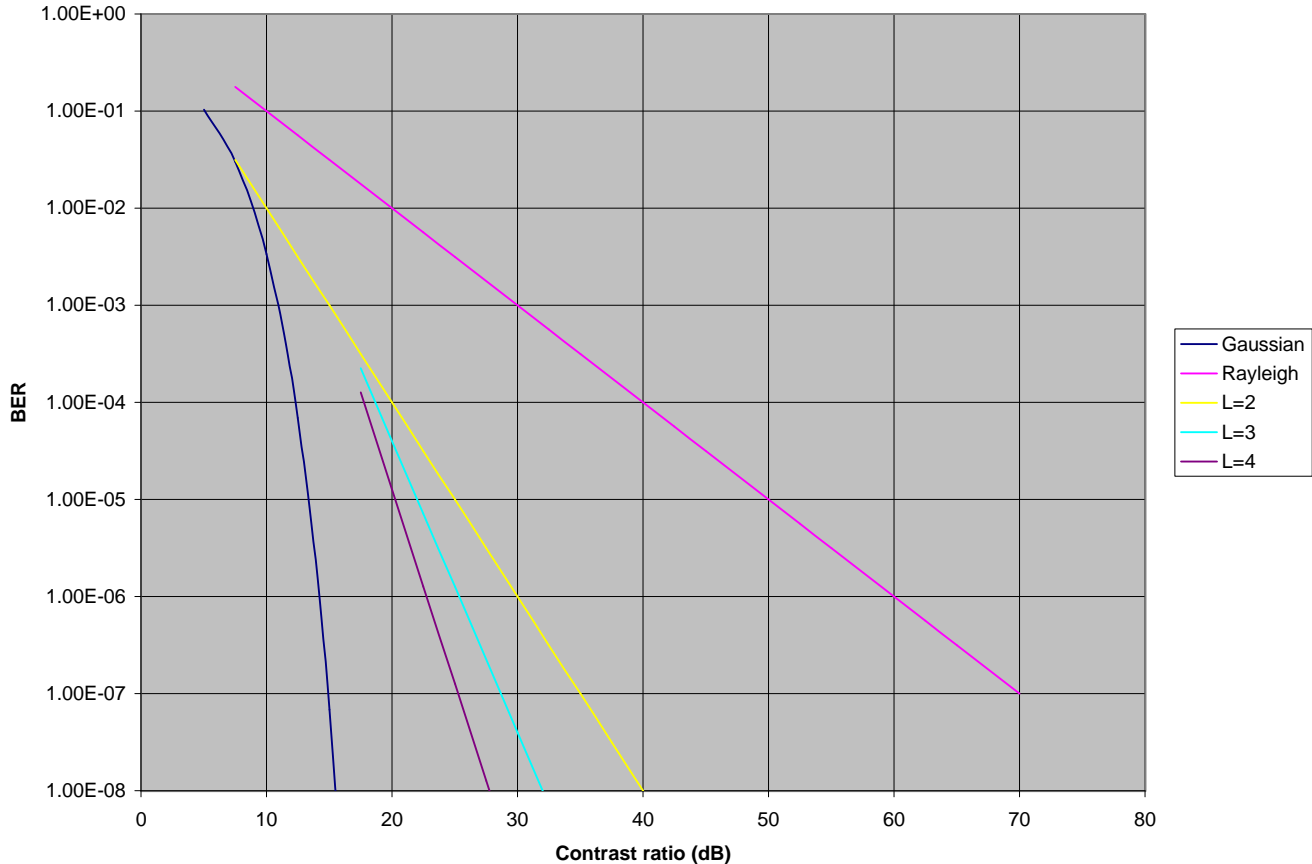
This implies a very different forward channel statistic than that due to the coherent scattering model associated with Rayleigh fast fading. It is well beyond the scope of

Exhibit 3. Ladder diagram for MT NPCS Message



this paper to present an analysis of this complex channel model, which is little known outside of engineers familiar with paging or public safety communications. Suffice it to say that the radio waves from any given transmitter are self-coherent, and subject to Rayleigh fading. However, during such a self-fade, the simulcast signals of other transmitters remain available at the receive antenna by linear superposition. There is a new effect though, “beating” of the signals of two non-coherent transmitters at their difference frequency. This creates a kind of deterministic fade process that can be pronounced under Ricean fade conditions at high MS elevations. A modern NPCS network optimizes this fade process by managing the frequency offsets of nearby transmitters such that the deterministic fade rate is dealt with by forward channel interleaving. A second channel degradation, simulcast delay spread (SDS), is managed by the use of non-linear optimization techniques in the setting of transmitter timing offsets, as well as DSP channel equalization in the MS receiver. In short, if one transmitter’s signal fades, the signal of another transmitter is there. If the signals of two transmitters beat against one another in a deep fade, then the signal of a third transmitter is still there.

Exhibit 4. Theoretical BER versus contrast ratio for non-coherent orthogonal FSK



Discussion of the NPCS inbound link is more straightforward, since it is a nearly pure diversity fading channel, which is well-known [9]. Each NPCS receiver is typically a two-branch diversity system with equal gain combining on the two optimal, independent, FSK FFT-based channel bank decoders. The typical slot sensitivity of these receivers is about -128 dBm. In addition to two-branch diversity at each base station receiver, the encoded traffic from each receiver is returned to a common logical management entity in the Output Messaging Switch (MS-O) which can perform either hard or soft decision diversity on the input of multiple receivers. This may be described as two-branch “micro-diversity” together with L -branch “macro-diversity” where L refers to the number of sites at which any mobile station’s slot transmission is decoded. The net effect is $2L$ diversity branches.

A well-designed NPCS system may offer expected macro-diversity on the order of 3 to 6 branches. The net effect on the inbound channel is dramatic. (See Exhibit

4.) While a digital cellular system operates in a Rayleigh fading environment at a point as close as possible to the raw 50% bit error rate (BER) in order to maximize cellular reuse, an NPCS inbound link operates much further down the raw BER curve. A raw BER of 3×10^{-2} is fairly standard for a cellular system at the carrier to interference ratio maximum.

It is clear from the Exhibit that the NPCS inbound link enjoys a large benefit in terms of raw BER versus contrast ratio due to diversity. This may amount to 30 dB or more under reasonably conservative assumptions about the effectiveness of inbound channel diversity. The net impact of the diversity structure of both forward and reverse channels in NPCS is a low message error rate relative to received signal strength, and certainly relative to the same operating conditions in a BPCS network.

It is typical in any zone to have a first attempt message success rate on the forward channel of about 90% to “available” devices, where “available” implies that the

device is actually active and registered in the market. With two retries allowed for in the logical MS–O, the message success rate to available devices becomes about 96% or better. Another way of viewing this process is entailed in the observation that about 75% of the forward channel traffic is “unique”; that is, has not been attempted before from the MS–O. The raw BER that will cause an error in a forward channel message of around 100 characters is about 1%, depending upon the correlation characteristics of the error process. Assuming that this rate happens then 10% of the time, and a negligible rate occurs the other 90% of the time, the average BER is about 10^{-3} . While this corresponds to a contrast ratio of about 15 dB, it should not be assumed that this is the expected operating contrast ratio on the forward link; rather, the non-linear mapping of contrast ratio to BER heavily weights the poorer operating locations of the serving area onto overall performance. *The same will be true for any messaging system.*

In a reference NPCS zone, about 10% of air time is used in messaging to unavailable devices, for the combination of new message attempts as well as device search procedures. This implies that about 4% of all message attempts fail, and must be resent from the MS–H. This resend rate, together with the 10% of air time to unavailable devices, as well as the MS–O retry rate all constitute a penalty function for guaranteed message delivery.

If we were to add this penalty function into the operational air interface efficiency of the NPCS network, we would wind up de-rating that measure by about $1/3^{\text{rd}}$. In other words, we would quote an effective BT per stream of 0.12 bit/s/Hz/stream, referenced to a guarantee of low-latency message delivery to any MS that is active in any zone of the NPCS network [10].

It is difficult to perform a comparative analysis of GSM SMS, since the network reference model does not deal with assured message delivery as a defined service option. Failed forward channel messages are discarded by the VLR/MSC and HLR/MSC, with an indication of failure and cause being transferred to the sending Service Center. Recovery of the link to the MS is in general driven by registration of the MS; in other words, the network does not engage in active search for the

MS. This will increase message delivery latency to units that happened to be out of good coverage temporarily during the message attempt. Since the network fall-back point for message re-transmission is the Service Center, the penalties for call setup, authentication, ciphering, message delivery, and call release are incurred for each message attempt. Note as an aside that message delivery will not even be attempted to MSs that are in a blocked state. For example, a device temporarily lost by the network during a message attempt will not be cleared for message delivery until it re-registers or makes a call attempt. In a network model that does not support time-based registration, this can take a while.

Whatever penalty function we might attempt to assign to the GSM SMS will be speculative. There are those operators who take no action to guarantee delivery at all, and this is within the scope of the standards. Others may implement a more persistent strategy of undelivered message attempts in their Service Centers.

Having said all this, let us assume that the first time attempt success rate to assumed available devices is not quite as good as NPCS; say 80%. Let's further assume that no MSC retries are configured. Also, let's assume that a much larger proportion of devices are unavailable at the time of message attempt, strictly on the basis that cell phones are more frequently “off” or out of coverage than an NPCS unit. Let's estimate this as 20% as against the NPCS network number of about 4%.

The net failure rate on first attempts is $0.2*1 + 0.8*0.2 = 0.36$. In the first case, the device being assumed unavailable, no air time is committed; the message is denied by the MSC/HLR. In the second case, the pattern of air time utilization is highly dependent upon the failure mode. In some instances, there will be the consumption of multiple pages to the MS with no reply. In other cases, the message delivery will fail on the SDCCH due to uncorrectable errors. The link layer protocol (LAPD_m) protocol may manage to deal with some failure modes, but will give up on significant link problems.

An estimate of channel bandwidth consumed on network retries is rather pointless, since these do not occur. The treatment of messaging is structurally similar to the Erlang B, block calls are dropped, model. In principle,

the GSM SMS network is engineered to “push back” at message attempts to MSs it assumes are blocked. The appropriate penalty function may then estimate at about 80%, but the service delivery model is not the same as in NPCS. The net delivery efficiency turns out to be about $0.8 \times 0.047 = 0.038$ bit/s/Hz/cell, referenced to a “blocked calls dropped” delivery model.

An economic comparison

The GSM’s transport efficiency difference is on the order of 3 to 4 times worse than NPCS in return for a less satisfactory message delivery service. From an economic point of view, one might argue that a service provider would then be indifferent to the cost of service in the two models at a point where the NPV of the costs of an NPCS stream was valued at about 3 to 4 times more than those of a GSM cell. However, this would not be comparing “apples to apples.” Let us make an attempt to normalize the notions of cell and stream in some way. Let’s start with the density of sites as a reference point. Assume for the sake of argument that we have an area served by 6 sites, in one configuration this will be an NPCS subzone, in the other an array of 6 GSM “center-fired” cells. Let’s assume also that the transmitters in both cases are of the same power. Our aim will be to compare two reference systems with the same overall bandwidth allocation.

We allocate a basic reuse pattern of spectrum to both systems; that is, 3 paired channels of 200 kHz, or a total of 1.2 MHz. In the GSM model, we have one paired channel at each of 6 cells, for a total of 400 kHz of spectrum per cell. Using the previous efficiency model, we get 0.038 bit/s/Hz/cell * 400 kHz = 15.2 kbit/s/cell. There are 6 cells with 6 transceivers, so the total bit rate in the serving area is 91.2 kbit/s.

In the NPCS case, the 600 kHz of paired spectrum is divided into 48 paired channels of 12.5 kHz both forward and reverse. Each stream is supported at the 6 sites by 48 transceivers for a total of 288. The net traffic in the serving area is 48 streams * 25 kHz * 0.12 bit/s/Hz/stream = 144 kbit/s.

If the cost of a transceiver at reference power were \$1,

then the GSM system generates data at \$0.065/kbit/s. Likewise, the NPCS system comes out at \$0.33/kbit/s. The GSM SMS wins by about a factor of 5. What value is obtained for this apparent cost?

Well, first of all, there is a difference in contrast ratio, or E_b/N_0 in the two systems. Returning to the equations that we began with, it is easy to show that

$$\gamma_G/\gamma_N = (T_{bG}/T_{bN})^2 = (6.4/270)^2 = -32 \text{ dB}$$

where γ_G is the contrast ratio in the GSM system, γ_N is the contrast ratio in the NPCS system, and the other two terms are the respective bit times. Also, to a reasonable approximation, power received falls off as the inverse 4th power of distance from a site at UHF.

$$P_r = P_t g_b g_m (b_b b_m / d^4)^2$$

where P_r is received power, P_t is transmitted power, g_b is base station antenna gain, g_m is MS antenna gain, b_b is BS antenna height, b_m is MS antenna height and d is BS-MS distance. Since coverage area will vary as the square of BS-MS distance, with a little algebra, we can take this last formula and derive that

$$Area_G/Area_N = T_{bG}/T_{bN} = 1/42$$

In other words, the coverage area of a GSM site is about 40 times less than that of an NPCS site, at a constant site density. Now, it goes without saying that cellular systems do not generally operate in a noise-limited environment. However, the serving area of a cellular site can certainly be no larger than that defined by fundamental noise limitations. That merely implies that a GSM transmitter must have somewhat greater power than that required to defeat the noise threshold at the cell boundary. Then, co-channel interference will almost invariably add a further reduction in effective contrast ratio. And we have not even touched on the irreducible BER factors associated with rms excess delay and Doppler. On the other hand, most NPCS streams will operate in a noise limited environment.

When we consider the incremental contribution of channel diversity over and above the contrast ratio gain due to lower bit rate in NPCS, the effective gain in NPCS to GSM is closer to 60 dB, a factor of 1,000,000 times! To make a ballistic analogy, if the penetrating power of a GSM bit were packed into a 1 gram projec-

Here is the crux of the issue: NPCS technology offers an economic entry into high quality messaging without having to construct a network with *either* too high a density and capital cost to be practical *or* too low a density and probability of message delivery to be practical. NPCS is cost effective at low scales and high scales, and avoids the “if you build it, they will come” syndrome.

The Evolution of GSM: GPRS

It is beyond the limited space available here to address in great detail the future evolution of both GSM and NPCS messaging. As is well known, the evolution of data transport in the GSM involves the General Packet Radio System (GPRS) and the Enhanced Data GSM Environment (EDGE.) GPRS adds a contention-based packet delivery model with an IP protocol stack at the MS. The EDGE provides a more efficient modulation method. We’ll spend some time on the GPRS and then limit ourselves to a few comments on EDGE.

The GPRS will introduce a new kind of call model, so that rather than being set up on an circuit-switched SDCCH for SMS, the MS will be assigned to a contention pool of radio resource. The aim is to “buy back” the air time on the dedicated channel by time sharing it among other potential users. In order to accomplish this, a new class of logical channel is introduced, Packet Data Channels (PDCH.)

An understanding of the impact of GPRS upon GSM overall can be seen from the reference model shown

in Exhibit 5 [11]. From the model, it can be seen that GPRS adds a parallel pathway for traffic to the MS that will interoperate, generally, with a packet data network like the Internet. The one source that can bridge into either the legacy circuit-switched pathways and the new packet-switched pathways is the Short Message Service Center.

The two new network elements are GPRS Support Nodes (GSNs). The Gateway GSN (GGSN) plays the role of protocol adaptation into the target packet network; for example, mapping native GPRS protocols onto Internet Protocols. The Serving GSN (SGSN) plays the role of managing data traffic flow to and from the MS. In that role, the SGSN is a network element equivalent to the Serving MSC on the circuit-switched side.

The protocol stack diagram for GPRS data flow is shown in Exhibit 6 [12]. From this, it can be noted that the MS does not “really” support an IP stack as is commonly believed. Rather, GPRS supports the Point to Point Protocol (PPP) between the MS and a GGSN, which can use a protocol like the Dynamic Host Configuration Protocol (DHCP) to assign IP parameters to a TCP/IP stack at the MS or within a host computer connected to the MS. Subsequently, IP datagrams can flow to the GGSN encapsulated within PPP over the various intermediate network elements. It is only at the output ports of the GGSN that IP routing between MSs and other hosts can begin.

Exhibit 6. GSM/GPRS Protocol Stack Diagram

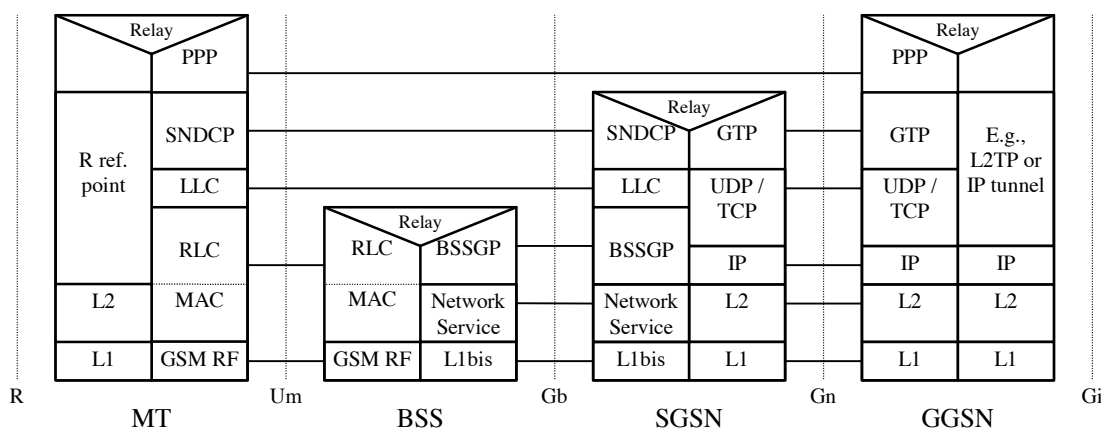
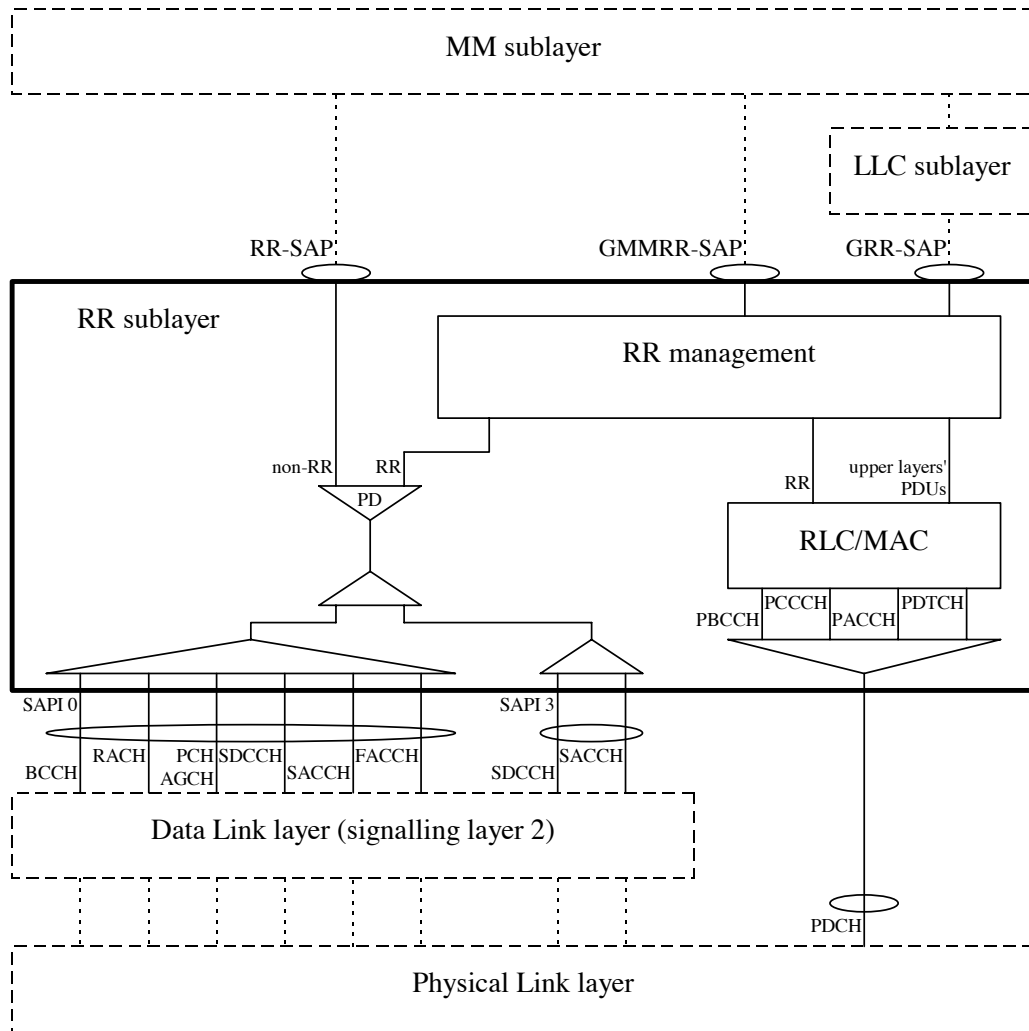


Exhibit 7. GSM/GPRS Logical Channel Model



To say this another way, the native address plan of the GPRS is *not* based upon IP. Rather, the GGSN is a proxy agent for IP address assignment and routing into the Internet, or into other packet data networks; for example, X.25/X.75 networks.

We have now seen the GPRS reference model and stack model. Let's consider its channel structure for a moment. This is shown in Exhibit 6 [13]. From a quick review of the diagram, it is straightforward to note that the new packet data channels constitute a logical mirror of the legacy circuit-switched channels already in the GSM. Furthermore, this new channel set is capable of being assigned a unique allocation of radio resource (RR) within every cell of a system.

The Flaw in GPRS

In the opinion of the author, this logical channel model, taken together with the previous network reference model, constitutes a significant architectural error. First of all, it creates contention between demand for radio resource allocation to circuit-switched or to packet-switched channels within every cell. Unfortunately, the points at which there may be effective and quantifiable measures for that contention do not communicate this information with one another. It is true that the GPRS standards define a kind of "capacity on demand" allocation model to deal with this problem [14]. However, the methods given to resolve the problem are not adequate to the task.

The two methods offered in the standards are “load supervision” and “dynamic allocation.” Load supervision involves monitoring the congestion or utilization state of packet channels. Dynamic allocation involves the recovery of unused packet channels and the allocation of new packet channels. The question is not whether dynamic allocation could be executed effectively, rather, it is that a running average estimation of packet channel usage may not constitute a “fair” estimator of overall demand for radio resource.

For example, moderately loaded packet channels may easily be improved with more bandwidth if circuit-switched channels are also under-used; but if circuit-switched channels are subject to heavy demand, moderately loaded packet channels could have radio resource removed.

More generally, efficient data transport across a communications medium is impossible without a queue to hold data and buffer it prior to delivery to the channel. Queuing is not a function embodied in the GPRS architecture. This is typical of the design of most IP routers and packet switches. In such a network layer model, datagrams that cannot be delivered due to media congestion are not queued, but are eligible for discard.

In the GSM/GPRS network reference model’s architecture, two very different classes of traffic contend for a common resource pool; namely, physical radio bursts. At the circuit-switch fabric of the MSC, an Erlang B call model is maintained: blocked calls are dropped. At the packet-switch fabric of the SGSN, a kind of datagram forwarding model is maintained: blocked datagrams are dropped.

Packet resource congestion can be signalled upstream from the BSC to the SGSN, and then to the GGSN. However, in an IP-based network interoperating with a GPRS network, this merely implies that TCP will retry packets and eventually signal link failure to hosted applications. While some IP switches (at layer 2) manage access contention, no IP router (at layer 3) manages queues.

Focussing again on the network reference model, it can be seen that there is no common point of communica-

tion between the SGSN and the MSC that could possibly gain access to radio resource demand measures except the BSC. And the BSC does not receive congestion measures from these network elements. It is merely left to solve a virtually intractable, multi-dimensional optimization problem that arises on every sector face of every cell within its scope of control.

Now, since radio resource allocation to packet channels must be done on a “whole channel” basis, these allocations look like circuit-switched calls to the MSC. Let’s consider some data “call models” to estimate the resource utilization efficiencies of these packet channels.

Data statistics

One of the most common statistical data models is the Poisson arrival model. This statistical model assumes that there is a mean arrival rate of data packets, λ , and then gives the probability that a given discrete number of packets will arrive in a specified observation time, T . In the limit, as the observation time becomes large, the Poisson arrival probability density function approximates a continuous Gaussian density function in which the mean, μ , and variance, σ^2 , are both λT . To account for the variability in packet sizes, it is often assumed in simulation work that packet size is also Poisson distributed, or some other discrete variation of a Gaussian.

More recently, there has been considerable interest in more advanced statistical models for data traffic. These include power law and power-tail distribution functions, for which it may not be possible to define a variance. The author’s observations of traffic in NPCS networks is completely consistent with power law distribution functions. Unfortunately, this topic is well beyond the scope of this discussion, and will have to be treated elsewhere.

Nonetheless, the amount of data, measured in bits, that arrives for transport in any given interval will vary statistically on at least two dimensions: number of packets and size of packets.

Given a finite amount of radio resource for the transport of user data, and in the absence of queuing at the

SGSN, the question then arises as to the peak to mean ratio of data arrival in a transport “window.” Given Poisson statistics, the shorter the transport “window”, the more variable the volume of traffic that arrives. Said the other way around, the longer the transport “window”, the less the variability in data arrival from time to time.

The explicit reason for this is that the relative variation from peak to mean can be expressed in terms of standard deviation units. In the case of the Poisson distribution, the standard deviation, σ , is $(\lambda T)^{1/2}$. But since the mean is also λT , the variation relative to the mean, decreases as the square root of the mean. λT .

To give a specific example, suppose that in some transport interval we expect 10 packets. The standard deviation is 3.2; and if we design to 1% probability of over-flow, we should allow for the mean plus 3 standard deviation units, or about 20 packets. On the other hand, if we expected 100 packets, the design limit would be 130 packets at a 1% probability of over-flow. In the first case, we need to design for twice as much data as expected in order to drop 1% of the packets. In the second case, we need to design for 30% over the mean number of packets in order to drop 1% or less.

This is with Poisson statistics, which is too bad, since that is not the way in which real data networks operate [15], [16]. With real data networks, variability remains high in spite of the increases in observation interval. Said in more mathematical terms, the variance does not converge to a finite limit, or at least not rapidly.

Scaling phenomena in data traffic

Said yet another way, data arrival rates are highly variable at all scales of observation. A corollary is that it is impossible to build a queue deep enough to avoid data loss at specified probability. This last statement can be taken to read on the futility of queuing, and in one sense that is absolutely true. Queuing cannot reduce the likelihood of data loss due to extreme bursts in a power law or power-tail distribution environment.

In another sense, queuing data still turns out to be a useful function: while it doesn’t remove the possibility of

packet loss, it does allow for the efficient use of channel bandwidth.

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

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

and where

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

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. 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,

$$E[R(n)/S(n)] \rightarrow \text{Constant} \cdot n^H \text{ as } n \rightarrow \infty$$

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.

To summarize some of the characteristics of this class of data:

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

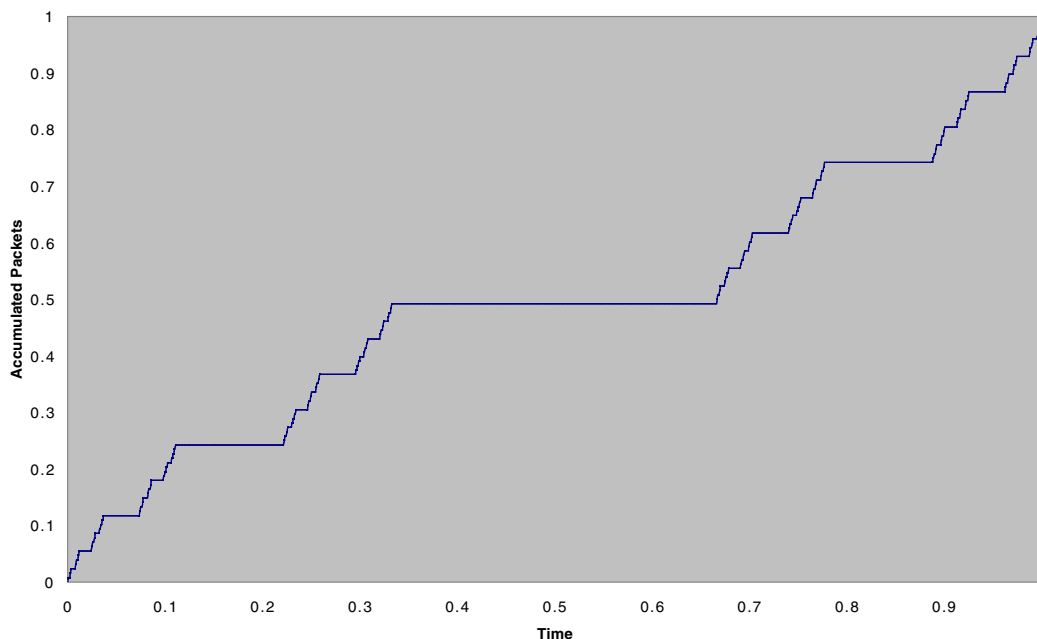
A full exploration of the implications of these scaling phenomena is beyond the scope of this paper. The interested reader is invited to explore the references. But there is one point that cannot be over-emphasized here: there is no strong convergence to an expected peak rate at any level of average flow.

A graphic example is given in Exhibit 8, a “devil’s staircase” [17]. Imagine this chart as the graph of accumu-

lated data flow on a channel. It would be feasible to compute an average data rate over the period of time in the graph; it turns out to be 1 data unit per time unit. But the curve is nowhere differentiable; that is, the slope of this curve does not converge to anything as the scale of observation is reduced. Also, the slope over any sub-interval bears no relationship to any sample of the slopes of sub-intervals within it. So, there is no well-defined local flow rate at any sub-interval, and no relationship whatsoever between the local flow rate in any sub-interval and the over-all average rate. This is what real data flows are like. It is also what error bursts are like. This is the mathematical face of the statement “data is bursty.” Although the curve shown is a mathematical abstraction, about the only things needed to make it a very practical approximation to a real data flow is to randomize it somewhat and add in a deterministic background of overhead routing protocol traffic.

Returning to the context of the GPRS logical channels, the upshot of this discussion is just that channel congestion can never be an adequate measure of channel demand in a bursty data environment with power law or power-tail statistics. The reason is simple: there

Exhibit 8. Devil’s Staircase



is no consistent relationship, statistical or otherwise, between congestion and mean data rate. In the GPRS model, congestion would be evidenced by packets dropped at the SGSN or BSC because of insufficient bandwidth to transmit them within the defined routing window.

These properties of real data flows are particularly problematic for any architecture that requires predictive estimation of load to act as a driver for bandwidth allocation. To return to the analogy of dam building, it is as if Hurst

had been asked to share concrete blocks for the Aswan High Dam with another dam on a nearby river. When the Nile was low, material for the Aswan Dam could be re-used on the other dam. When the Nile was in flood, the shared concrete could be returned to Aswan. Good luck, Mr. Hurst. Too bad everything floods at once! It is lucky that no one asks civil engineers to operate in the same way as communications engineers.

There appears to have been a view in the cellular world, since the inception of the Cellular Digital Packet Data (CDPD) system, that data transport is “for free”; that there are interstitial, unallocated “chunks” of bandwidth that are available for use to transport data without cost to the voice service. CDPD made the same architectural error as the GPRS, requiring a special logical channel for data, a kind of “data taxi.” And like people observe with real taxis, the streets are full of them until you need one. Apparently, demand for taxis comes in floods as well.

In purely data-oriented networks; for example, Ethernet local area networks, there is a trend to higher and higher back-bone bandwidths relative to customer demand. The operators of these networks have learned a lesson which will likely come as a shock to those who venture into GPRS; namely, by the time you realize that you are out of bandwidth, your customers are already either angry or gone. And when you have a lot of bandwidth to spare, chances are good that there will be a “data drought.”

Prediction versus “Retrodiction”

Let us contrast the predictive approach of GPRS radio resource allocation with that most used in NPCS and similar networks. There are three paradigms used in the broad field of signal processing: prediction, estimation, and smoothing. In predictive filtering, information available up to the present instant is used to predict some parameter at some time in the future. In estimation, information available up to the present is used to make a best estimate of a parameter’s “true” value now. In smoothing, information available up to the present is used to estimate some parameter’s “true” value at some point in the past. Implicit in all of these statements is the fact that the available information is likely to

be incomplete, distorted, subject to additive or other noise, generally, in one or more ways, questionable. By “parameter”, we more properly mean the value of some “state variable” of a system.

In the context of estimation, one is usually interested in some minutiae of a signal in real-time; for example, whether the energy of its content in some small frequency range is greater or less than that in another. This example would be pertinent to the design of a binary FSK receiver.

In the context of prediction, one is usually interested in the future values of a system’s state variables so that one may control them, or take advantage of them. In the systems theory of predictive filtering, the evolution of a system is broken down into two parts: first, its behavior in the absence of any new inputs; and second, the “innovations” that might occur because of unforeseeable events. Analysts attempt to extract the maximum amount of information that can be obtained from the first class of behaviors, since, in principle, the “innovations” cannot be forecast. An example might be prediction of the future position of a target aircraft. Given accurate information about its present velocity and mass, Newton’s laws of motion and gravitation would tell us where it will be at some future in the absence of pilot action to apply new forces. The “innovations” of the pilot are what simple physics cannot account for.

Finally, in the context of smoothing, one is usually interested in making maximum use of all available information that may be brought to bear on a task. Performing the task in “real-time” is not relevant. An example might be removing noise or distortion from a stored audio recording. The “smoothing” filter can access information in the recording that occurs both before and after the instant for which the output signal is being generated. Since such a filter uses “future” information, it can only be “causal” if its output is delayed in time.

Inasmuch as the GPRS aims to optimize its allocation of radio resource to either circuit-switched or packet-switched logical channels on the basis of local congestion metrics, it is predictive. Since blocked circuit-switched calls and packets subject to congestion are

dropped, increase radio resource allocations can only be of use to new incoming calls or packets, as the case may be.

In strict contrast, NPCS and similar networks will enqueue some volume of traffic, inspect the content for particular classes requiring special treatment, and then and only then, allocate radio resource to the traffic. This is “smoothing” or “retrodiction.” Bandwidth is allocated to traffic that arrived at some point in the past, by intent.

Given our earlier comments about the fractal structure of data traffic, its arrivals are very nearly a pure “innovations” process. This is not to say that total busy hour loads are dissimilar from day to day, or that one month’s aggregate traffic is not like another’s. Rather, it is to say that if the aim is to provide sufficient bandwidth to manage an expected load of X , then one better have on hand bandwidth for bursts of from $2X$ to $5X$.

Perhaps the reader will now have a stronger appreciation of the issues inherent in the “devil’s staircase” model for accumulated data traffic arrivals. As one inspects the curve of accumulated arrivals at decreasing time-scales, one finds little correlation between the accumulations in any adjacent time frames, and rapid “switches” between high and low flow rates. And because the curve is a fractal, this pattern is observed at all time scales.

It is as if data arrivals obeyed a scaled distribution of Poisson distributions; that is, as if data arrived from a population comprised of independent and uncorrelated subsets, each with its own unique mean arrival time, λ . This is practically consistent with some obvious models of population usage segments; namely, that some users send or receive data every few minutes, and others every few months. The problem is in assuming that these wide ranges of behavior can be captured by a single population average, rather than a distribution of populations with a corresponding distribution of averages.

The Flaw in GPRS, another view

We have observed that the designers of the GPRS introduced new data services by introducing a new suite of logical channels for packet transfer, in parallel with the

existing circuit-switched channels. We have called this an architectural flaw. Before we leave this point, we would like to offer an opinion as to why this structure exists and what might constitute an alternative approach.

Our belief is that the introduction of these new packet channels was done in order to minimize the impact upon legacy networks and devices. Any more revisionist approach would possibly have stranded millions of subscribers with legacy devices.

Our architectural alternative proposes the introduction of a common link layer bearer element that would support both voice and data content. This concept is not so wildly unusual. It is, for example, of the essence in asynchronous transfer mode (ATM). It is also the model in voice over IP (VoIP) as defined within the ITU-T H.300 series, specifically H.323.

Finally, it is a proposal that has arisen before in the context of wireless standards, certainly within North America. The link layer of the Telecommunications Industry Association TR45.3’s now-defunct IS-7x series of standards, written by this author in another stage of his professional career around 1995, was based upon the same concept. That set of standards was designed to replace the TDMA IS-54 protocol for D-AMPS. It aimed to support both circuit-switched voice and IP to the MS. It was supplanted by IS-136 given arguments in defense of legacy networks. Suffice it to say that little evidence has accumulated over the last 6 years to convince this writer that the allocation of radio interface bandwidth in a media-specific (i.e., voice or data) manner should be made below the network layer.

By avoiding any network architectural distinction in the logical channels that support given end-to-end applications (voice or data), the resource allocation problem simply never arises.

GPRS channel structure

Having presented a critical view of the logic of the GPRS channel model, let’s investigate their physical composition. GPRS channels operate over the same bursts as we saw earlier for the GSM. Groups of four

Exhibit 9. Basic GPRS Coding Schemes

<i>Scheme</i>	<i>Code rate</i>	<i>USF</i>	<i>Pre-coded USF</i>	<i>Radio Block excl USF & BCS</i>	<i>BCS</i>	<i>Tail</i>	<i>Coded bits</i>	<i>Punctured bits</i>	<i>Data rate kbit/s</i>	<i>Full channel data rate kbit/s</i>	<i>BT</i>
CS-1	1/2	3	3	181	40	4	456	0	9.05	72.4	0.362
CS-2	2/3	3	6	268	16	4	588	132	13.4	107.2	0.536
CS-3	3/4	3	6	312	16	4	676	220	15.6	124.8	0.624
CS-4	1	3	12	428	16	-	456	-	21.4	171.2	0.856

frames are grouped together to form blocks. A full-rate packet channel is “multi-framed” with 52 GSM frames. These frames carry 12 blocks for user data or channel control, and 4 additional frames that provide for timing and channel synchronization, the Packet Timing Control Channel (PTCCH.) A multi-frame lasts about 240 ms.

As we saw earlier in the basic GSM, each group of four slots, or 1/2 a frame, transports 456 bits of encoded data. Uniquely in the GPRS, four distinct coding schemes, CS-1 through CS-4, are defined that provide for different trade-offs between channel coding and user data throughput.

Some parameters associated with these schemes are shown in Exhibit 9. The basic scheme, CS-1, is identical to what we have already seen for the SACCH and SDCCH. The code rates given for CS-2 and CS-3 are approximate, after the removal (or “puncturing”) of coded bits as shown in the table. CS-4, which yields the highest data rate, provides no channel coding at all, just the Block Check Sequence (BCS), to test for bit errors.

The reference in the table to “USF” is for an overhead parameter called the Uplink State Flag. Its use is not important to our discussion, other than to note that each of the coding schemes treats USF somewhat differently.

More interesting are the quoted data rates, which refer to the allocation of 1/8th of a channel. Use of all of a physical channel yields the highest possible data rates of

72.4, 107.2, 124.8, and 171.2 kbit/s, for each of the four coding schemes respectively. Various data rates based upon fractional allocations of the total bandwidth of a packet channel are allowed for.

The corresponding raw values of *BT* for these channels are 0.362, 0.536, 0.624, and 0.856 bit/s/Hz, respectively. Note that the value of 0.362 bit/s/Hz is somewhat lower than our earlier value of 0.399 bit/s/Hz for the SDCCH. This minor difference is due to the overhead associated with the USF and the PTCCH.

Adjustment for channel efficiency

Earlier, we considered a simple Poisson traffic model that indicated a peak to mean ratio of 2:1, assuming 10 packets arriving in the observation window and a 1% probability of over-flow. We also considered a data model based upon “fractal time” [18], in which peak to mean ratios were essentially arbitrary. The author has observed peak to mean ratios of between 2:1 and 5:1 on commercial networks. These values imply a channel efficiency of no better than 50% or 20%, respectively. The practical impact of operating with less bandwidth on hand to manage peaks is a lower grade of service due to dropped packets. For CS-1, *BT* becomes 0.072 to 0.18 bit/s/Hz. This is about 2 to 5 times the *BT* for current GSM SMS.

Interoperation with TCP/IP

One of the noted features of GPRS is its support for a TCP/IP stack at the MS. More correctly, we have noted that the MS is capable of exchanging PPP encapsulated IP datagrams with its GGSN. Naturally, TCP protocol data units (PDUs) would be encapsulated within the IP datagrams. Now, TCP is a session-oriented protocol that supports peer-to-peer communication. About the only class of intermediate network element that inspects and deals with TCP is a firewall.

Our first comment in this regard is to essentially discount the value of the uncorrected CS-4 coding scheme in the table above. Since the destination IP address and TCP port will be carried in a GPRS PDU without error correction to the SGSN, the likelihood of receiving routable IP datagrams is slim in all but a fraction of locations. This scheme does not even allow for the MS to establish its own error correction procedures with the remote host, because IP datagrams in which the destination address cannot be trusted should not be routed. The flaw here is in the assumption that the IP control information is "user data".

In [19], Pentikousis discusses the impact of wireless communications channels on the performance of TCP. Among other degradations, he considers both the higher error rates and attendant retries as well as round-trip delay times (RTTs). This last effect tends to decrease the size of the "congestion window" (a kind of buffer) that TCP will manage at a wireline host communicating with a wireless MS. The smaller congestion window will reduce host-to-host throughput via TCP relative to communications with a wired host over links of the same bandwidth, but lower transport delay time.

On the topic of error rates, Pentikousis estimates the dropped packet rate at about 12% in an environment with about 10^{-3} BER after correction, quite typical of GPRS CS-1.

Pentikousis [19] also notes the impact of short data transfers on the flow rate as measure by the user. A typical TCP session will require at least 4 round-trip cycles to recover a short data message of, say, 500-1,000

octets. If the one-way time delay is of the order of a GPRS multi-frame, about 240 ms, then the entire transfer will take nearly 2 s, for an average throughput of 2.0-4.0 kbit/s, even if the allocated channel bandwidth were 170 kbit/s. In other words, *the availability of high bandwidth is irrelevant in the context of short message transfers anyway, given that MSs are "electrically distant" from Internet message servers.* Infinite bandwidth would not change this one iota, and the effect is even more pronounced for shorter messages.

In fact, GPRS RTTs can fluctuate through several orders of magnitude depending upon a variety of factors including message size, use of compression, the number of active users in a cell, and radio coverage. Shelton [20] describes a range from 20-30 ms at the low end to 20-30 s at the high end, making our estimate of a few hundred milliseconds quite plausible.

This effect also argues in favor of the use of CS-1 only for short message transfer. The alternative, the use of a higher bandwidth scheme, merely increases the probability of error, hence retry, hence longer RTTs, hence lower real bandwidth! Even with CS-1, and a dropped packet rate of 12%, the likelihood of one or more dropped packets in the previous 4 RTT transaction is about 64%. The time to retransmit dropped packets within the GPRS network marginally increases the one way delay time. The time to retransmit dropped packets at the TCP level, due to IP datagrams being lost within the GPRS network because of channel congestion is far worse. And yet, a "congestion driven" mode appears to be the design point for GPRS.

The expected number of packets lost due to errors, at a 12% drop-rate, will be about 1 in 8. The expected number of packets lost due to physical channel congestion will be dependent upon the actual bandwidth available to packet channels relative to demand bursts. For some short period of time, bursts of offered load can be enqueued and delivered later, so the pattern of actual throughput will be a "time-smeared" version of offered load. When the buffering time of network elements is exceeded, packets are dropped. In a true packet switch or router, hold times are very short, no more than

several milliseconds. So, peak to mean ratios of *throughput* can be forced to be as low as desired, by limiting the upper limit of bandwidth relative to average offered load.

Pentikousis [19] finally makes some observations concerning MS power consumption given TCP over IP as the communications mechanism. Suffice it to say that a MS engaged in TCP retries is a MS consuming its own battery life. Now consider two MSs communicating with one another via some TCP-based protocol for peer-to-peer messaging. Such a configuration would imply that the MS in better RF coverage would drain its battery in order to assure end-to-end communications with the MS in poorer coverage.

Of course, almost every practical proposal for MS to MS communication involves the intermediary of a messaging server of some form or other on the network; a mail server, an SMSC, or an Instant Messaging server. These intermediaries resolve the power drain issue to some extent by providing an “always available” network proxy for the destination MS. However, all of these expedients contribute to end-to-end round-trip delay times that further exacerbate the “short message” effect just described.

In the final analysis, given TCP/IP, high bandwidth per modulated carrier is for the service provider, not the user, until data content per transaction begins to exceed a threshold of about 10,000 octets, which is a typical threshold for TCP to begin to “stream” data.

For the sake of making a point about the inter-operation of the Internet with NPCS, the reader might note that all interworking is maintained through a network-connected proxy agent, which maps any TCP/UDP/IP protocol onto FLEXsuite. It is not beyond the pale that, at least in some circumstances, this proxy agent model might not yield lower interworking latencies than would arise with TCP being hosted at the MS, since its protocol adaptation functions could be optimized on the one side for the Internet and on the other side for the wireless network.

Corrected BT

We have not yet accounted for some of the other transactions that will drive down the *BT* for short message transfers in the GPRS. These include initial time and power alignments at the start of transmission, bandwidth requests and MS paging, authentication and ciphering, DHCP for IP address assignment, and grade of service negotiation, to name a few. We have already adjusted *BT* in the case of CS-1 to something in the range of 2 to 5 times what we found for basic, circuit-switched SMS. We now contend that accounting for these remaining overheads drives the final *BT* back to where it started, at about 0.018 bit/s/Hz, for messages of the order of a few hundred characters.

In short, there will not be much difference in the economic comparison of a pure GSM SMS network and a GPRS network for short message transfer, especially where the short message model is based upon TCP over IP. This is not to say that GPRS is completely without merit. It is only our point that its value is enhanced in the context of a user paradigm that involves relatively few, but large data transfers. Short, frequent, TCP/IP oriented, server-based, messaging is an Achilles’ heel.

Bluntly, this is not how to do wireless IM.

EGPRS, EDGE and 3G

To be fair, there are improved modulation and coding methods available in the so-called Enhanced GPRS (EGPRS) that are of merit. We have discounted these methods based the complexity of their actual deployment.

The EDGE’s more efficient modulation method would make another contribution to the *BT* efficiency of the GSM air interface. This author would argue that some of that raw bit/s/Hz efficiency will inevitably be given up in the heavier channel coding needed to support the poorer BER to contrast ratio of the linear modulation. Likewise, the same “short message” effect will apply to EDGE as it does to the GPRS, in other words, higher streaming rates are of little value when communicating peer hosts are sending short messages over “electrically

long” pathways.

Lest the IS-95 CDMA proponents stand up and argue that much of this paper is merely a recitation of their claims about the benefits of operating on Gaussian channels and that the 3G networks based upon direct sequence spread spectrum will have similar operating features as NPCS, with a tip of our hats in their direction, let us say, with some limitations, “yes”. The two most fundamental limitations will be that contrast ratio must still suffer as the bit rate per spread carrier is increased by simple physical arguments, and that the diversity achieved in a RAKE receiver is time-scale dependent.

We have addressed the first point already here. To the second point, not that in any situation where the natural excess path delay is less than a chip time, the RAKE receive cannot resolve distinct diversity channels. Likewise, if the absolute time delay or excess time delay span is large with respect to the scale limits of the RAKE delay line, energy in available paths may be degraded or lost. Mitigation of the first effect is expensive of network equipment, and basically requires a simulcast solution similar to NPCS in any case. Mitigation of the second effect requires greater complexity in the population of devices.

In the final analysis, many of the same effects will occur in the 3G versions of CDMA as do in the TDMA standards we have discussed here. None of them are targeted at, or optimized for, short message.

NPCS standards are also undergoing review and change. Under the auspices of the PCIA Paging Technical Committee, the author of this paper chairs a group looking at the next generation of NPCS. Whatever comes of this process, it is important to remember the fundamental distinctions between narrowband and broadband services.

Conclusions

This paper has presented a comparison of messaging services over broadband and narrowband PCS networks. The comparison has shown that while NPCS has a moderate advantage in terms of air interface efficiency, the cellular reuse capabilities of the GSM can erode the economic efficiency of NPCS to where the two are roughly at par. However, NPCS message delivery mechanisms have 3 to 6 orders of magnitude better contrast ratios than comparable services on the GSM. The impact of this on practical message delivery services in the NPCS case must be either improved grades of service for message delivery success or lower bandwidth consumption to complete messaging, or both.

In the broadband arena, we have considered both the current SMS over GSM as well as GPRS. We have found that for all of its added complexity, the GPRS brings little of value for short message transfer. Its strength would be in the transfer of larger and longer data content with infrequent arrival rates. We have especially noted that the higher advertised bandwidths of GPRS are virtually pointless to users in the context of short message transfer.

This is not to say that GPRS will not, or ought not, be deployed. Neither statement is true. Practically, GPRS may be the best available path for TDMA service providers to take to a broader range of data services than they can presently offer.

We have seen that GSM/GPRS and other 3G services being advanced by public PCS service providers will likely require far higher site densities than are presently used in North America in order to achieve satisfactory voice and data services for their customers. In fact, the desirable site densities may be very difficult to achieve with the present technology of cellular siting.

In contrast, that “density niche” is already being occupied by another species of wireless network; namely, IEEE 802.11b wireless LANs. The public broadband PCS providers are extremely likely to find themselves face to face with an array of IEEE 802.11b incumbents offering better pure data services for free by the time

they have worked out the problems with their data standards.

And as this niche of standards is further expanded in the direction of IEEE 802.11a and HIPERLAN, the situations under which public high speed wireless data access will be supplanted by superior private high speed wireless data access will naturally increase.

Perhaps the worst case, highest risk scenario for those engaged in the deployment of GPRS/GSM and the various flavors of 3G would be as follows. The customers that they attract dominantly want short message services, and that those who want high bandwidth wireless services migrate instead to the range of Wireless LANs now being deployed at no direct cost to the users. In this case, the factors that yield efficient SMS delivery in NPCS may become very relevant to 3G service providers, since this market segment is among the most price sensitive.

Indeed, building out a high density, high complexity network with an “if we build it, they will come” mindset, only to discover too late that customers are not willing to provide a sufficient revenue stream to yield a satisfactory return on the marginal investment for data services ought to be seen as a serious risk proposition.

In the final analysis, we argue that the niche will remain for highly reliable, genuinely ubiquitous, power efficient, short message transfer via NPCS. This is a service that private wireless LANs do not offer at all, and that the increasingly fractured set of public “2.5G” and 3G broadband services doesn’t do well.

Endnotes

1. $\sqrt{2kT}$ where k is Boltzmann’s constant, 1.38×10^{-23} J·s, and T is absolute temperature in Kelvin, about 300K at room temperature. The $\sqrt{2}$ comes from assuming optimal thermal noise power transfer between matched antenna and receiver input resistance.

2. We shall variously use outbound, forward, and downlink to refer to the BS transmit scenario; and inbound, reverse, and uplink to refer to the MS transmit scenario.

3. For those interested in network economics, the notion of time-bandwidth product may be monetized in a number of ways that can account for both network capital and operating costs, including spectrum license costs.

4. Note that the unit “bit·s⁻¹·Hz⁻¹” is really dimensionless, as is “s·Hz” or” s·s⁻¹”. This figure of merit could as easily be expressed as a percentage.

5. Ignoring details of sectorization.

6. GSM 03.40 , *Technical realization of the Short Message Service (SMS) Point to Point (PP)*, Figure 03.40/5.

7. GSM 04.11 *Point-to-Point Short Message Service support on the mobile radio interface*, Figure F1/GSM04.11.

8. Having said that, it is common to adjust delay and frequency offsets on a transmitter by transmitter basis to optimize delay spread issues within a serving area (or subzone), and to create serving area boundaries.

9. Proakis, J., *Digital Communications*, McGraw-Hill, Chapter 7.

10. We have applied the same penalty function to both forward and reverse channel. This is close enough purposes of this analysis.

11. GSM 03.60 *Digital cellular telecommunications (Phase 2+) General Packet Radio Service (GPRS) Service description; Stage 2, version 7.4.1*, Figure 2. ETSI 1998.

12. GSM 03.60 *Digital cellular telecommunications (Phase 2+) General Packet Radio Service (GPRS) Service description; Stage 2, version 7.4.1*, Figure 49. ETSI 1998.

13. GSM 04.60 *Digital cellular telecommunications (Phase 2+) General Packet Radio Service (GPRS); Mobile Station (MS) - Base station System (BSS) interface; Radio Link Access Control (RLC/MAC) protocol* Figure 1, ETSI 1999.

14. ETSI TS 101 350 *Digital cellular telecommunications system (Phase 2+); General Packet Radio Service (GPRS); Overall description of the GPRS radio interface; Stage 2 (GSM 03.64 version 8.5.0 Release 1999)*, Chapter 6.

15. W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson, “On the Self-Similar Nature of Ethernet Traffic (Extended Version)”, *IEEE Trans on Networking* (1994).

16. B.B. Mandelbrot, “Self-Similar Error Clusters in Communication Systems and the Concept of Condi-

tional Stationarity,” *IEEE Trans Communications Technology* COM-13, 71-90, 1965.

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

18. B.B. Mandelbrot, *The Fractal Geometry of Nature*, W.H. Freeman, NY, 1983, Ch8, p79 and Ch31, p280.

19. Pentikousis, K., “TCP in wired-cum-wireless environments”, <http://citeseer.nj.nec.com/436764.html>.

20. Shelton, D., “General Packet Radio Service (GPRS)”, <http://www.aethersystems.com/wireless/papers.asp>.



Allan Angus
Senior Member of Technical Staff
WebLink Wireless

For more information, contact:

Allan Angus, PE,
Senior Member of Technical Staff
E-mail: allan.angus@weblinkwireless.com
Phone: (214) 765-3470
2-Way: (800) 970-9325

Eric Van Steenburg
Public Relations Manager
E-mail: eric.vansteenburgh@weblinkwireless.com
Phone: (214) 765-3979
2-Way: (888) 816-4616