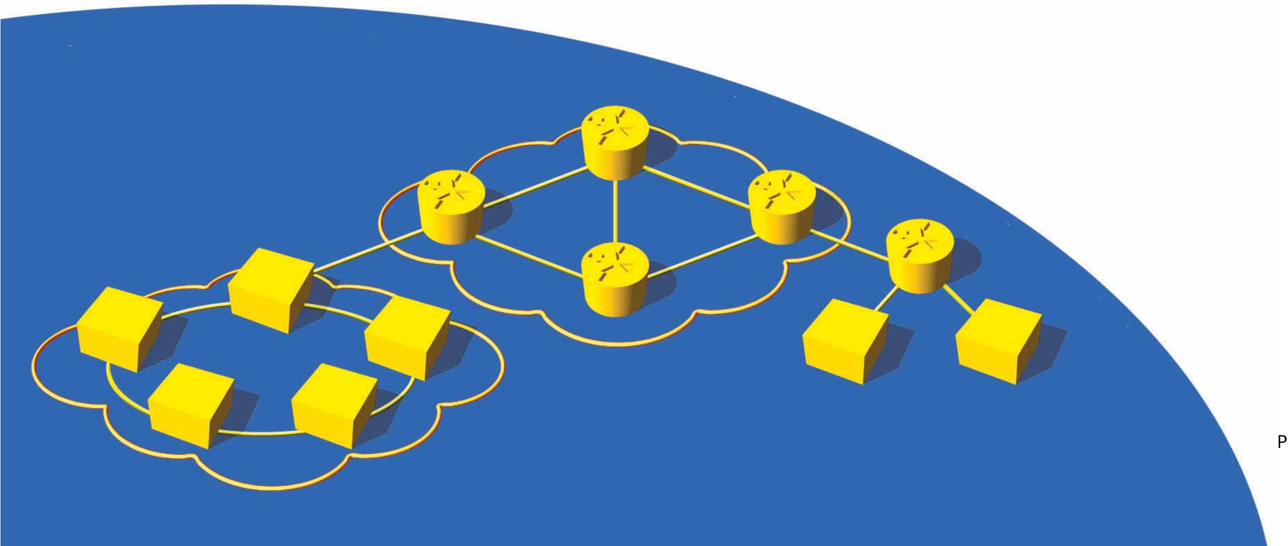
Analyze | Assure | Accelerate™

White Paper

Three Hidden Dangers of GPRS Deployment and How to Avoid Them

March 2004



P/N 340-1278-001 REV B

Spirent Communications, Inc.

26750 Agoura Road
Calabasas, CA
91302 USA

Email: productinfo@spirentcom.com
Web: www.spirentcom.com

North America

+1-800-927-2660

Europe, Middle East, Africa

+33-1-6137-2250

Asia Pacific

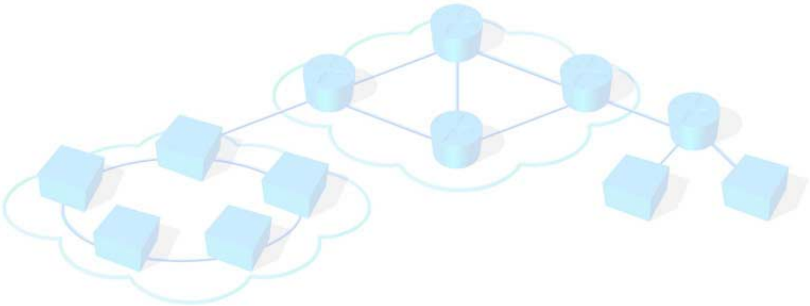
+852-2166-8382

All Other Regions

+1-818-676-2683

Copyright

The information contained in this document is subject to change without notice and does not represent a commitment on the part of Spirent Communications. The information in this document is believed to be accurate and reliable, however, Spirent Communications assumes no responsibility or liability for any errors or inaccuracies that may appear in the document.



Three Hidden Dangers in GPRS Deployment and How to Avoid Them

Growing demand for wireless Internet data access is fostering double-digit growth in new mobile data delivery technologies, especially GPRS. Indeed, GPRS subscriptions continue to grow at a rate of over 30% quarterly. [1] This paper discusses specific benchmarking recommendations for determining the performance of the core GPRS packet network.[2]

Contents

[1] Source: World Cellular Data Metrics – Quarterly Datasheet – December 2003, EMC, part of Informa Telecoms Group.

[2] Published in the 5th World Wireless Congress, San Francisco, May 25-28, 2004.

Introduction

GSM, the Global System for Mobile communications,[3] represents 70 percent of the world's wireless market, serving nearly 1 billion subscribers[4] in nearly 200 countries.[5] With such global coverage, GSM has become a leading focus for overlaying wireless data services onto an existing service provider network. Today, that data service is predominantly GPRS.

However, increasingly, operators are migrating to EDGE and even to UMTS while maintaining or lowering prices. This presents an increasing strain on the core packet networking as more users transfer more data more quickly and connect for longer periods of time. This, in turn, demands carefully-planned benchmarks to ensure proper capacity planning and robustness.

Hidden Dangers in GPRS Deployment

In spite of the attractions of GPRS and its successors, the sudden explosion in subscriber populations coupled with new models for flat-rate v. content-based billing is threatening both the robustness of the network and the operators' power to charge for services that they successfully deliver. In particular, there are three specific areas that demand special attention and examination.

Danger #1: Unknown GGSN Performance Ceilings

While GPRS's air interface throughput maximum of only ~150 kbps may seem inconsequential, mobile subscriber populations could quickly reach tens to hundreds of thousands. In such cases, even "low bandwidth" subscriber access links can quickly overwhelm even a Gigabit Ethernet GGSN uplink.

Consider that most of the largest GPRS operators are reporting quarter-over-quarter growth in double-digit percentages[6] – many exceeding 50%. Such explosive growth challenges even the most diligent capacity planning efforts and thus precipitates an urgent need for benchmarking for at least an order of magnitude greater subscriber populations.

Of particular concern is the concentration of subscribers on the GGSN. *Figure 1 on page 3* depicts the many-to-one cascade of base stations (BS) to base station controllers (BSC) to SGSNs to the GGSN.

[3] Originally, GSM stood for "Groupe Spéciale Mobile," the organization that spearheaded the initiative. However, the group soon changed its name to "Global System for Mobile communications" to give the new technology a more universal tone.

[4] Source: GSM Association, January 2004 press release at http://www.gsmworld.com/news/press_2004/press04_06.shtml.

[5] Source: GSM Association, April 2003 statistics.

[6] Source: World Cellular Data Metrics – Quarterly Datasheet – December 2003, EMC, part of Informa Telecoms Group.

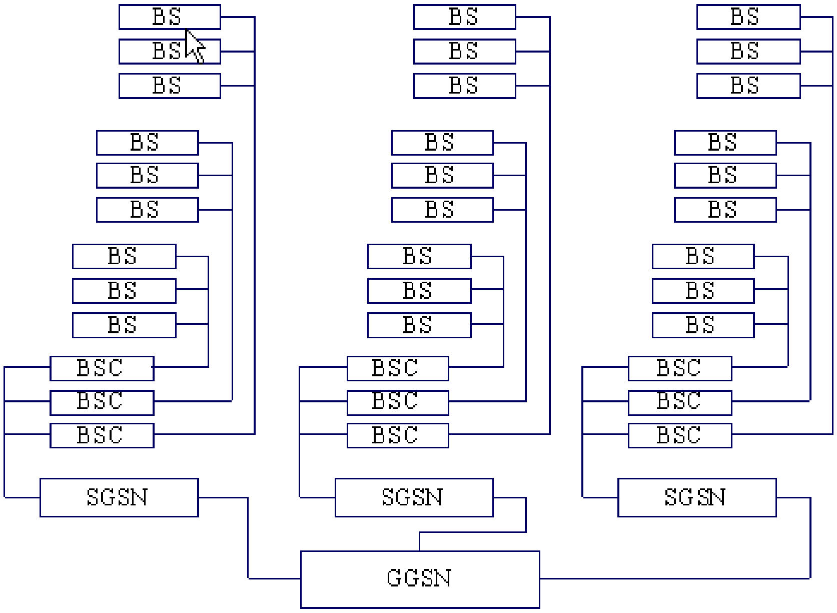


Figure 1. Many-to-One Cascade from Base Station to GGSN

Consider, too, that there is a cluster of mobile stations attached to each base station, and that each mobile station will generate at least one GTP tunnel and may generate multiple tunnels.[7] For this reason, the actual number of GTP tunnels to the GGSN may exceed the number of active subscribers. Moreover, subscribers who are attached to the network but inactive still consume resources on the GGSN, especially as those inactive subscribers roam between SGSNs.

It is clear, then, that GGSN performance remains a critical factor in GPRS network scalability, throughput, and availability[8]: The GGSN supports more GPRS connections than any other device in the network and therefore occupies a central role in the performance of the network. One could even argue that customer satisfaction – and therefore customer churn – is increasingly dependent upon the GGSN.

[7] Specifically, each mobile system will generate one or more primary tunnels and zero or more secondary tunnels.

[8] Eventually, UMTS will place even further demands on the SGSN. However, the current predominance of subscribers accessing the network via GPRS and EDGE places the burden on the GGSN for the foreseeable future.

Benchmarking the GGSN's performance to determine these ceilings requires that the test environment:

- Emulate thousands to millions of subscribers originating behind tens to hundreds of SGSNs.
- Exercise both steady-state and burst patterns for both bearer/data and control traffic.
- Emulate both single and multiple primary tunnel configurations per mobile station, and both zero and non-zero secondary tunnels per mobile station.
- Include inter-SGSN roaming, validating that all primary and secondary tunnels remain active during hand-off.

Figure 2 depicts a sample test of burst control and data plane activity. The filled triangles indicate the number of active GTP tunnels (in hundreds) while unfilled circles indicate instantaneous data throughput in kbps.

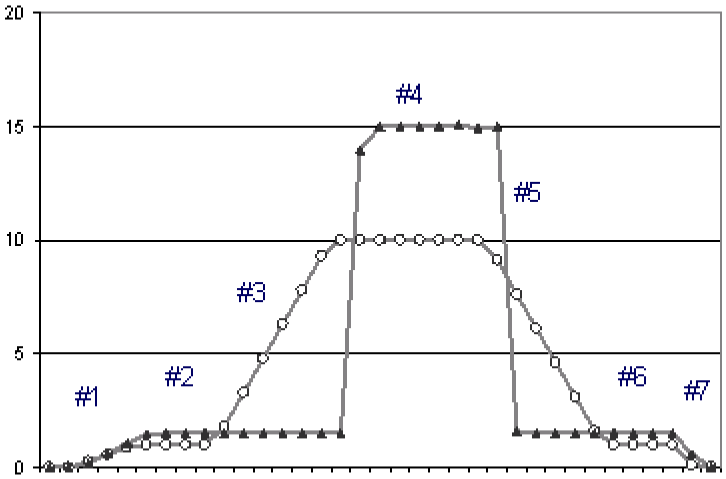


Figure 2. Sample Benchmark: Control and Data Plane Burst

1. The test begins with a few nodes attaching to the network and sending WAP data.
2. Once all GTP tunnels are active, data throughput reaches a steady-state plateau.
3. The test then initiates a burst of new GTP tunnels, simulating a wave of mobile users attaching to the network.
4. Once all of subscribers attach, the test immediately initiates HTTP requests across all of the active tunnels.
5. When all HTTP transactions are complete, the test begins to deactivate the more recent group of GTP tunnels, simulating users detaching from the network.
6. Shortly thereafter, the test returns to the earlier steady-state (*Step 2*).
7. After a brief period, the test ceases the remaining data transfers and deactivates the GTP tunnels.

Of course, this is merely one example of how control and data plane burst tests may operate – in this case benchmarking the number of active calls and the aggregate data quantity. Two other valuable performance metrics are GTP tunnel create (i.e., set-up) rate and tunnel create time. That is, inasmuch as tunnel creation is not necessarily a serial processes, the create time may not necessarily be the geometric inverse of create rate.

The graph in *Figure 3* illustrates this point. GTP tunnels are created every 25 ms (i.e., at an average creation rate of 40 GTP tunnels per second). However, each tunnel create process takes 125 ms to complete. Thus, the time between successive tunnel creations is independent of the time to complete a single tunnel creation.

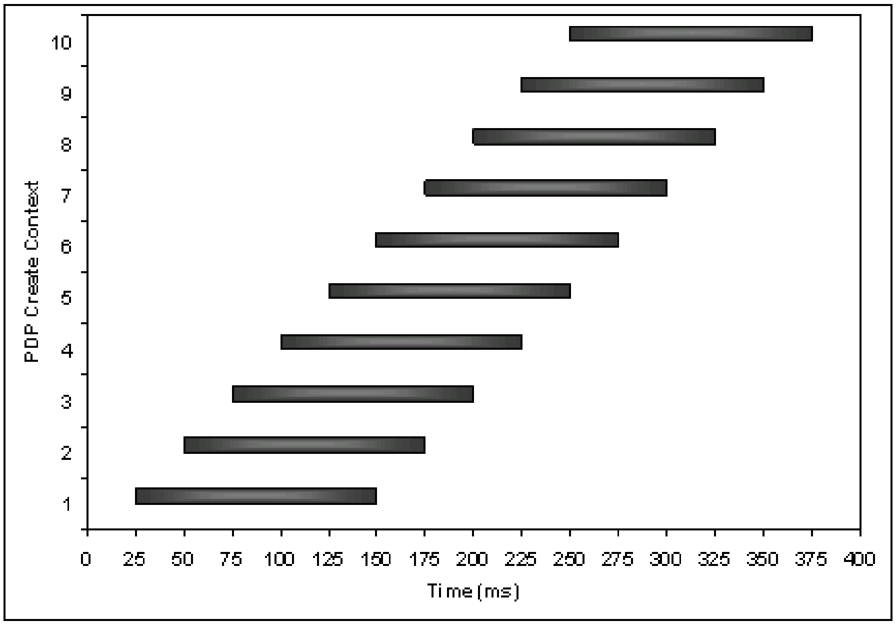


Figure 3. Tunnel Creation Rate and Time are Unrelated

Of course, the same model may be applied to tunnel delete rate and time, as well. Performance tests such as these enable accurate capacity planning through the knowledge of the GGSN's maximum performance and scalability.

However, beyond that maximum, the GGSN may exhibit different behavior in response to periods of oversubscription. This brings us to the second danger.

Danger #2: Nonlinear Oversubscription Response

Considering the maturity of the installed GSM voice network, one might expect that (a) GPRS data equipment is likewise capable of scaling to similar subscriber densities and (b) that its responses to oversubscription are reasonably well understood. Unfortunately, such response characteristics are seldom prescribed by operators. Indeed, some GGSNs may crash and/or reboot in response to certain combinations of oversubscription conditions.

Determining how the GGSN responds to different types of oversubscription demands benchmarking very specific oversubscription conditions, including:

- Steady-state tunnel creation and/or deletion at rates that exceed the GGSN's ceiling by at least 20%.
- Burst tunnel creation and/or deletion with a floor value near the expected high-water mark combined with periodic bursts of 20% to 100% over-subscription.
- Traffic bursts with a floor value near the expected high-water mark and periodic bursts of 20% to 100% oversubscription. Such traffic is likely to be asymmetric.
- Data/bearer traffic that includes both (a) a packet generation function to emulate streaming data and to ensure continued oversubscription, and (b) real application traffic to determine actual application response including possible session loss.

Such tests will prepare the operator for the network's response during periods of oversubscription. Furthermore, anticipating the likely severity and/or frequency of such events is absolutely necessary for proper capacity planning.

This discussion thus far has focused exclusively on basic network performance and response. This reflects the network pressures from emerging flat-rate mobile data packages that are becoming particularly popular among the business sector.

However, operators are increasingly selling into the non-business "personal" mobile data market with lower cost-of-entry packages coupled with "pay-as-you-go" options.[9] Such endeavors have precipitated an emerging trend toward "content-based billing." In this model, mobile data pricing is based upon a laundry list of packages, applications, and pricing strata. That brings us to the third danger.

Danger #3: Complex Application Tracking Requirements

The business case for per-application pricing models has already proven itself with today's plethora of charging models for messaging v. browsing v. traffic services v. news updates, etc.

For instance, many operators assign per-transaction charges for MMS (e.g., picture-messaging) based upon the expectation that an MMS message is likely to fall into one or two generally predictable file sizes.

The following illustration depicts a network configuration in which (a) SMS (text messaging) and MMS (picture messaging) are forwarded by the GGSN but not accounted and (b) HTTP is accounted differently depending upon URL – either per transaction or per kilobyte.

[9] Pre-paid services fall under "pay-as-you-go" inasmuch as payment received prior to service delivery is applied based on cumulative mobile data activity.

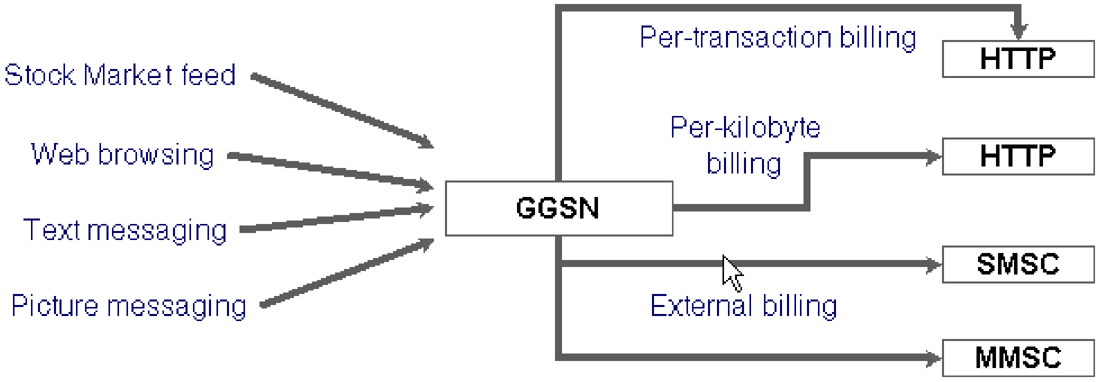


Figure 4. Sample Content-Based Billing Model

Similarly, many operators charge a per-message fee – albeit a lower one - for text messaging[10] and instant messaging. By contrast, operators often charge per minute or per byte for browser functions (e.g., WAP, HTTP).

If operators are to charge differently for different services, then they need a way to track such activity accurately and a way to validate that accuracy. For instance, they may need a method (a) to ascribe HTTP traffic per kilobyte to the subscriber's account but (b) to exclude SMS and/or MMS transactions since the latter will be charged separately and tracked by the SMSC or MMSC.

Therefore, benchmarking the GPRS network means exercising the accuracy of accounting and charging metrics. This means that the test application must:

- Generate real application traffic (i.e., not merely protocol headers), including multiple unique transactions per application (e.g., multiple URLs for WAP).
- Measure number, duration, and payload per application and, where appropriate, per content (e.g., per URL).
- Exercise content-based billing features when the subscription density and/or the data traffic load is near and above the capacity of the system under test.
- Exclude network retransmissions (e.g., TCP window retransmissions) from ascription to test results counters inasmuch as operators would usually not penalize subscribers for the operator's network anomalies.
- Compare data throughput and subscription density per application when content-based billing is enabled and disabled for that application.[11]

[10] Traditionally, text messaging relied upon Short Message Service operating on the SS7 signaling network. However, operators are increasingly migrating text messaging to data services while maintaining the same user interface.

[11] The operator must determine if accounting and charging for a transaction or application will have a negative impact on the number of such transactions or applications that it can support.

Preparation for Deployment

The requirement for adequate performance, even during heavy network loads, creates special challenges for the mobile data network: Service providers cannot risk offering revenue-generating services unless they can first assess how effectively their systems will deliver these services and how accurately they will track them during peak network loads.

Today, the mobile data industry lacks the necessary foundation of empirical data describing the performance, scalability, and reliability of GGSNs in the face of exploding traffic growth. Service providers are therefore taking significant – and unnecessary – risks by deploying services in advance of such data.

The three dangers that we have identified here significantly increase the risks associated with GPRS service delivery. Wireless data operators require accurate and scalable benchmarking and validation prior to service provider deployment. Such tests must include controlled and reproducible assessment of performance ceilings, network availability during oversubscription and accounting accuracy.

Looking Forward

The benchmarks described in this paper presume that the operator has implemented policies to ensure adequate performance of the data network. However, if the GPRS network shares the same infrastructure as the voice network, then voice traffic may be given precedence to transit lower-delay, lower-loss network routes and queues than data services (such as GPRS). This puts GPRS at a disadvantage during periods of congestion. Therefore, as push-to-talk and other packet-based voice services grow in popularity and ubiquity, GPRS data services will warrant additional benchmarking in the presence of competing voice-over-packet services.

Today, the recommendations described here will provide a solid foundation for characterizing and designing reliable revenue-generating GPRS networks.