DETERMINING READINESS FOR 2-WAY INTERACTIVE SERVICES

Bob Harrison
Spyglass Integration

Abstract

A team of RF, system integration, quality assurance testing, and software development engineers at Spyglass Integration has recently created a comprehensive suite of testing and measurement tools and methodologies that characterize the downstream and return path bandwidth utilization for different classes of 2-way interactive services such as VOD, tCommerce, and unified messaging. These characterizations can be mapped to an operator's existing broadcast and return system network as a means to identify and mitigate bottlenecks and realize a balanced delivery of services for both steady-state and peak subscriber access.

In this paper, we will introduce these tools and methodologies that may enable operators to determine their current network's suitability for deploying 2-way interactive services, and identify where infrastructure investment or bandwidth allocation modifications may be considered to achieve required quality of service for subscribers.

Background

For interactive TV applications it all comes down to the subscriber's perceived quality of the service. Does the interactive guide fully populate with program information? Is a VOD purchase request properly provisioned, enabled, and billed? Is an acknowledgement for a commerce transaction quickly provided? Do messaging services such as multi-player gaming chat and e-mail provide a responsive user interface?

When the data transport bandwidth for downstream and return systems in cable plants is exploited for emerging interactive services, it is important to know how well the systems which support these services function, perform, and scale. Bandwidth bottlenecks in the network topology need to be anticipated and identified. Application server response, as it is integrated within the network, needs to be stressed and measured. The consumer set-top terminal's ability to receive and transmit messages influences perceived performance.

It is possible to functionally test an interactive service by configuring a test bed consisting of an application server which hosts an interactive service integrated with a digital head-end on an isolated RF network with a representative set top terminal. Anticipated command and control messaging and data flow between the application server and the set top may be observed and analyzed, and a service's functional goals may be validated with respect to an agreed upon service specification. But this functional validation is not sufficient for deployment consideration by a network operator. The operator is concerned with the stability, performance, and scalability of the service functionality as tens or hundreds of thousands of customers subscribe to and use the service. Will the newly introduced service fail? Even worse, will the resources absorbed by the new service break existing and stable revenue

generating services such as core digital video broadcast and PPV?

How can a network operator or a vendor of a new interactive service predict that the new service will not impact current network operations, and secondly, provide a level of service quality that will meet the expectations of all the subscribers who are offered the service?

Current Service Evaluation Practices

Once a vendor of an interactive application has demonstrated that the service meets its specified functionality through a thorough validation or acceptance test plan, operators engage in a phased approach to understand issues of performance, stability, and scalability of the service, as it applies to their unique network environment, without negatively impacting currently deployed services.

Operators have created laboratories that attempt to replicate their operating network so that they may stage the service in a familiar environment. For the first time, the service is integrated in a head-end which maps the component versions, configurations, third party video distribution and data network infrastructure products that represents the operator's deployed operations. Service functionality may be revalidated at this point, but what about performance, stability, and scalability?

Internal "Friendlies"

Possibly ten to a hundred operator employees will be given access to the new service and asked to "give it a try". If the service fails or performs poorly, these non-subscribers (friendly users) will report their observations and impressions in a qualitative way. Rankings on a scale of "1 (poor) to 10 (excellent) " are solicited. These friendlies are not quality assurance specialists performing evaluations based on formal test procedures. They are considered representative of exercising the kind of service interaction that can be expected of subscribers. Are the friendlies all accessing the service at the same time ? Are they accessing all the features offered by the service? Are they examining boundary conditions or service inflection points as a means to examine extreme stress scenarios? Not necessarily.

The goal of this internal friendlies trial process is often to ascertain the stability of the application server and set top client application over a long period of time (weeks to months) and to understand major issues of service stability (does the service crash or become unavailable) to anticipate subscriber acceptance of the service. This level of testing, performed on an isolated network (an internal laboratory head-end) does not predict service performance, stability or existing network integrity as downstream and return path data communication bandwidth by the service approaches the nominal or peak utilization of a subscriber population in a specific property. Nor does it address the load of the application server itself (ability to service transaction requests). However, the level of confidence that the service may one day be considered deployable may be enhanced, because the service is consistent with the configuration and version of deployed network elements.

Bank of Set Tops

Within the laboratory evaluations, operators (with cooperation from their network infrastructure vendors and the interactive application service provider) often attempt stress testing by configuring many set-top boxes in a scripted or automated test harness. Using tools such as TestQuest, Inc.'s

TestQuest Pro that can replay streams of scripted IR commands, monitor the results produced on screen, and provide comparison with reference images, it is possible to repeatedly and deterministically emulate viewer behavior and create a methodology to invoke all service features across a finite number of set tops which have been allocated for the task. Even if hundreds of set tops are provisioned for this process, it still falls short of the subscriber population that will be expected to be supported by an operational head-end system.

Limited Operational Field Trials

Once the internal friendlies evaluation has been performed and (optionally) laboratory stress techniques have been analyzed, the service may become a candidate for a field trial. The operator selects a candidate property, and the service in integrated within an operational head-end. A small subscriber population is selected to evaluate the service. These subscribers are again friendly to the evaluation; it is not expected that they will discontinue service should they experience service disruption or other anomalies. The greatest value of the limited operational field trial is that the service functionality may be validated within an operational network. Again, confidence for total subscriber scalability has not been gained.

A New Approach

When evaluating a 2-way cable plant's suitability to support the introduction of an interactive service, several characteristics need to be studied:

1. The service introduction will not impact the actual or perceived delivery existing deployed services.
2. The server that supports the interactive application service must be shown to scale for the expected subscriber population request load (both nominally and during peak utilization)
3. Bandwidth limitations in the data network (downstream in-band and out-of-band) and return system must be identified so that bottleneck issues may be alleviated by network element upgrade or addition or topology reconfiguration.

A *Meaningful* Load Tester

Raskin and Stoneback suggest, "HFC network performance monitoring is likely to be done most effectively by collecting and coordinating communication performance information from the applications running over the network"[i]. This implies that network performance monitoring needs to be performed in the context of the applications that the network is expected to support, not simply loading a network with variable volumes and frequencies of data payloads. In response to this suggestion, Spyglass Integration created an application load tester and IP network interactivity tester which provides the flexibility to coordinate application oriented communication messaging and collect the relevant statistics with the goal of understanding HFC data network performance in a meaningful context.

Load testing addresses the objective to interject significant packet data in a cable data network in an attempt to load the system with the level of transaction traffic that can be expected by a realistic subscriber population. Load testing can be designed to be a vehicle to provide insight and analysis for throughput for either a single set top or many concurrent set tops. Throughput in this context is defined as the time it takes to receive a response at a set top for each message request or

acknowledgment sent by the set top to an application server.

Throughput analysis comprises a technique and measurement capability to create a *meaningful* request from a set-top to an application server and measure the time for a *meaningful* response to be received by the set-top. The actual interactive service is invoked and satisfied by the application server.

Concurrent throughput analysis is the ability to measure processing speed from multiple set tops (the time it takes to receive a response at a set top for each command request or acknowledgement sent by "N" number of set tops to an application server.

Concurrent throughput analysis comprises a technique and measure capability to create a *meaningful* request from multiple set-tops to an application server and measure the time for a *meaningful* response to be received by the set-tops. The actual interactive service is invoked and satisfied by the application server. The behavior for each set top configured for concurrent throughput analysis must be separately identified and measurable.

Packet Characterization

Message packet characterization comprises the ability to record downstream packet characteristics from an application server communicating with a single set top for the following attributes:

- Size of packets
- Frequency of packets
- Information contained in the packets (content sensitive)

Concurrent packet utilization requires the ability to record downstream packet characteristics for "n" number of set tops for the following attributes:

- Average size of packets
- Average frequency of packets

To capture packet characterizations it is necessary to probe (or sniff) communication between a set top and an application server. A model can be built which represents the size, frequency, and content for a message set between the set top and application server. This model can be used to build script testing scenarios to generate meaningful requests and responses. Scripts that use these packet characterizations may be invoked to create repeatable network load that is representative of true interactive service messaging.

System Loading

In addition to characterizing a specific interactive service under test and evaluation, it is important to interject load that represents the delivery other data or video services. Therefore, a load testing environment must also provide the ability to simulate traffic unrelated to the application service under investigation, yet representative of other network functions (conditional access messaging, program guide data carouseling, PPV purchase polling, etc.).

A Load Tester for Motorola Networks

A load testing simulator had been designed and built by Spyglass Integration to drive application level requests in a Motorola DigiCable environment. The DCT 2000 set top is a proprietary platform that provides fundamental services on a network (UDP message packetization, DAC-6000 communications, and NC-1500 communications) as well as providing the fundamental RF network interfaces with out-of-band modulators (OM) and return path demodulators (RPD). It is possible to create a communication proxy application for the DCT

2000 that enables a PC based application simulator to use the DCT 2000 as a HFC RF gateway. The proxy essentially provides the out-of-band and return path network communication services requested by an application on the PC.

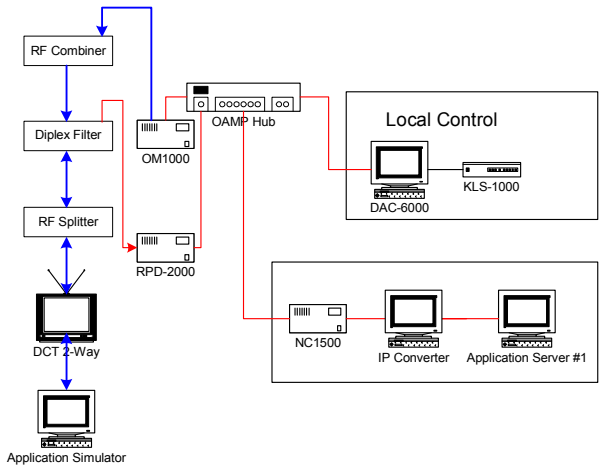


Figure 1

Figure 1 illustrates the integration of a PC based application server with the DCT 2000 through the DCT 2000's serial port. The DCT 2000 can be thought of as a tethered out-of-band and return path modem for the PC. The DCT 2000 is used only as a network modem. There is no need for compute intensive on screen display functions, IR remote interrupt service routine handling, or other data manipulation.

The application simulator may be scripted to create the various request messages a specific interactive application server expects. Commands to the DCT 2000 proxy place the messages on the HFC network for routing to the application server. The scripting is specific to each application server, and generally takes the form of an ordered series of messages as defined by the messaging protocol used by the set top client application developed for the service. For example, if the application server is a VOD server, the messages scripted for the PC application simulator would be the series of VOD commands that can be expected by a native VOD client on the DCT 2000 (session establishment, stream control, etc.).

A single application simulator PC and DCT 2000 proxy can be used to emulate multiple set top sessions if the IP address associated with the DCT 2000 proxy can be altered prior to delivering a message to the application server. This has been accomplished by creating a hardware / software gateway (indicated in figure 1 as the "IP Converter"). By changing the IP address of the packet prior to delivering it to the application server, the server can be spoofed to handle multiple logical sessions with multiple set tops. Through an IP mapping management technique, application server responses can be redirected back to the originating DCT 2000 proxy / application simulator, or another network device to collect the characteristics of the response.

The proxy service for the DCT 2000, the IP converter, and data (packet) collection and analysis utilities resident on the application simulation are created once and are independent of the class of interactive service which is being evaluated. The only variable is the messages and state transition protocol for the interactive service as defined by the vendor of the application server, emerging open standards, or by packet characterization probing.

Although we have implemented this environment with a DCT 2000 in a Motorola DigiCable network, the technique may be applied to virtually any infrastructure provider's network. The prerequisite requirement is for a set top application development environment to expose the network services required for return path and

downstream (in-band and out-of-band) communications, as well as serial or Ethernet communications to a tethered PC.

Scalability of the Load Tester

The load tester can be used to simulate the realistic request / response application messaging of hundreds of set tops from a single set top. If one can characterize the anticipated consumer generated frequency of message requests (often represented in units of requests per minute or requests per hour), the scripting engine within the application simulator PC may be configured to generate the requests expected in nominal and peak utilization times by hundreds of subscribers. The IP converter spoofs the application server to believe that these requests originate from different set tops.

By positioning multiple load testing systems as front ends to independent collections of multiple out-of-band modulators, return path demodulators, and network controller elements in a network, it is possible to stress and identify bottleneck conditions with respect to independent network segments. The capacity, as indicated by set top population, of a out-of-band or return path segment can be measured, with respect to the application service being evaluated and the nominal loading of the delivery of existing data and video services. This is determined by creating scripts for each application emulator that provides realistic application server requests. Data (message packet size for each request and response and server response time) is collected for independent network segment to evaluate the equilibrium point between request and response message cycles. This equilibrium point represents the traffic scenario where the response time delay is deemed to be unacceptable to user's experience. This equilibrium point correlates to the number of set tops that can support a required level of service quality given the network configuration of an independent network segment. It is possible to use this data to develop a model of how set-tops may be distributed among multiple OM / RPD / NC network configurations to empirically achieve the performance goal required for an expected subscriber population.

We have found that the communication proxy application within the DCT 2000 can be extended to provide more than a protocol gateway. Since the DCT 2000's serial port can serve as a bottleneck for requests issued by the application simulator, request messages from the application emulator may be pre-cached in the DCT 2000, sequenced, and issued on a scheduled basis to an application server.

Discounting Application Latency

Sometimes the response latency of the application server is the bottleneck. Degradation of application response times is a function of application server performance, rather than the network. To remove application latency from a network performance characterization study a second tool has been developed, called the "Interactivity Tester".

The Interactivity Tester can be though of as a client message generator and server reflector of known UDP packets between a set top and the physical network location of an application server. Figure 2 illustrates the components of the Interactivity Tester, and its integration in an HFC network.

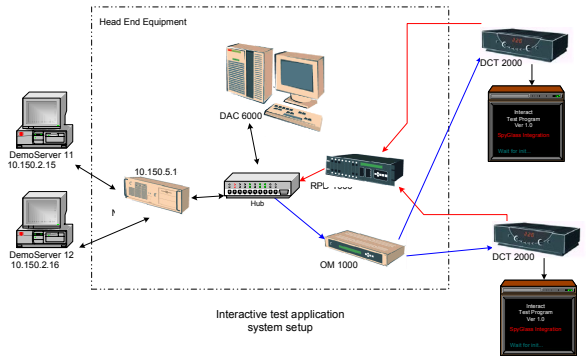


Figure 2

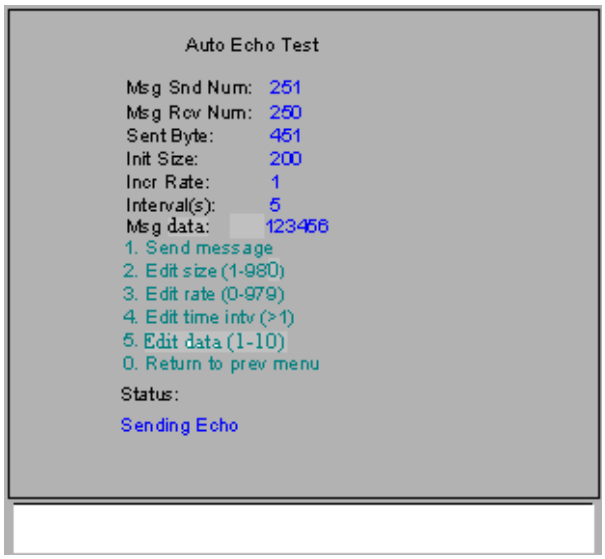


Figure 3

To remove application server latency, a PC-based network echo server (labeled "Demoserver" in figure 2) is positioned at the same network topology location as the application server. This echo server may be flexibly configured to provide UDP packets with message patterns identical to those that are expected to be provided by an application server in response to a request. A client component, installed on the set top, can be configured to initiate requests to one or more echo servers. The messages sent comprise payloads that emulate application message requests, and can be throttled at variable periodic frequencies (within a 1 second resolution). Figure 3 represents the typical on screen display that the set top client agent provides.

The echo server is configured to provide a selected response to the client's request. Figure 4 illustrates the method to build the response message.

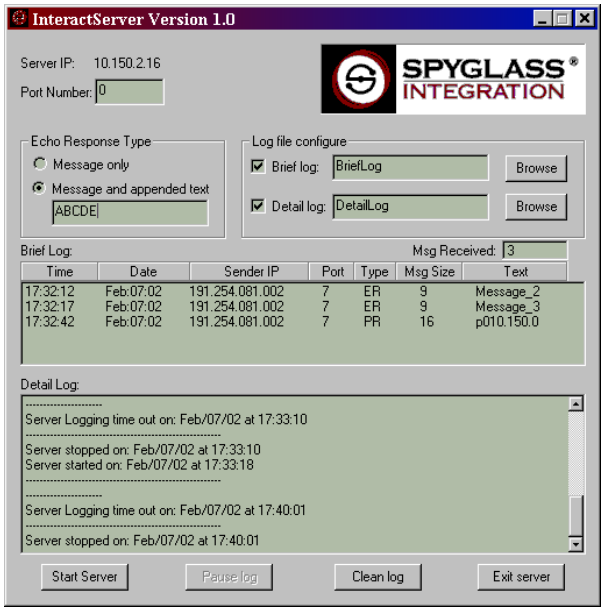
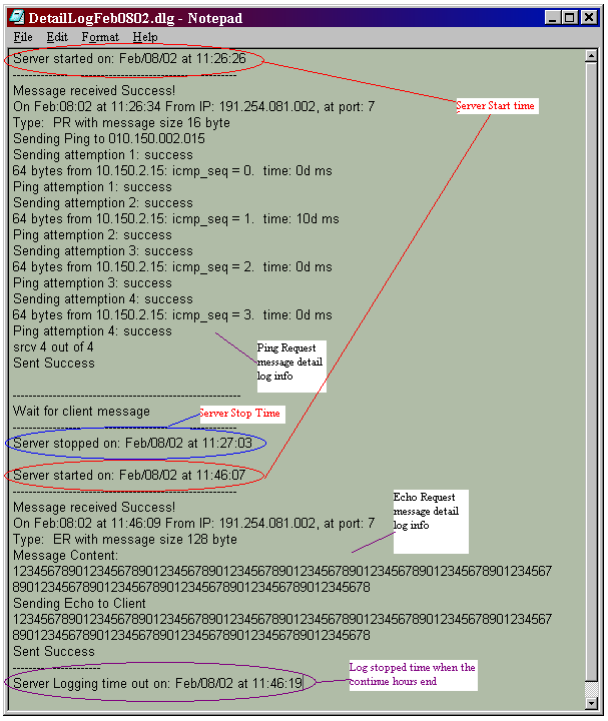


Figure 4

The echo server also provides a detailed logging facility with respect to transactions generated by the set top client, and responses

issued by the echo server. Figure 5 illustrates an example of this detailed log file.



Conclusion

The Load and Interactivity Testers are tools that may be applied to a laboratory HFC network to economically introduce a level of interactive service utilization to evaluate the network's and application service's abilities to scale for a subscriber population representative of a deployment environment. The tools provide the ability to characterize the data packet messaging between a set top and an application server, to script and invoke multiple concurrent sessions, and collect empirical data that represents messaging behavior. By analyzing this data, it is possible to identify either network bandwidth constraints or the capability of an application server to support a specific subscriber load over varying utilization assumptions.

Credits

The following Spyglass Integration team members contributed to the concept, design, and implementation of the Load and Interactivity Testers: Robert Cleroux, Tony Curran, Jim Desmond, Mike Foss, Bob Frankel, Ben Li, Ed MacDonald, George Sarosi, Martin Wahl, and John Zhang.

Bob Harrison is a Solutions Architect with Spyglass Integration located in Lexington, Massachusetts. He may be reached at bharrison@spyglassintegration.com

---

[i] Donald Raskin and Dean Stoneback, *Broadband Return Systems for Hybrid Fiber/Coax Cable TV Networks*, (Upper Saddle River, NJ : Prentice Hall PTR, 1998) p. 241