Jennie Grosslight
Thomas Dippon
April 9, 2002

InfiniBand Compliance and
System Testing

Switch Performance Measurements
This presentation provides a brief overview of the test requirements for an InfiniBand (IB) system, examples of standards compliance and systems test, and new methods of measuring performance using Agilent test solutions.
Supporting InfiniBand requires a huge, dynamic range of technologies—from microwave down to the lowest physical layers, all the way up through many layers of the protocol stack.

At the physical layer it is important to verify signal integrity with jitter and bit error rate measurements. Transmission line measurements are critical for signals moving at 2.5 Gb/s. Important tests as you move up the stack of InfiniBand layers include:

- Functional verification
- Transport tester
- MAD (management datagram) tester
- Packet tester
- Code group tester
Problems can occur in multiple nodes of an InfiniBand fabric or within systems on a fabric.

• Tracing Packets
• Latency Through Components
• Throughput

Broad visibility, cross-correlation, and stimulus diversity are critical for effective system-level validation.

Broad visibility: Visibility across a range of business functions and links within the system is necessary for tracking down system-level problems. You need insight into the behavior of each link at the picosecond level all the way up to the highest packet level. Broad visibility is important for matching up request/response pairs in a protocol, for example.

Cross-correlation: You need the ability to correlate activity in different parts of an InfiniBand system. The symptom of a problem may appear in one part of the system, but the root cause may lie somewhere else. We’ll examine two important kinds of correlation for system-level test.

Spatial correlation: You want to see how an event in one part of a system relates to an event in another part of the system and view both events simultaneously. Spatial correlation allows for latency measurements and tracing traffic through a system. Another form is temporal correlation—that is, how an event at one point in time relates to an event at another point in time. A good example is matching packets.

Stimulus diversity: You can’t model 100% of the real world to cover all possible conditions and corner cases that your system will handle. You need to test your system with real-world stimulus, but you need a way of providing stimulus that is focused, directed, and controllable in order to reproduce and debug problems. Setting up regression tests that have very specific test cases, with very specific stimulus, that exercise the whole system, is a good practice.

In summary, you need to be able to work all across the system, to be able to correlate activities—from picoseconds to packets—and to be able to not only handle the billions of events that occur in the real world, but provide very focused, directed, controllable stimulus to your system.
Now we will discuss several examples of standards compliance and system-level testing.
The example shown here is a surface mount adapter board connected to a length of 1x cable, a 1x to 4x adapter, followed by a length of 4x cable, and another connector.

The InfiniBand specification has rather tight tolerances on transmission line impedance. “Normalizing” or calibrating your time domain reflectometry (TDR) system provides the most accurate measurements. Systems that do not provide for complete calibration, or failure to calibrate, may result in measurement errors.

Vector network analyzers (VNAs) or TDR scopes make accurate and in-depth measurements useful for:

- Finding elusive signal integrity problems
- Extracting superior model parameters
- Removing fixture effects
- Measuring trace or cable length
Example #2: Characterization/Validation of InfiniBand 1x Serializer/DeSerializer Components Using ParBERT

We can characterize InfiniBand serializer and deserializer (SERDES) components by measuring bit error ratios with user-defined data, pseudo-random bit sequence/pseudo-random word sequence (PRBS/PRWS), or mixed data.
The Compliance and Interoperability Working Group (CIWG) of the InfiniBand Trade Association (IBTA) is responsible for compliance testing specifications. Agilent has implemented CIWG test descriptions in its E2950 Series test software to provide a push button solution for compliance tests. Several test scripts are included with its exerciser software. Additionally, with the Agilent InfiniBand exerciser, users can customize their testing environment with the C++ or TCL scripting language. The compliance test software is available as InfiniBand exerciser option #100.

An easy-to-use graphical user interface (GUI) for the Agilent E2950 Series test system as well as a COM-Interface are specialized for InfiniBand real-time performance measurements. Other features include:

- Functional validation
- Compliance testing
- Error injection (deterministic generation of error conditions)
- Stress testing
- Regression testing of corner cases
- Application/vendor specific testing
- Emulation of host channel adapters (HCAs) / target channel adapters (TCAs) / switches with fully programmable characteristics

Agilent logic analyzers and protocol analyzers can help you examine areas of interest in detail. The logic analyzers are optimized for cross-bus measurements and provide a consistent way of looking at data, no matter what bus you’re on.

The protocol analyzers offer a hierarchical view of protocols. You can click on specific packets and pieces of packets and expand them for more detail.
Example #3: Compliance and System Testing (Cont.)

- Set sequential triggering on logic analyzer
  - TS1 followed by TS2 followed by Link packet
- Expand triggers or leave at high level
- Set exerciser to initiate TS

This is an example of how you might set up an InfiniBand trigger on an Agilent 16700 Series logic analysis system. Sequential triggering helps to narrow in on specific events.

In this example, the trigger is set for a comma, followed by a training sequence, followed by a link packet. The exerciser initiates the training sequence. The logic analysis system triggers on the first link packet after the training sequence.

Trigger descriptions can be expanded to the bit level or viewed at a high level of user-definable names, such as “Event #1” or “Link packet”. Event # is the default value.

This way of setting up triggers and looking at results is very similar to the way that it traditionally has been done with microprocessors and other kinds of buses in a system. The logic analysis system is optimized for cross-bus measurements and provides a consistent way of looking at data across buses.
Example #3: Compliance and System Testing (Cont.)

- View packet decode, 10B, 8B, etc.
- Filter idles (optional)
- Time-correlate to compliance tool listing, other system buses, or multiple InfiniBand nodes

The logic analysis system provides a view of the InfiniBand packet decode along with character codes for 10B, 8B, etc.

The logic analysis system has triggered on the first link packet after the training sequence. You can easily place global markers to move between a listing on a compliance tool and the InfiniBand packet decode, as shown. Global markers also track across all buses or InfiniBand nodes connected to the logic analysis system.

If the InfiniBand link had failed to come up and the logic analysis system had not triggered, we would reset the trigger to look for either the first TS2 or perhaps TS1. In this manner you could determine how far the system had progressed in the link power-up sequence. Traffic between the requestor and responder would point you to the defective node.
Compliance tests on the logic analyzer include:

- Link initialization and training (InfiniBand Specification V2, r1.0.a, page 97)
- Clock tolerance compensation (InfiniBand Specification V2, r1.0.a, page 96)
- Flow control packets (InfiniBand Specification V2, r1.0.a, page 183)

In this example:

- No complete TS1 groups were found in the trace (rather shallow 32k trace)
- One full TS2 group with 17 sets was found
- There were 29 compliant skip ordered sets found
- There were 2 compliant (Init) link packets found
This diagram shows Agilent test instruments for complete system test and validation attached to multiple nodes in the InfiniBand fabric, a central processing unit (CPU) front-side bus, and a peripheral component interconnect (PCI) bus. We can probe different channels and buses to find problems and trace them to root causes, providing visibility into the different parts of the system.

Signal tracing of a faulty direct memory access (DMA) transfer is a straightforward example of system testing. The easiest way to debug a problem like this is to follow the data through the system by moving the scope probe from point to point through the circuit. We set up an analyzer to look at the InfiniBand traffic, trigger on the remote direct memory access (RDMA) operation, and follow the data as it goes through the PCI bus into the memory system. We also take a look at the CPU-to-chipset communication because that's probably how the memory management units (MMUs) are configured and programmed. Remember, InfiniBand DMA transfers are located in the virtual memory space, so MMUs are involved.

Using a traffic generator, shown here, we can generate the RDMA request and tell the traffic generator to use a very specific data pattern in the DMA buffer to make it easy to follow the signature as it flows through the system.

We can trigger the 16700 Series logic analysis system on the DMA request and trigger the PCI and memory analyzers on the data pattern. If the request is making it out to the PCI bus or the memory bus, we'll find it and trigger. We can see where and when the request went. We also can look at the CPU bus for some insight into how the MMU is programmed.
Examples of InfiniBand 4x triggering, viewing, packet decode and compliance testing are similar to 1x examples, except that Lane deskew is unique to 4x.

The InfiniBand specification provides systematic support for compliance testing. Distributed throughout the specification are statements that highlight key conditions all products must satisfy to comply. These statements often stand independently, but they also can refer to other parts of the specification that contain details supporting the compliance. The following is an example taken from Volume 1, Release 1.0.a of the InfiniBand specification:

“C7-53: Flow control packets shall be sent for each VL, except VL15, upon entering the LinkInitialize state. When in the PortStates LinkInitialize, LinkArm or LinkActive, a flow control packet for a given virtual lane shall be transmitted prior to the passing of 65,536 symbol times since the last time a flow control packet for the given virtual lane was transmitted.” ©IBTA

Each compliance statement is numbered for reference, and all statements are summarized into tables that appear at the end of each volume of the specification. Associated with each compliance statement are a set of test assertions, which are defined as atomic statements resulting from the decomposition of a compliance statement into individual components that can be shown to be either true or false by running a test or set of tests. For example, an assertion associated with statement C7-53 above could be stated as: “Flow Packet transmitted at least every 65536 symbol times.”

From the assertions, a set of actual tests is defined by the InfiniBand specification. Each test may demonstrate the truth or falsehood of one or more assertions. A test identifies the assertion(s) it covers, the test equipment setup, and the test procedure. This information then can be used as a springboard for developing instrument control and analysis commands that implement a specific test.
Agenda

- Overview of InfiniBand Test Requirements
- Compliance and System Tests
- Performance Measurements
- Summary
Switch Performance Measurements Overview

- The performance parameters of switches are key to overall performance of fabric topologies such as InfiniBand Architecture.
- Due to the number of input factors and the complexity of switch cores, it is hard to predict actual performance under real-life load conditions.
- Therefore, measurement of switch performance parameters is critical for manufacturers to verify simulated results under real-life and worst-case load conditions.
- End-users can benefit from measurements to compare various switch implementations and verify claims of manufacturers.

This slide gives a couple of reasons why switch performance measurements are important for fabric topologies. Performance measurements can be interesting for both designers of switch chips, switch products and systems as well as end-users of the technology.
Switch Performance Measurements Overview (Continued)

- The goals of this presentation are:
  - To show a method for measuring InfiniBand switch performance
  - To show how well InfiniBand switches work - even in this early stage of the technology
  - NOT a competitive analysis

Summary of presentation goals.
A definition of terms that are used throughout the next couple of slides.
Test Method (1)

- InfiniBand exercisers are connected to the ports of a switch.
- All exercisers have precisely synchronized time stamp clocks (± few nanoseconds).
- Exercisers generate "instrumented packets" by inserting a time stamp in the payload of each packet just before transmission.
- Upon reception, the difference between the receiver's clock and the time stamp value in the packet indicates the latency.
- In addition, counters keep track of the number of transmitted/received bytes/packets.

With multiple InfiniBand exercisers, it is possible to conduct switch performance measurements under deterministic load conditions, using various packet sizes, VL, and so on. The basic idea behind measuring latency through a switch is to transmit so called "instrumented packets" from one port to another. Instrumented packets have a time stamp in their payload which is generated by the exerciser that transmits the packet and evaluated by the exerciser that receives a packet. With multiple exercisers generating traffic on multiple ports, latency can be measured under stress conditions.
The concept of multiple "streams" allows performance measurements not only for a single packet traversing an otherwise idle switch, but to measure the switch's behavior in congestion scenarios, or when multiple ports are in use. Another important aspect is the dependency of performance and the type of "traffic" going through the switch. Are there many small packets or fewer large packets? What is the inter-packet delay? Does the switch have to perform P-Key checking. Is the linear or random forwarding table used?
This slide shows a series of latency measurements on two existing InfiniBand 1x switches. In this measurement, only two ports of the switch were connected to InfiniBand exercisers and traffic was sent from one exerciser to the other (single stream).

The results are quite different for the two switches in the test. Whereas switch "A" on the left shows a constant latency independent of the load offered on the input, switch "B" shows a significant increase in latency above ~90% load, especially with small packets. On the other hand, below 90% load the performance of switch "B" is extremely good—around 300 ns independent of the packet size. Switch "A", on the other hand, shows an increased latency with larger packets. This is due to the store & forward behavior of switch "A". The first byte of the packet is only transmitted after the last byte has been received.
Early testing shows excellent results up to ~70…80% load
Throughput limitations are typically seen with small packets
Switch Performance Testing - Summary

- Even 1st generation InfiniBand switches show very impressive performance numbers
  - 300 ns cut-through latency!
- It is important to understand the performance characteristics and limitations of switches
  - Even early switches perform excellently up to ~70…80% load
- Many other scenarios (multiple streams, multiple partitions, P_Key checking, SL to VL mapping, weighted VL priorities etc.) have to be taken into account to get a realistic picture
- InfiniBand 4x silicon and switches are now available - stay tuned for great performance numbers
Agenda

- Overview of InfiniBand Test Requirements
- Compliance and System Tests
- Performance Measurements
- Summary
Agilent provides the most complete set of solutions for testing InfiniBand devices and systems (fabrics). Testing can be done at the physical layer and at every layer of the InfiniBand protocol.

Here are examples of tools that can help you at various levels. At the physical layer TDR, scopes, bit error rate testers, VNAs, and spectrum analyzers provide insight.

Protocol analyzers work at the 10B level and go up through the networking, transport, and application levels.

Logic analysis systems straddle the middle. They start working at the 10B level and then move up into the networking and transport layers. The strength of the logic analysis system is in the ability to perform cross-bus analysis. As you move higher up you really want to transition to protocol analyzers as the tool of choice.
Additional information on the Agilent test tools for InfiniBand is available. The brochure “Test Tools for InfiniBand” provides an overview of the Agilent test solutions. For more detailed information, refer to the available data sheets. The current E2950 software with the online manuals are available on our WebPages. The protocol analyzer software includes InfiniBand traffic samples to help you evaluate the product.