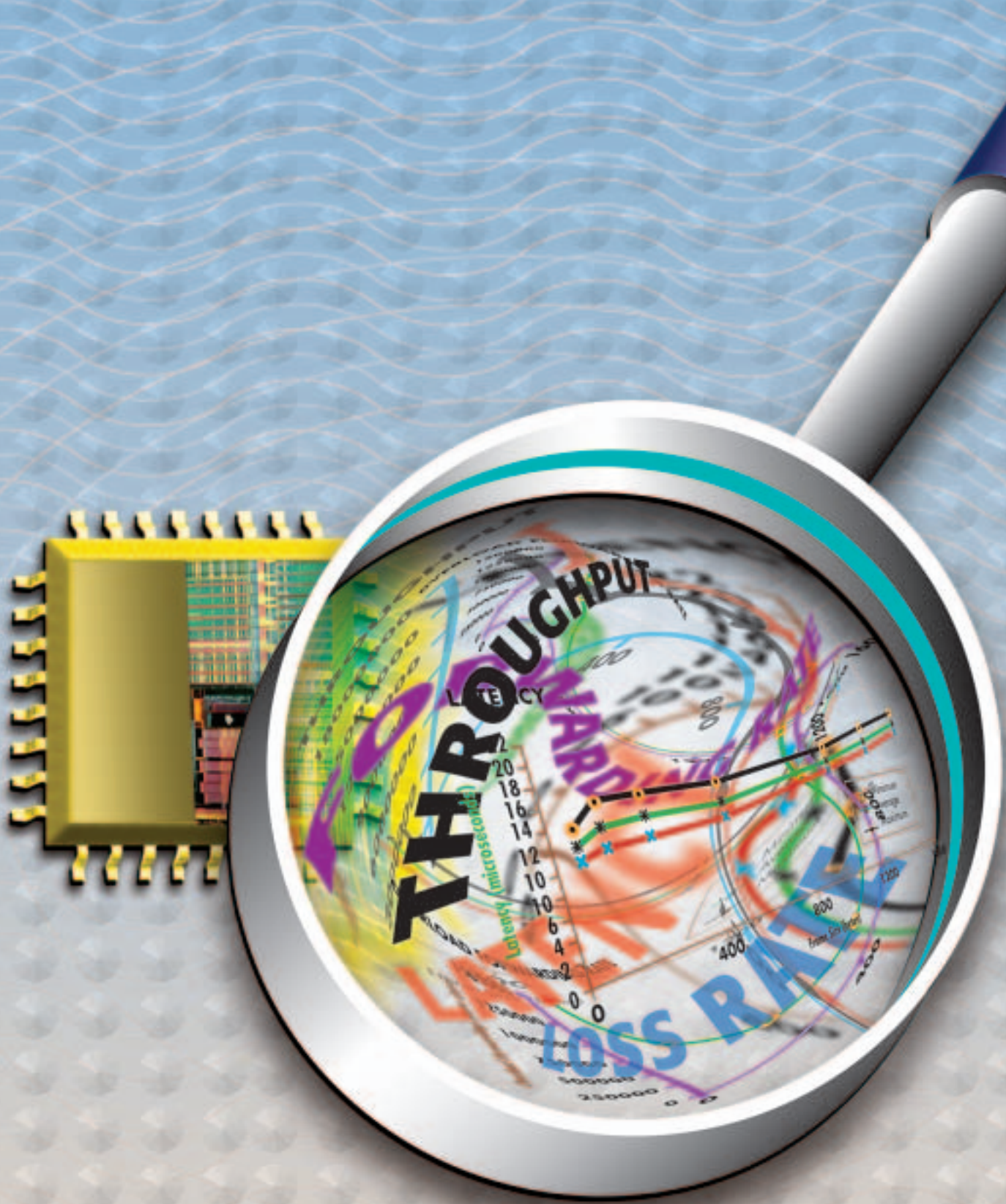


ILLUSTRATION AND DESIGN BY MIRA RAMJI-STEIN

MPLS BENCHMARKS DEFINE NET PROCESSOR PERFORMANCE

BY GANESH BALAKRISHNAN AND RAVI GUNTURI

Network processor vendors are lining up to claim performance pre-eminence when running the MPLS forwarding scheme. Fortunately, the Network Processing Forum's new "black box" MPLS benchmarks are already in place to level the playing field and return the real-world processor performance data that designers need.

The deployment of multiprotocol label switching (MPLS) by carriers and service providers and within the enterprise is rapidly increasing, driven by the ability of the protocol to speed network traffic, manage traffic flows and provide quality of service. With this increasing deployment has come a call from networking equipment manufacturers and network-processor product vendors for a means of impartially and accurately char-

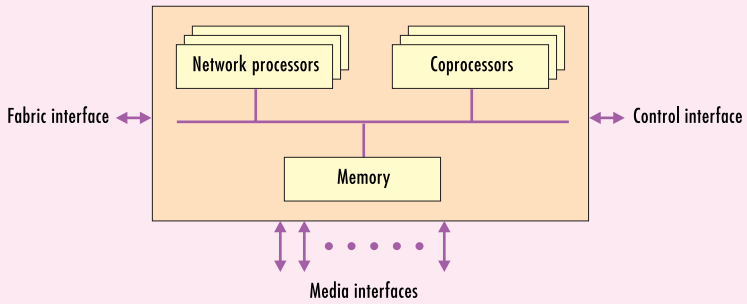


Fig. 1 Benchmarking starts with a reference design identifying a specific device under test. The design may consist of one or more media interfaces, a fabric interface, as well as multiple network processors and coprocessors.

acterizing MPLS performance on the myriad network-processor-based systems that claim to be the system of choice for running MPLS.

The Network Processing Forum (NPF) has stepped up this call with the MPLS Forwarding Application Level Benchmark Specification Implementation Agreement (IA). This is an industrywide open specification that takes a "black-box" approach to achieve impartiality when measuring the MPLS data-plane performance of the various network processor systems. Focused on performance and not protocol conformance, the specification outlines the necessary requirements, tests, testing parameters and reporting formats and includes a detailed implementation kit with readily portable scripts. The specification also incorporates several key implementation methods designed to minimize skewing of results.

The black box surrounds the network processor subsystem, enabling network-processing elements to be evaluated solely on observed input/output behavior. The approach enables industrywide standard performance benchmarks and gives vendors the ability to make their own design choices. But all of the relevant subsystem information inside the black box, such as the list of hardware components, total power consumption and mechanical size, must be provided with the test results.

The benchmarks focus on testing the performance of fundamental functions, as opposed to protocol conformance. A basic level of conformance is assumed, meaning vendors do not have to implement every feature in a protocol specification in order to use the NPF benchmark tests. The benchmarks are open standards developed in a consensus process that invites input from all of the NPF's member companies. All of the forum's benchmark specifications are available free on the forum's Web site at *www.npforum.org*.

The MPLS Benchmark IA itself is an application-level benchmark targeted at network box vendors that will use network-processing vendor-provided application functions out of the box for value-added data-plane functionality. This is a hardware specification; no stipulations are made about what software is used on the system. Once the software is loaded, however, it must be used throughout the duration of the test. This prevents software changes that may be more favorable in one testing scenario than another.

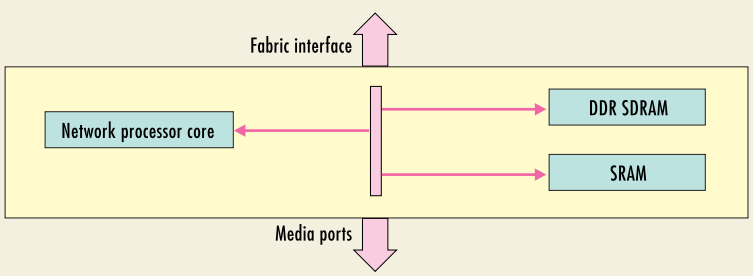


Fig. 2 The NPF provides a complete template for reporting the MPLS Forwarding Application Level Benchmark results. The template contains reference design details as well as ingress, transit and egress traffic details.

Benchmark outline

MPLS domains contain three distinct forwarding points: ingress, transit and egress (for more on MPLS, see sidebar titled "*MPLS raises bar for network speed and management*," page 20).

Each of these functional areas has its own unique processing tasks that can be implemented differently. Therefore, in order to get a complete picture of MPLS performance, the MPLS benchmark tests cover each area independently. Each has its own set of parameters, tests and reporting formats. This comprehensive approach minimizes the ability of participants to gear systems to one particular aspect of the MPLS specification.

The MPLS benchmark specifications are based on the existing benchmark methodology found in RFC2544 and the NPF Internet Protocol version 4 (IPv4) Forwarding Level Benchmark Implementation Agreement. The tests have been tailored to address the unique aspects of testing MPLS technology on network processor systems.

The MPLS IA is a performance-testing benchmark and not a verification of MPLS protocol conformance. Only the basic aspects of MPLS are tested in this implementation agreement to determine the relative performance of MPLS in network-processing systems. The MPLS benchmark specification does not exhaustively test every possible feature or parameter. In situations where multiple options exist that do not have significantly different performance impacts, the most widely implemented method was chosen as the performance metric for that group of functions.

For example, FEC implementations could be network prefix classifier, IP host address classifier or 5-6 tuple classifier. The longest-prefix-match (LPM) or network prefix classifier was selected for the benchmark because it is the most popular

classifier in MPLS today.

Another important attribute of the MPLS benchmark spec is that it addresses only data-plane functionality. MPLS control-plane updates do not stress the system enough to affect data-plane forwarding performance. Thus the result of a control-plane benchmark test is not that useful to a vendor, and tests to measure control-plane performance were not included in the benchmark. Instead, an optional test to measure the maximum number of label switch paths (LSPs) that can be supported by an MPLS router at the throughput rate was added. This measure is useful to vendors evaluating an MPLS router for deployment in traffic engineering and flow management network solutions.

Test specifics

The first step in testing a network-processing subsystem using the MPLS benchmark implementation agreement is to create a reference design identifying a specific device under test, or DUT (see Fig. 1). This test design may consist of one or more media interfaces and a fabric interface. It may include multiple network processors and any number of coprocessors connected to the network processor in any way. The choice of speed and media type is left to network processor vendors or customers comparing different network processors.

The reference design must detail the significant components of the DUT including a block diagram, component list and other elements such as mechanical size, media and fabric interfaces, ports, processor and coprocessor details, and memory.

The MPLS test setup utilizes a data-plane network-traffic generator. The traffic generator must be able to send and receive MPLS traffic with different label values and varying label stacks. It should also be able to calculate the number of packets transmitted and received and should have the ability to measure the end-to-end latency through the DUT. A reference implementation kit is available to run the NPF MPLS benchmark. The implementation kit contains a mandatory Tcl script to generate IPv4 traffic for one of the specific test configurations. The rest of the implementation kit is not mandatory but can serve as a good starting point for vendors that wish to implement the benchmark. All the MPLS and IPv4 forwarding tables for all tests, which maintain the state necessary to forward packets, must be loaded before the tests are conducted.

The benchmark tests come in three basic configurations—ingress, transit and egress—corresponding with the three

> The MPLS benchmarks address only data-plane functionality and test ingress, transit and egress.

scenarios of operation in an MPLS domain. The key tests include forwarding rate, throughput, latency and loss rate. The tests and methodology are detailed in the MPLS benchmark specification and follow the tests and methodology in RFC 2544 and RFC 1242. Only the salient differences in tests, test setup and test parameters are detailed for each test configuration in this article. The ingress and transit configurations add an optional test to measure maximum LSPs supported at the throughput rate.

The forwarding rate measures the rate at which the system can process packets at full line rate. A specific number of IPv4 packets are sent to the DUT at line-rate speed, and a count is made of the number of frames received.

Throughput measures the maximum rate at which packets can be forwarded

> Benchmark test parameters include frame size, routing tables, traffic patterns and label stack depth.

by the DUT without packets being dropped. This measurement is achieved by sending a specific number of frames at a specific rate on all of the media interfaces to the DUT.

Latency is a measure of packet delay as a percentage of the throughput rate. Measuring packet-forwarding latency requires a traffic generator that can uniquely tag and identify packets sent to and from the tester, and recording the time when the packet was transmitted and received. The receive and transmit time stamps are used to calculate latencies for the packet.

The benchmark specification requires latencies reported for 90 percent and 100 percent of the throughput rate. Those rates were chosen because most DUTs experience higher latencies at forwarding rates close to the throughput rate when the system is stressed to maximum.

The loss rate measures packet loss as the traffic rate exceeds the throughput rate. This test evaluates packet loss over the throughput rate until line rate is reached. The loss rate test is conducted in a manner similar to the test for throughput.

An optional test to measure the maximum number of label switch paths supported at the throughput rate was also added to allow vendors to report the maximum number of flows that could be managed without affecting throughput. A minimum of 200 LSPs must already be exercised by the DUT in the course of performing MPLS operations and forwarding packets.

This test is measured by sending frames through the DUT at the rate determined by the throughput test. The number of LSPs in the next-hop label-forwarding entry (NHLFE) is progressively increased in fixed increments uniformly across all interfaces until frames are no longer being forwarded at the throughput rate. The test requires the data-plane MPLS and IPv4 tables to be populated between iterations. Therefore, the method to add LSPs must be synchronized with the procedure to send traffic to conduct the tests. But no control and data-plane interaction is required for this test.

Configuration parameters

Each configuration has its own set of test parameters and results, but the parameters generally include frame size, routing tables, traffic patterns and label stack depth and operation. Each configuration also has some unique setup parameters.

A subset of the RFC 2544 frame sizes is used. The size of a packet is selected to be the maximum packet size so that any corresponding link-level frame size never exceeds 1,518 bytes when a frame is going into or coming out of the DUT. That prevents packet fragmentation from affecting the packet-forwarding performance during the tests.

Studies predicting future deployment of MPLS suggest that this technology will be used widely in BGP-based VPNs, which require a label stack depth of three. The maximum label stack depth for the benchmark was chosen with these future deployment scenarios in mind, so all IPv4 routing tables, MPLS forwarding tables and traffic patterns must be set up to support label stacks of depth three in each configuration, if possible. Not all parts will be able to support all the tests, however, so vendors will have to enter a zero value for tests they cannot conduct.

The test setup for each configuration is very similar. The appropriate MPLS and IPv4 forwarding tables must be loaded before testing. The traffic tester must be configured to run the benchmark scripts and must be connected to send traffic to the relevant ports on the DUT. The traffic patterns specified for each configuration should be tested. Both the exact set of ports and the test setup will vary depending on the DUT.

Ingress configuration

The salient parameters for each configuration are outlined below.

• Frame sizes: The ingress configuration pushes one or more labels in the packet's label stack. The benchmark specification uses a number of IPv4 packet sizes, specified in RFC 2544, which range from 40 to 1,500 bytes. The link layer frame sizes are computed from the IPv4 packet sizes based on the media interface. Fixed-sized streams and mixed streams are used. The mixed stream sizes are based on real-world analysis of IP packet size distributions.

• Routing table: The ingress configuration uses a static routing table based on a snapshot of the Mae West routing table. The snapshot must be used to perform all ingress traffic tests for all of the label stack operations. The entire Mae West snapshot route table must be loaded into the table of entries for the classifier in the DUT. A longest-prefix-match classifier will classify each packet based on the destination IPv4 address of the packet, obtained from the snapshot of the Mae West table. A mandatory script is used to select the Mae West entries for the data traffic to hit. That script, in combination with a large number of LSPs and a large

MPLS raises bar for network speed and management

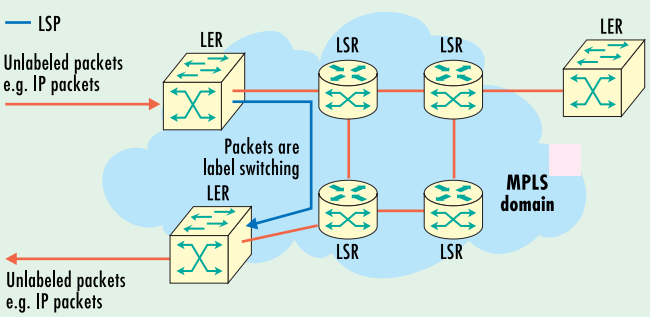
Multiprotocol label switching (MPLS) is a forwarding algorithm that utilizes labels and label switching to significantly enhance the speed and manageability of network traffic. By integrating a label-swapping forwarding paradigm with network-layer routing, an MPLS domain makes faster forwarding decisions than traditional IPv4 forwarding. This improves the performance of network-layer routing, increases the scalability of the network layer and facilitates traffic engineering through an IP network. MPLS integrates the key features of both Layers 2 and 3, but it is not limited to any Layer 2 or 3 protocol. It can be extended across multiple product segments.

An MPLS domain (see figure) consists of two or more label-edge routers (LERs) connected by multiple label-switched routers (LSRs). Label-switch paths (LSPs) are set up between ingress and egress LER pairs to transfer packets across the MPLS domain. As such, an LSP consists of an ingress LER, one or more LSRs and an egress LER. Multiple LSPs can be set up between any ingress and egress LER pairs. Packets that traverse an MPLS domain undergo varying degrees of processing, depending upon the location in the LSP where the packet is being processed.

For every packet entering the MPLS domain, the ingress LER determines whether it should enter a particular LSP and then pushes one or more labels in the packet's label stack. LSRs swap existing labels in the label stack, or swap and push one or more new labels on the packet's label stack. Egress LERs are responsible for terminating one or more LSP by popping the corresponding label from the packet's label stack. If an egress LER pops all the labels, the original packet (for example, the IP packet) is obtained. If multiple MPLS domains are nested, a packet's label stack contains a label for each nested MPLS domain.

Control protocols such as LDP, CR-LDP or RSVP-TE are used to establish, maintain and terminate LSPs in an MPLS domain. These control protocols allocate, distribute, assign, release and withdraw labels used to realize the LSPs. The control protocols also provision the establishment of constraint-based traffic-engineered LSPs called CR-LSPs. The constraints could be resource or path requirements through the MPLS domain. CR-LDP is a variant of LDP that is used to establish the CR-LSPs. In order to create LSPs, ingress LERs, LSRs and egress LERs contain tables that need to be populated with control information related to the LSPs being created.

➥ *Continued on page 35*



An MPLS domain consists of two or more Label Edge Routers (LER) connected by multiple Label Switched Routers (LSR). Label Switch Paths (LSP) are set up to allow MPLS transfers.

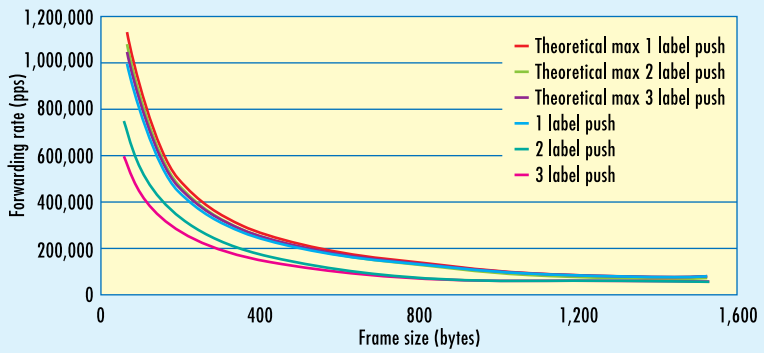


Fig. 3 A sample forwarding rate graph shows what is required of benchmarking participants. Systems that cannot complete all of the tests must have a zero entered as the results of those aforementioned tests.

routing table, prevents any manipulation, hard coding or caching of the routing data that could impact the test results.

• Traffic patterns: The traffic patterns are different for each configuration. In the ingress configuration, the tester must generate traffic on the ports with IPv4 destination addresses drawn randomly from the Mae West snapshot routing table. All of the traffic should be uniformly distributed over the ports. All frames must consist of IPv4 datagrams with no options.

• Label stack depth, label stack operation and FEC types: In the ingress configuration, the label stack depth of an incoming packet should be zero. The MPLS forwarding operation will push 1, 2 and 3 labels. An IPv4 LPM classifier must be used. The classifier is an essential component on an MPLS forwarding pipeline and cannot be removed from the system while measuring performance.

Transit configuration

The transmit configuration parameters are as follows:

• Frame sizes: The same frame sizes will be used in the transit tests as were used in the ingress tests. Since these frames will also have an MPLS label, the IPv4 datagram size is shrunk accordingly to accommodate the entire frame within 1,518 bytes.

• Traffic patterns: In the transit configuration, the tester should generate traffic on all the media ports with labels drawn randomly from the labels assigned in the incoming label map (ILM). The traffic should be uniformly distributed over all the ports. In this scenario, the individual tests start with all frames consisting of labeled traffic with a single label. The traffic must hit the NHLFE entries that swap a single label. In the next iteration, the traffic must hit the NHLFE entries that swap a label and push an additional label onto the label stack. In the final iteration, two additional labels must be pushed onto the label stack.

• Label stack depth, label stack operation and FEC types: In the transit configuration, the incoming packet label stack depth should be one. The Swap and Swap-Push MPLS forwarding operations will manipulate the incoming label stack.

Egress configuration

To perform the egress tests, the DUT should be set up in Penultimate-Hop Pop (PHP) mode. Most MPLS domains will be set up in this fashion. Further, setting up the DUT in this mode allows measurement of the performance of the MPLS point-of-presence operation, without being masked by an IPv4 address lookup to determine the next hop. If the PHP is not supported, performance of these tests may be reported in the regular mode of operation.

• Frame sizes: The frame sizes and route tables remain the same for these tests as for the other tests.

• Traffic patterns: The tester should generate traffic on the ingress ports with labels drawn randomly from the labels in the ILM. The traffic should be uniformly distributed over the ports. The first traffic pattern sends frames with a single label. The next pattern includes two labels and the third pattern, three.

• Label stack depth, label stack operation and FEC types: In the egress configuration, packets sent to the DUT must have a label stack depth of one, two or three, depending on the label stack operation. A single point-of-presence operation must be done on all stack labels.

Reporting, certification

Specific reporting formats for the individual tests within each configuration are explicitly stated in the MPLS Forwarding Level Benchmark IA. Note that systems that cannot complete all of the tests are not excluded from the benchmark. They must simply enter a zero value for the results of the tests in which they cannot participate.

In addition to these instructions, the NPF provides a complete template for reporting the MPLS Forwarding Application Level Benchmark results. The template contains examples of the reference design details, such as the block diagram in Fig. 2. The template also shows examples of the graphed test results for the ingress, transit and egress traffic tests. A sample forwarding rate graph is shown in Fig. 3.

Network-processing manufacturers that wish to certify the performance results of their benchmark tests must submit their products to a third-party independent auditor and certification authority such as the Tolly Group. Once testing is completed, the NPF provides the "NPF Certified" mark to the manufacturer to validate that the benchmark results are in complete compliance. ■

For more on nework processors, see:

"Steering Your Way Through Net Processor Architectures"; *www.commsdesign.com/story/OEG20020724S0079*

Ganesh Balakrishnan (ganeshb@us.ibm.com) is an engineer with IBM Corp. He has a master's degree in Electrical and Computer Engineering from Purdue University.

Ravi Gunturi (ravi.gunturi@intel.com) is a network software engineer at Intel Corp. He has BSEE and MSEE degrees from the University of California at Berkeley.

The authors are co-editors of the NPF MPLS benchmark specification at www.npforum.org.

Kit aids in development of Virtex-II Pro systems

A development kit for networking-equipment designers includes the Xilinx Virtex-II Pro FPGA, 128 Mbytes to 1 Gbyte of DDR SDRAM, 32 Mbytes of SDRAM, 2 Mbytes of SRAM and 16 Mbytes of StrataFlash memory. The platform also comes with an embedded Linux OS, pads for the XPak module, two HSSDC2 connectors, receptacles for two SFP modules, a 32-bit PMC bus and four 140-pin GPIO expansion connectors. Designers can use this board to develop IP for packet processing, security processing or other networking designs. Available now, the kit is priced starting at $1,995. Avnet, Phoenix; *www.avnetavenue.com*

Sonet/SDH receiver can detect down to –23 dBm

Operating at up to 3.125 Gbits/second, the ZL60011 receiver is aimed at 2.5-Gbit/s Sonet/SDH equipment designs. The receiver comes with a 1,310-nanometer InGaP PIN photodiode and can detect signals as low as –23 dBm. The optical receiver also includes a transimpedance amplifier with integrated limiting amplifier and a photocurrent monitor.

Available now, the device is housed in a five-pin TO-46 package and priced at $20 each per 1,000. Zarlink Semiconductor, *www.zarlink.com*

High-speed D/A delivers 72.5-dB signal-to-noise

The LTC1743 is a 50-Msample/second, 12-bit digital-to-analog converter that delivers a 72.5-dB signal-to-noise ratio, making it a fit for direct-IF applications. The device offers a ±1- or ±1.6-volt input. It delivers a 71-dB SNR and 90-dB SFDR when working from a ±1-V input and a 72.5-dB SNR and 85-dB SFDR when working with a ±1.6-V input. A separate digital-output supply pin allows easy connection to DSPs and FIFOs. Available now, the D/A converter comes in a 48-pin TSSOP priced at $9.30 each per 1,000. Linear Technology, Milpitas, Calif.; *www.linear.com*

Amplifiers for mobile phones take little current

The LMV321 (single), LMV358 (dual) and LMV324 (quad) amplifiers consume 120 microamps off a 2.7-volt supply and 100 microamps off 5 V. Targeting mobile devices, the amplifiers have a 2.5- to 5-V supply range, a 1.4-MHz gain bandwidth at 5 V and a 1.5-V/microsecond slew rate at 5 V. Available now, the devices are priced at 18 cents, 24 cents and 29 cents each per 10,000, respectively. Fairchild Semiconductor, San Jose, Calif.; *www.fairchildsemi.com*

➥ *Continued from page 20*

Each ingress LER contains a forward equivalency class (FEC) to next-hop label-forwarding entry (NHLFE-to-FTN) table. An FEC is used by the ingress LER to direct packets onto the appropriate LSP. Some typical FEC implementations include network prefix classifier, IP host address classifier or 5-6 tuple classifier. Based on this classification, the ingress LER determines the NHLFE to use. The NHLFE determines the packet's next hop and the label operation to perform, thereby placing it in an LSP. LSRs and egress LERs also contain an incoming-label-map table, to map the topmost label on an incoming packet's label stack to an NHLFE. The NHLFE contains the operation(s) that must be performed on the packet's label stack. A single MPLS node will usually function as both an egress LER and an LSR.

Current Layer 2 and Layer 3 MPLS implementations may have their own unique MPLS requirements, such as maximum label-stack depth and number of LSPs supported. These requirements vary widely depending upon the network application. ■