The Hybrid Optical and Packet Infrastructure (HOPI) Testbed

The HOPI Design Team

Table of Contents

Introduction

When Internet2 was first organized in October 1996, one defining mission was to provide scalable, sustainable, high-performance networking in support of the research universities within the United States. The resulting infrastructure, comprised of campus, regional, and national components, is a successful, robust, shared packet switched network. In the next few years, however, it must evolve to take advantage of new infrastructures. The HOPI testbed will examine how those infrastructures can be molded into future network architectures.

In planning for the Internet2 networking architecture needed beginning in 2006, a hybrid consisting of shared IP packet switching and dynamically provisioned optical waves is under consideration. The term HOPI is used to denote both the effort to plan this future hybrid and a testbed facility that will be built to test various aspects of potential hybrid designs. The goal of that facility is:

> **The HOPI testbed is to provide a facility for experimenting with future network infrastructures leading to the next generation Internet2 architecture.**

The future architecture will require a rich set of wide-area waves together with switches capable of very high capacity and dynamic provisioning, all at the national backbone level. Similarly, regional initiatives are under way to encourage the creation of RONs (regional optical networks) by the same consortia that currently operate the gigaPoPs that connect campuses to Abilene. Finally, the planned hybrid infrastructure will require new capabilities of campus LANs.

To enable the testing of various hybrid approaches, the HOPI testbed will make use of resources from Abilene, from the RONs, from National Lambda Rail (NLR, of which Internet2 is a key participant), and from other internationally connected facilities such as MAN LAN (Manhattan Landing) in New York, NY.

A HOPI design team, composed of engineers from the Internet2 and international research and education communities has been assembled to plan the HOPI test facility. This document is the work of the design team and summarizes the recommendations for the testbed.

The HOPI Design Team

The members of the HOPI design team are:

- Linda Winkler, Argonne National Laboratory
- Rick Summerhill, Internet2
- Peter O'Neil, NCAR
- Bill Owens, NYSERnet
- Mark Johnson, North Carolina Research and Education Network
- Tom Lehman, USC Information Sciences Institute
- David Richardson, University of Washington
- Chris Robb, Indiana University
- Phil Papadopoulos, University of California San Diego

- Jerry Sobieski, Mid-Atlantic Crossroads
- Steven Wallace, Indiana University
- Bill Wing, Oak Ridge National Laboratory
- Cees de Laat, University of Amsterdam
- René Hatem, CANARIE
- Sylvain Ravot, Caltech
- Guy Almes, Internet2
- Heather Boyles, Internet2
- Steve Corbató, Internet2
- Chris Heermann (Facilitator), Internet2
- Christian Todorov (Scribe), Internet2
- Matt Zekauskas, Internet2

Future Networks and Architectures

The current Abilene network provides a shared packet switched infrastructure for the research and education community. It has been, and continues to be, extremely successful. Recent experiments have demonstrated single flows on the order of 6.2 Gbps across the Abilene network, and total traffic usage on particular links has approached 9.5 Gbps. In looking to the future, however, the community will require greater capacity and networks that are dedicated to particular disciplines or individual services. How these networks evolve at the campus, regional, national, and international levels is under investigation.

Perceived Problems with Packet Switched Infrastructures

While Abilene has been extremely successful, the future effectiveness of a shared packet switched network in meeting the needs of all communities is being challenged. At the heart of the scrutiny is the fact that Abilene is a fixed bandwidth, shared IPv4 and IPV6 infrastructure. User communities have expressed the requirement to sustain multiple simultaneous 6 Gbps flows. While enhancements to Abilene in the past have been accomplished by increasing the bandwidth of single links from 2.5 Gbps to 10 Gbps, it is unlikely such links can be upgraded to 40 or 100 Gbps by the time the next generation network must be implemented. Investigation is needed into alternative architectures to support large traffic requirements.

Increasing demands are being placed on the network for more deterministic capabilities. One expected benefit of adding deterministic capabilities is that applications will have greater flexibility to utilize alternative transport protocols not restricted by the requirements to share resources fairly with other users. Alternative technologies to support deterministic behavior need to be explored. Some disciplines require lower layer facilities, such as dedicated waves or circuits, to accomplish their tasks.

The packet switched infrastructure has been extremely successful in solving a multitude of problems, but some are looking to circuit switched capabilities to provide additional features, including deterministic paths and dynamic provisioning of such paths.

Facilities Based Infrastructures

The ability to obtain access to fiber assets and/or multiple waves at both the national and regional levels has dramatically changed the landscape and allowed us to look at new architectures. The campus level of the hierarchy has had such facilities for some time, but the ability to connect to regional and national infrastructures via dark fiber at a variety of different layers in the network stack has only recently come to pass. The ability to create waves from campus to campus, or to build Ethernet paths from a physics department on a campus to the CERN Large Hadron Collider, has only recently become realizable.

The Future Infrastructure

Given these newly available facilities, what is the future likely to provide? With dark fiber available at all levels of the hierarchy, the ability to provision and switch waves will almost certainly exist as part of the infrastructure. The ability to provision and switch waves from campus to campus, or even from user to user, will undoubtedly be available to the community. It is likely there will continue to be a need for a shared packet based infrastructure and this packet based network may well be provisioned from the set of waves available to the community.

Therefore, the future infrastructure is likely to consist of the following:

- **A pool of waves that can be provisioned and switched dynamically as needed.**
- **An IP packet based infrastructure for general communications, wave/circuit provisioning control plane, and data transport for those applications that do not require the dedicated resource paths provided by the wave infrastructure.**

It is this set of resources that motivates the **HOPI** project, and in particular, the creation of a testbed that can model these resources for future development.

Architectural Issues

Given the expected new infrastructure, there are many issues to explore. Should deterministic paths or optical waves extend from host to host? If deterministic paths or optical waves can extend to the edges of the new network, how close to the end-points should they extend? Should they extend from campus router to campus router, departmental router to departmental router (layer3), departmental switch to departmental switch (layer2), or host to host? What is the topological extent of the deterministic path across the new infrastructure?

Even more basic, what are deterministic paths? What properties should be important for a deterministic path? What latency, jitter, or loss values are acceptable for a path to be considered deterministic? How do these compare with properties associated with a packet switched infrastructure and under what conditions might the packet infrastructure provide such properties? How do they compare with a raw optical wave?

How dynamic should the deterministic or optical paths be? Should the time scale be in weeks, days, hours, minutes, seconds, or even faster? How fast should dynamic provisioning occur in the new architecture? For example, in the distant future, will a network consist of wave switching where a user runs software that builds the

equivalent of a phone-switched path at the optical layer, does a file transfer, and then disconnects? What direction should new architectures take with respect to dynamic provisioning?

If dynamic provisioning of some sort is to be part of the new architecture, what is the average duration of a path setup? That is, if part of the architecture is circuit switched, what call duration can we expect? Switched path duration might well depend on discipline specific requirements.

If dynamic provisioning is to be part of the new architecture, then there will be cases where a deterministic path is not available. What scheduling requirements are necessary to ensure users can accomplish their tasks, and what resources must be available to ensure that scheduling problems don't become a dominant problem on the network?

What is the interaction between the packet based infrastructure and the circuit switched infrastructure? Will it be necessary to dynamically provision deterministic types of paths across the packet switched infrastructure?

What does the classic backbone/regional/campus hierarchy become in this new environment? Does a backbone simply become a layer1 resource that supports layer2 or layer3 services on top of the network? How will regional infrastructures be designed and how will they be connected to a backbone infrastructure? Do they connect at layer1, framed with a layer2 protocol, or do they connect at layer2 with framing at layer3?

How will two national infrastructures connect to each other, to national networks supporting federal programs, or to international networks? Will these be at layer1, layer2, or as they exist today at layer3? What framing will be used on such interconnects, if at all?

There are many questions about the new architecture and yet there is limited infrastructure available at this time. How can these questions be examined in the near future?

The HOPI testbed will provide a vehicle for examining the above questions, not necessarily by providing a rich set of optical waves, but by taking the available resources and modeling the future infrastructure with a set of deterministic paths that can use existing facilities and commonly available equipment. It is the fact that the HOPI testbed can provide a model for what is expected to be the future infrastructure that makes it a viable resource for looking at future architectures.

Scaling Issues

It is clear that a completely circuit switched architecture, even one based on a large number of optical waves, cannot scale to the full Internet2 community. However, scaling to the full community may not necessarily be a goal – a user sending a short email does not need a deterministic path. Since the goal of the testbed is to understand new architectures, scaling will be examined whenever possible and clearly noted as part of the results of the testbed.

The HOPI Testbed

The HOPI testbed will consist of a variety of available infrastructures, including Abilene and a wave on the NLR facility. The testbed may incorporate particular layer2 components, but the emphasis will not be on those components per se, rather to examine the basic architecture. The design of the basic service, the HOPI node, connector interfaces, or the control plane may involve particular technologies from the existing infrastructure, but the study of basic architectural issues is of primary interest. Knowledge gained from working with particular technologies will undoubtedly be very valuable, but the goal is to study architecture.

The Goal of the Testbed

The goal of the testbed is to examine potential future network architectures.

Therefore, it will create a pool of near deterministic paths to model future infrastructures. The paths must be capable of being dynamically provisioned. The testbed must also consider how packet switched and circuit switched infrastructures might be combined into a hybrid architecture for the future.

The key elements of the testbed are deterministic or near deterministic paths (to be called lightpaths in the remainder of this paper, without strict definition), dynamic provisioning, and how the packet switched infrastructure relates to these two elements.

While different types of experiments might be conducted on the testbed the overriding goal remains a study of new architectures.

Testbed Resources

The testbed will draw resources from a variety of different facilities, including:

- The Abilene Network
- The Internet2 NLR Wave
- The MAN LAN Experimental Exchange Facility
- The Regional Optical Networks

The well known Abilene network is a 10 Gbps IP backbone network with production qualities. It is the primary Internet2 backbone today and provides a rich packet switched infrastructure to the testbed. Recent experiments on Abilene have incorporated MPLS L2VPN tunnels across the network to provide dedicated paths for demonstrations and dedicated networks. For example, a L2VPN is currently being constructed between the Abilene NOC and the Internet2 Technology Evaluation Centers. Through the use of MPLS L2VPN tunnels, Abilene is expected to provide immediate access to the HOPI testbed for all current Abilene connectors.

The second resource for the HOPI testbed is a national scale 10 GigE wave on the NLR footprint. Phase One of this footprint is expected to be available in September 2004 and will be connected, as described below, to the Abilene network. The wave will be framed with 10 GigE, providing interfaces facing east and west at each of its terminal nodes. At each terminal node on the NLR footprint, either a HOPI node will connect to the east and west facing interfaces, or the wave will be passed through by cross-connect. See Figure 1 for a map of the Abilene and NLR resources available to the testbed.

The third resource available to the testbed is the MAN LAN experimental facility. The core component of the facility consists of an Ethernet switch that provides 1 and 10 GigE interconnection and peering for traditional research and education packet based networks. It is a research and education based exchange point that provides international networks with a variety of classical peering options. The MAN LAN experimental facility, consisting of an ONS device, adds the potential for interconnecting at layer1 using either SONET or Ethernet framing. A variety of deterministic paths are available between Abilene and international networks. The current connectors to the experimental facility include the CANARIE network in Canada and the Surfnet network in the Netherlands. The experimental facility has already been used in several experiments and can be used as part of the HOPI testbed.

Figure 1

Finally, an essential component of the testbed will be the emerging regional optical networks. They will interconnect the campus LANs with the testbed. A growing number of RONs are obtaining fiber to support their regional facilities and they will play a crucial role in the HOPI testbed. Potential interconnect between a RON and the testbed will either be direct to the node via dark fiber or Ethernet services, or through the Abilene network.

Basic Services Definition

The HOPI facility will provide dedicated unidirectional data paths between applications facilities. These applications facilities may be user code running on individual end systems, core IP routers, entire networks, or some combination thereof. Whatever the actual service termination points, the connection established

between the points – and the technologies developed to establish the connection – should be capable of providing **deterministic** data transport across these lightpaths.

The HOPI Design Team decided to start with a very simple and limited definition. Initially, HOPI will provide deterministic lightpaths with characteristics as follows:

- The bandwidth capacity of a given path will be fixed at either 1 Gbps or 10 Gbps.
- A latency formula will be developed and validated that bounds data transit time as a function of propagation delay through all network elements along a given lightpath under optimal propagation conditions – that is, an idle network. Latency will then be bounded by this formula plus some constant.
- A jitter function will also be devised and validated to bound the variance in the latency, again, assuming optimal propagation conditions.
- A loss function will also be devised in a similar manner.

The Design Team expects the services to be expanded in scope and refined in accuracy as the project evolves. The desire is to develop the deterministic lightpath service parameters in a manner that allows extension and application to many different network environments – not just in the HOPI testbed. Furthermore, monitoring techniques will need to be developed to provide accurate platforms for verification and fault isolation for lightpaths.

The Client Interface

A HOPI client is broadly construed. A client could be a single workstation for an experiment, a Regional Optical Network participating in the HOPI experiments, or a service provider providing static commercial transport circuits that HOPI may wish to integrate or otherwise utilize.

The physical client interface is nominally a single fiber connection to a HOPI node port. The architecture of the node is described below, but the port can be assumed to be a 1 or 10 GigE port on a high performance Ethernet switch or TDM (time division multiplexing) device.

The Ethernet connection hints at the need for the formulaic approach to predicting latency, jitter, and loss, and to the provisioning complexities that must be addressed to insure optimal propagation conditions for the data path.

Bandwidth Capacity

One approach to capacity would be to simply assert 10 GigE LAN PHY as the basic service. All connections into, and all lightpaths provisioned over HOPI would be 10 Gbps. While this simplifies many aspects, it also limits the practical use of the testbed, and would provide little insight into how connection oriented services can be provisioned for data applications with needs other than just bandwidth. The lower level service consisting of 1 GigE segments provides sufficient lightpaths to experiment with rich architectures.

Other bandwidth granularities could be used as well. For example, 100 Mbps is possible. There is nothing that prevents the service definition from including additional discreet bandwidth levels later, or even continuous bandwidth specifications at a later time. However, starting with 1 and 10 GigE simplifies the project.

It should be noted that the basic service definition does not specify Ethernet as an absolute requirement. Despite the Ethernet physical client interface described above, the HOPI design includes considerations for interconnecting deterministic lightpaths across many different types of framing and transport layers. An example would be to concatenate path segments provisioned over HOPI to path segments provisioned across either Abilene, the hybrid data path, or across other networks. Such hybrid paths will have enormous value to RONs that may not have direct HOPI access except thru Abilene. Another primary goal is to provision hybrid lightpaths between Europe and North America – routing the path over Ethernet framing to SONET framing and potentially through MPLS tunnels.

Latency

Latency will vary according to the length of the physical data path, the number and types of network elements along the path, and the characteristics of those network elements. HOPI will attempt to understand latency, and to deterministically control and bound it. In the first approximation, the HOPI desire is to provide TDM-like performance to lightpaths regardless of their underlying transport path. To this end, HOPI would like to provide transit performance as if the data had the network all to itself. Specifically to provision a path that is not interfered with by other traffic or other lightpaths.

By allowing smaller granularity paths (1 Gbps) to be aggregated into larger segments (10 Gbps), bursty traffic can in fact create transient queuing delays resulting in higher latency and higher jitter. So the multi-rate service definition requires sophisticated buffer and queue management capabilities in those network elements performing the aggregation along the lightpath route.

Jitter

Jitter is the variance in latency. It is a function of many characteristics along a path: packet size, framing technology, and queuing delays are just a few examples. Jitter can be affected by collateral or cross traffic as well. The objective of the jitter service definition is to understand and to be able to predict the jitter bounds based upon a-priori knowledge of the lightpath's route and data characteristics, other collateral path characteristics, and then to bound that jitter using queue management techniques similar to those used to address latency issues.

Loss

Loss is also a function of many characteristics along the path. Loss can be affected by layer1 transport issues related to bit errors or framing issues at layer2. Ideally, loss should be zero on a lightpath or at least as close to zero as possible, to ensure the ability for higher layer flow mechanisms to provide robust data transfers. The ability to monitor and measure loss along a lightpath and understand its characteristics must be examined by this project.

Layer3 Transparent / Layer2 Opaque

While the customer hand-off is framed as Ethernet, the initial service offering is not some type of extended, private Ethernet service. The HOPI service will provide a facility that is layer3 transparent, but layer2 opaque. The bandwidth, latency, jitter, and loss guarantees described above will be provided for any layer3 protocols (IPv4, IPv6, non-IP, etc.). However, HOPI will not pass through the various Ethernet layer2

signaling protocols and features, such as VLAN tags and spanning tree. In this way, HOPI provides a generic layer2 that serves only to pass packets between two points on a deterministic data path, but does not provide the features of any particular layer2 technologies.

HOPI may choose to use Ethernet signaling as a UNI control protocol between the user and the HOPI edge interface. For example, it is conceivable that HOPI will allow a connector in Chicago with a single physical interface to mark packets with a VLAN tag that indicates that the packets should be sent either to a HOPI connector in Denver or a connector in Florida. This technique allows a user to choose one deterministic data path or another, but it does not mean that a user's VLAN tags will be passed through the network and appear on the other side. HOPI may also choose to use Ethernet signaling protocols to build the HOPI service, such as marking each deterministic data path with a unique VLAN and applying 802.1p prioritization, but this is just an artifact of implementation and not extended to the user.

At a future date, this could be changed, and HOPI could explicitly offer users an extended layer2 service. Such a service would need to specify which layer2 features are extended to the user, and which are reserved for HOPI use.

HOPI Node Design

The HOPI node design must support the basic service definition above and connect to the HOPI resources in a manner that provides a rich potential for experimentation.

A primary technical resource is the Internet2 NLR wave. It can be viewed as essentially a series of point to point wavelengths statically provisioned along the NLR fiber footprint. Along that footprint, there is one wavelength provisioned between each pair of adjacent NLR terminal locations. The termination points of these waves are transponders that assert a 10 Gbps Ethernet framing onto the wavelength, and express this Ethernet framing to the client side interface. There is no economical means at this time to redefine that framing and therefore it must be viewed as essentially a series of point to point 10 Gbps Ethernet segments.

A subtle but important constraint on these point to point Ethernet segments is that the end points of these segments are fixed. That is, the Ethernet segments that comprise the wave cannot be changed at the **wavelength layer** to other locations around the country. The Ethernet segments could potentially be concatenated, effectively moving the endpoints of a given connection, but this is done at the expense of excluding all access to the waves from the intermediate transit terminal locations.

While these waves taken individually as point to point connections will have 10 Gbps of deterministic capacity, this determinism does not necessarily propagate down to smaller capacity lightpaths that might be mapped over a particular link simultaneously. For instance, interconnecting HOPI Ethernet segments via 10 Gbps interfaces on an Ethernet switch located at the terminal locations will insert switch buffering, possibly switch backplane architectural constraints, and possibly other traffic characteristics into the path performance equation.

In addition to the wave on the NLR footprint, the Abilene network will be connected to a HOPI node via dark fiber and provisioned with Ethernet, most likely at 10 GigE,

but possibly at 1 GigE in some cases. The connection will be directly to an interface on the closest Abilene core node router.

The HOPI Node

A HOPI node should then allow for a cross-connect between the following types of connections without manual reconfiguration:

- Directly between a HOPI connector and either an east or west facing interface on the NLR wave.
- Directly between Abilene and an east or west facing interface on the NLR wave.
- Directly between Abilene and a HOPI connector
- Directly to an interface on either an Ethernet switch or TDM device.

The ability to reconfigure the connections to a HOPI node will allow different network components to be connected to the waves at different times. This reconfiguration capability provides a high degree of flexibility in allocating different resources at different times. While introducing substantial complexity into the testbed, it provides the ability to dynamically allocate 10 Gbps resources to particular experiments. A major goal of the HOPI project is to explore the tools and techniques that are being developed to control and manage dynamic deterministic data path services. The systems and software that provide for dynamic reconfigurability of the lightpaths is called the control plane.

The HOPI Node is initially composed of three key hardware components:

- A fiber cross-connect switch (FXC) – as the name implies, this device allows pairs of fiber to be connected to each other, and configured such that other pairs of fiber are connected to each other.
- An Ethernet switch or TDM device – must be 1/10 Gbps capable, and support both large MTUs and aggregation of smaller Ethernet based lightpaths into larger trunks.
- Control, management and support components – these devices include an IP router for out of bandwidth control and potential control plane connectivity, a PC to support ad hoc control plane and management activities, a measurement device, and a general device to support other potential user services. Power controllers are also an essential ingredient.

Other components will be included as part of a HOPI node as needed.

Fiber switches are relatively new, but they provide direct cross connects at the fiber level as well as remotely manageable reconfiguration capability between the HOPI client interfaces and other users attached to the fiber switch. Such switches will not provide sub-rate aggregation, nor will their framing agnostic capabilities be of much direct value to the HOPI waves (since the HOPI waves are fixed as LANPHY Ethernet). However, these switches can provide a basis for dynamically reconfiguring the HOPI node to support direct and dedicated access to the east-west waves, Abilene, and interconnecting other devices into the physical fiber path (e.g. optical splitters, TDM equipment, etc) that will augment the overall HOPI dynamics.

Perhaps the most practical means of providing circuit-like cross connectivity is to place either a high-end Ethernet switch or a TDM device at the node locations.

Either device will allow a relatively broad set of users/networks to establish port connections at the node, and then have those Ethernet ports cross connected to the HOPI backbone segments as necessary or appropriate. Further, these cross connects can be established manually through the fiber cross-connect and switched as needed through the larger capacity paths.

An Ethernet switch, if used at a HOPI node, should be capable of providing non-blocking 10 Gbps port-to-port capacity such that configuring 10 Gbps paths through the switch are essentially jitter free with latency bounded only by the maximum MTU size. Figure 2 is a schematic representation of the HOPI node:

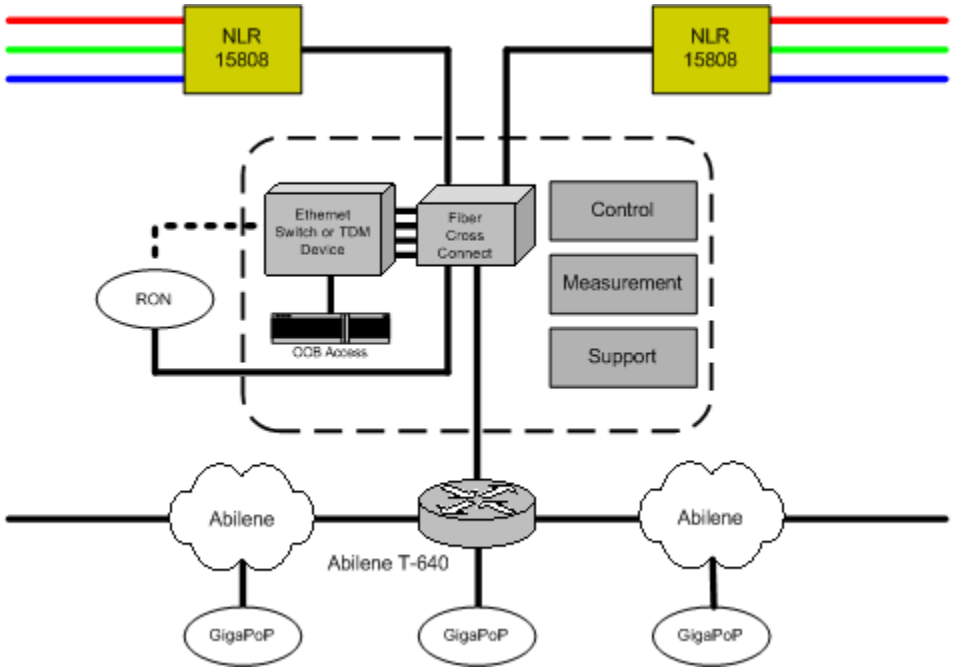


Figure 2

Ultimately, the architecture at these terminal locations should be targeted at being able to dynamically switch wavelengths, of different framing, under direction of mature routing and signaling protocols in the control plane. Such aspirations are down the road yet, but HOPI has been designed with such capabilities in mind.

The HOPI project, its purpose and implementation, should not replicate the programmatic goals of other projects. While replicating technical configurations of other programs may be unavoidable, the HOPI Design Team has attempted to identify programmatic goals that are unique:

> **The HOPI project should focus on architecture, and in particular deterministic lightpaths and dynamic provisioning – where Ethernet framing is just one of several possible data link layers that could be used.**

This HOPI node would also serve as the demarcation interface for local networks to HOPI. The basic interconnection would be 1 or 10 Gbps Ethernet framing. Any conversions necessary to present this framing would be the responsibility of the

access network. For example, a network delivering SONET framing (even if it is a 10GE WAN PHY) will need to terminate the OC192 so as to present the Ethernet client handoff to HOPI. Similarly, time division multiplexing of two 1 Gbps Ethernet links over an OC48c would need to be presented as two separate 1 Gbps Ethernet clients rather than a single OC48. The HOPI terminal architecture fiber cross connect allows for connection of such TDM devices to the HOPI node. The cross connect will assist and support dynamic allocation and provisioning of these resources.

Management and Control

The HOPI switches will be accessible out-of-band by a remote operations center. Out-of-band access will be via a conventional IP network and will allow remote access to the switches from any convenient location for management, reconfiguration, control plane functions or other operation reasons.

This proposed configuration allows the network operators to remotely access each switch along a desired path and manually configure that switch to cross-connect appropriate ports, perhaps associate certain policing, shaping, or queuing strategies with particular cross-connects, etc.

While manual reconfiguration initiated by phone call or email may seem a bit arcane, it will be adequate for initial testing, verification, and study of the performance value of the infrastructure. However, the goal should be to ultimately have a dynamically reconfigurable, connection oriented transport layer that is fully routed and signaled end to end. Such routing and signaling protocols and their end-system interfaces are not mature – indeed, reasonable people argue the merits of several competing standards, and those standards are not fully developed, deployed, or tested. A dynamic connection oriented transport system that peers with and interconnects with other networks is not something that can be deployed easily or reliably in the near term, and just getting the base resource in place will take considerable work.

HOPI Node Locations and the Abilene/NLR Interconnect

Figure 1 illustrates the geographic locations of Abilene router nodes and NLR terminal nodes. For each city common to both facilities, an indication is given under the city name that describes the mileage between the two PoPs.

A HOPI node might be located at an Abilene router node, or at an NLR PoP, or some third party location. Exact locations will be accurately described during the HOPI testbed implementation phase. Connectivity between the NLR wave and the Abilene packet infrastructure will consist of dark fiber to the fiber cross-connect switch at the HOPI node.

A HOPI node may not be initially provisioned at every Abilene/NLR interconnect location, depending on HOPI participants at that location. For example, a direct cross-connect between the NLR terminals may be provisioned in the beginning, and then replaced with a full HOPI node later as required by local connectors.

Moreover, in some locations where there are a minimal number of connectors to HOPI, the initial interconnect between Abilene and the HOPI node may be through a 1 GigE interface, with an upgrade to a 10 GigE interface as the node evolves and grows.

Connector Interface

The initial connector interface to HOPI will be through either a 1 Gbps or 10 Gbps Ethernet interface. Connectors can connect directly to the HOPI switch or to the fiber cross-connect. Connectors can also connect to HOPI through an MPLS L2VPN from Abilene. This type of connection will probably be the most common during the initial stages of the HOPI testbed.

Connectors who come directly to a HOPI node can do so via dark fiber and connect directly to the fiber cross-connect, or use a provisioned Ethernet service and connect either to the Ethernet device or to the fiber cross-connect. The expected and optimal path to a HOPI node is via dark fiber from a connector, but an Abilene connection may also be used. The connector should provide the necessary infrastructure to experiment on deterministic paths and dynamic provisioned services.

In connecting through an Abilene MPLS L2VPN, the connector must work with networks behind the gigaPoP to provision the MPLS tunnel through to the HOPI node Ethernet switch.

The following is a summary of connector interfaces to a HOPI node:

- 1 or 10 GigE connection to the fiber cross-connect device. The connection may be through an Ethernet service provider, including NLR, or through dark fiber directly from a RON
- 1 or 10 GigE connection to the Ethernet Switch or TDM device. The connection may be through an Ethernet service provider, including NLR, or through dark fiber directly from a RON.
- An MPLS L2VPN service through Abilene to the Ethernet or TDM device. Connection is through existing national/regional/campus facilities.

Control Plane

In the context of modern switched digital networks, the control plane constitutes those signaling and routing protocols and related services that allow the network to allocate and dedicate specific network resources of links from one location to another. A similar term data plane refers to said links that actually carry the user's data.

Control Plane Ideas

The HOPI testbed will provide dedicated network resources in the form of lightpaths to the participating projects and applications. This necessarily requires the ability to allocate and reserve resources on some time variant basis to different projects. This provisioning process will be the focus of the HOPI Control Plane architecture.

The control plane roadmap will need to address several issues:

- Routing. Given a set of user specified lightpath characteristics, the network must select a path supporting the appropriate interconnecting functions to provision the requested lightpath.
- Signaling. This is the process by which the network elements allocate and map the switching relationships at each switching element along the path.

- User Interface. How does the user access the control plane to submit a request for a lightpath? And once the data plane path is established, how is it accessed by the user application?
- Policy. Mechanisms are needed to determine which users or projects (or other networks) are allowed to use this facility and under what conditions. Authentication, Authorization, and Accounting (AAA) network security services provide the primary framework through which we will examine access control issues.

Significant effort will be required to deploy a flexible and robust control plane for the HOPI testbed resources.

Initially, the HOPI testbed will be a relatively simple collection of Ethernet segments. To make them useful, there must be a means to allocate subsets of these segments to specific requests, and concatenate them appropriately to construct cross-country lightpaths. Further, constructed lightpaths will need to be interconnected with other networks (e.g. CANARIE, GEANT, Starlight, or the RONs). This will require reconfiguring the HOPI Terminal Nodes along with interconnecting HOPI and other optical network control planes.

The HOPI Terminal Node Architecture provides for a dynamic switching capability at the terminal nodes through Ethernet devices and/or fiber cross-connects. The switching capability was established explicitly to support reconfiguration of the overall resource. However, there is no standard set of routing, signaling, security, interface protocols and services that provide this functionality. There are several evolving efforts to develop such control plane services - some academically based, some industry based, and some focused on one type of framing or another, or one specific environment or another. But all of these activities are in early stages and none seem to address all of the issues associated with providing these capabilities on a global extensible end to end basis.

The process of allocating dedicated lightpaths is complex. The software required to do this in a robust fashion is not yet available, and there is no consensus within the community as to how control plane services should be implemented.

Therefore, the proposed HOPI control plane will be implemented in three phases:

- In the first phase, the initial control plane will be manually configured by the HOPI engineering and operations team for each request.
- The second phase will be a more automated means of setting up lightpaths through the HOPI environment.
- The third phase will deal with the inter-domain issues associated with providing services across multiple networks, with different service and policy definitions, different technologies, and doing so on a persistent and reliable basis.

Some of the projects and organizations that are working in the control plane area include

- CANARIE User Controlled Lightpath Software developments from University of Waterloo, Ottawa University, Universite Quebec a Montreal and Carleton University
- The Optiputer Project

- The Omninet Project
- The DRAGON Project
- The CHEETAH Project
- IETF GMPLS standards

Of course, there are other projects and an attempt will be made to coordinate activities with these projects. There is a great deal of work being done in this area and it is too early to define specifically what the control plane will be in the next two years.

The Phase One Control Plane

The Phase One control plane will use manual configuration to affect a control plane. That is, HOPI Engineers will be responsible for scheduling and reconfiguration of the backbone waves to meet specific user requests. The terminal node switches can be remapped to either a) concatenate the East and West waves in a pass-thru scenario, b) drop either or both of the East/West waves to a RON interface at that node, or c) drop either of the East/West waves to Abilene. This reconfiguration will be done by remotely logging into the appropriate elements and modifying the switching to meet the lightpath requirement.

Manual configuration is only expected to provide a near term capability to set up lightpaths for the initial set of project tests. It will remain a capability through the life of the project, but is only expected to be used during the first phase of the project.

One could envision a less human intensive version of manual reconfiguration in the form of perhaps a web based front-end to a set of scripted configuration commands and/or scheduling, routing, and provisioning engines. Such an interface would reduce manual intervention by the operations and engineering staff, and would provide a better platform for developing more advanced and robust models for managing the overall network. It is anticipated that this capability will be implemented as Phase One evolves.

Such interfaces do not address some important aspects of lightpaths. The HOPI service defines lightpaths of both 1 and 10 Gbps. This provides greater flexibility by allowing potentially many 1 Gbps lightpaths to be in existence on a particular segment at one time. The control plane will need to consider the performance issues/limitations associated with aggregating multiple 1Gbps access lightpaths over a 10 Gbps backbone wave.

The control plane model must address not just the resource definition between end points, it must also address the interconnecting requirements of different end points; that is, the mating of paths with dissimilar characteristics (e.g. Ethernet versus SONET) into a single lightpath with some common end to end service definition. Much of the Phase One control plane work will be to develop a generalized model of the service termination points and service definitions – the parameters that need to be specified, the manner in which the service is ultimately delivered to (or accessed by) the user, etc. This experience will feed into the Phase Two control plane deployment.

Phase Two Control Plane

Phase Two will be characterized by a considerably higher degree of automation and service sophistication. The routing architecture and appropriate protocols can be deployed in a fashion that allows for considerably more sophisticated lightpath planning – defining the necessary path characteristics, locating suitable paths, scheduling the resources in an environment of competing priorities, etc. The actual process of provisioning the paths is expected to be more sophisticated as well, employing a more distributed process.

Phase Two will include services such as a bandwidth only lightpath, i.e. a dedicated data path that is characterized only by the amount of bandwidth that is guaranteed to be available within certain latency, jitter, and loss bounds. Such a path could be provisioned over a heterogeneous set of network technologies, including Ethernet framed waves, IP routed networks, SONET circuits, etc. Such lightpaths rapidly exceed the ability of operations personnel to manually configure them.

For these reasons, Phase Two is expected to include distributed and automated routing services and protocols complemented by signaling protocols that can streamline the entire process. Distributed models are more robust, and automated protocols simplify the process of hardware network modifications (either planned or unscheduled outages). The HOPI environment will include a growing list of backbone segments, peering networks, terminal node hardware and service offerings, and the environment will no longer be easily manageable on a manual basis.

While all this sounds complex and foreboding, there has been and continues to be movement towards these objectives within the standards bodies. In particular, the IETF has been hard at work standardizing extensions to routing protocols such as OSPF and ISIS that provide for the path selection through the network based on a "label type". These label types specify different framing types such as Ethernet, SONET, fiberchannel, MPLS, etc. Standards also exist for similar extensions to signaling protocols such as RSVP. And other new protocols have been specified to fill in holes in the overall picture such as Link Management Protocol (LMP). These standards taken as a whole constitute the Generalized MultiProtocol Label Switching architecture (GMPLS).

There are many people and projects in the R&E community looking at GMPLS technologies to address the needs of advanced e-science. Programs such as DRAGON and CHEETAH are extending and complementing the existing GMPLS capabilities with adding functionality for the control plane, user interfaces, and inter-domain policy application. HOPI will incorporate these technologies for evaluation when and where they make sense and deploy them as appropriate. The collaboration between HOPI and these other research and experimental networking programs is expected to accelerate the benefit to the user community perhaps by years.

Phase Three Control Plane

The issues related to the third phase of the HOPI control plane deal less with particular node architectures or specific service definitions, but attempt to address the issues associated with building lightpaths end to end on a global basis and doing so in an operationally sustainable fashion.

This means dealing with issues such as practical mechanisms of defining meaningful service policies for a network, being able to advertise these service definitions to inter-connecting networks, expressing policy on when and how these services can be accessed, and reconciling the allocation and utilization of the dedicated resources to recover costs.

Currently, there is no consensus within the networking community on how to establish peering relationships between different administrative domains. Lightpaths change the fundamental meaning of the inter-carrier interface. The ability to provision lightpaths end to end, with potentially many different service requirements, in an automated fashion, in a secure and authorized environment, on demand and near-real time, and recover costs is difficult to do on a single campus, never mind across multiple independent networks and on a global scope.

Phase Three is intended to extend the basic routing, signaling, and user interfacing deployed in Phase Two into the Phase Three multi-domain nearly-production realm where reliable intra-domain services meet the market plane of global inter-domain service practicalities.

Authentication, Authorization, and Accounting will play a crucial role in control plane functions. The HOPI project will attempt to document and collect likely AAA requirements as it implements control plane software, and to incorporate AAA functions whenever possible, adhering to standards based approaches.

Management

NOC Services – Engineering and Management

Engineering and Management of the HOPI testbed should proceed much like any backbone network, emphasizing engineering design, installation, monitoring, change-management, call center functions, debugging, and user support. The testbed brings special requirements, emphasizing experimentation more than production services. Whatever NOC services are utilized, experimentation and the ability to support specific tests will be crucial to the success of the testbed.

There are several obvious possibilities for NOC services. Since both Abilene and NLR will be essential elements of the testbed, Internet2 should investigate the services provided by these two organizations to obtain basic NOC services. It is essential, however, that the NOC services deployed by NLR and Abilene be capable of working closely together to solve potential HOPI testbed problems.

Control Plane Software Development

Since the devices that will become part of the HOPI testbed are unlikely to have built-in control plane and signaling support, most of the tests performed during the early stages of the testbed will require ad hoc software. This software will most likely be simple scripts deployed on control computers that can signal across the IP network and also talk to the devices, through proprietary interfaces or SNMP. There are two important factors in the design of this software:

Every attempt should be made to coordinate with other groups developing this type of software, including both the campus and international communities, and those from other projects. A catalog of freely available software should be kept and made

public. Coordination between other projects and testbeds will be crucial to the success of the project.

Every attempt should be made to follow the guidelines of existing standards. For example, if control of a fiber cross-connect is to be defined and it requires signaling through the control plane, a standard signaling interface should be used whenever possible.

Monitoring and Measurement

Monitoring and measurement may well be a difficult problem. At the early stages, basic and common monitoring techniques can be used, but as the testbed evolves, it may be necessary to monitor and measure either layer1 or layer2 attribute types that are not often done in today's layer3 environment. New techniques will need to be developed and new ideas put forth on how to provide this essential capability. If future networks are to provide dedicated and deterministic paths to users, it is essential that we be able to monitor the paths for problems, and measure the paths' capabilities to insure that the expected performance is presented to the user.

HOPI Test Cases

Dynamic Provisioning Test Cases

1. **Host to Host Lightpaths**. Determine the requirements for implementing the dynamic setup of lightpaths from end to end. Consider the necessary layer1, layer2, and layer3 requirements for such a path and how to implement those requirements dynamically. Incorporate software to provide the dynamic configuration across all types of lightpaths, including MPLS L2VPN tunnels and Ethernet paths. Determine the required scheduling mechanisms in cases when lightpaths are not available, and understand the degree to which dynamic configuration can be implemented – is it days, hours, minutes, or seconds? Over the lifetime of the testbed, a goal might be to provision host to host lightpaths measured in some number of seconds.
2. **Cluster to Cluster Lightpaths.** Experiment as in 1) above, but from a computer cluster to a computer cluster, with Ethernet switches as the endpoints of the lightpath. For example, consider similar experiments emphasizing the added complexity introduced by having many hosts on each end of a point to point path.
3. **Dynamic Backup for the TeraGrid Backplane.** Consider a network such as the TeraGrid and model it using lower bandwidth lightpaths. Create these as backup point-to-point paths for individual segments on the TeraGrid, understanding the necessary components to detect outages on the TeraGrid and how to dynamically setup backup paths. The idea here is not to actually backup the TeraGrid, rather to understand how a future architecture might provide redundancy features to a separate network, one with possibly different types of goals and features.
4. **Dynamically Implemented Testbed for the Internet2 NOC and ITECs.** Implement a testbed across the HOPI infrastructure for the Internet2 NOC and ITECs, and provide dynamic control for bringing the network up and shutting it down. The network would not be up all the time – much of the HOPI testbed will be eventually be dynamic - but could be created under NOC and ITEC control when needed. What are the software requirements to

dynamically control such a network and what are the problems to be solved for implementation?

5. **Testbed Responsiveness**. Examine the responsiveness of the testbed provisioning capability. Investigate the efficiency with which HOPI resources can be dedicated for experimentation.
6. **Dynamic Allocation and Control of Lightpaths using Dragon.** The DRAGON project is an NSF funded effort to develop tools and techniques required to allocate and manage wavelength services. The DRAGON Project PI's propose to port the tools to the HOPI testbed for validation and evaluation.

   The DRAGON Virtual Label Switched Router (VLSR) is designed to integrate GMPLS control plane protocols with devices that do not have native GMPLS protocol support. The VLSR will be configured to perform the connection setup/teardown functions for the HOPI terminal equipment - specifically the Ethernet and fiber switching components, as well as the assignment (allocation) of the HOPI backbone waves to requesters. The VLSR will initially use SNMP to manage the switching components, but is extensible to include TL1 or other interfaces as well. DRAGON is also developing inter-domain protocols for advertising Transport Layer Capabilities between autonomous domains. This Network Aware Resource Broker the DRAGON project would like to incorporate HOPI into a set of optical network domains where such inter-domain interconnecting can be deployed and tested in a broader sense. This effort will leverage HOPI in two ways: First, as simply another domain peering with optical networks in the Washington DC area (specifically DRAGON, ATDnet, and HOPI). In this first stage, we hope to simply iron out bugs and verify functionality. The second stage will be to utilize HOPI as a core long haul domain acting as a transit domain for transport layer services between geographically dispersed optical network initiatives in Washington, Chicago, Raleigh Durham, or California.

Deterministic Path Test Cases

1. **10 GigE Wave/Switch Comparisons.** Accurately test packet delivery across a coast-to-coast 10 GigE links in two modes:
   - East and West transponders cross connected at each terminal town along the path.
   - East and West transponders drop into 10 GigE switch ports through VLANs. Measure the delta in terms of jitter, latency, BER, and loss. At the end points, attach the 10 GigE to test devices to source and sink the streams.
2. **Compare Point-to-Point Wave Paths to Abilene Paths.** Set up a cut through path between Abilene routers: one on the East Coast and one on the West Coast. Study the RTT characteristics of packets using 50-byte and 9000-byte packets. Compare the following:
   - The concatenated transponder HOPI path.
   - The HOPI VLAN path, through Ethernet switches.
   - The standard Abilene IP routed backbone path.
3. **Understand MPLS as a deterministic path and compare it to Ethernet paths and IP transport across Abilene.** Compare coast-to-coast lightpaths at both the 1 and 10 GigE service levels across multiple paths:
   - Across the HOPI wave
   - Through an MPLS L2VPN tunnel with premium service on Abilene

- Through the normal IP routed infrastructure on Abilene.
- Through other services, including the routed IP premium service and through GRE tunnels.

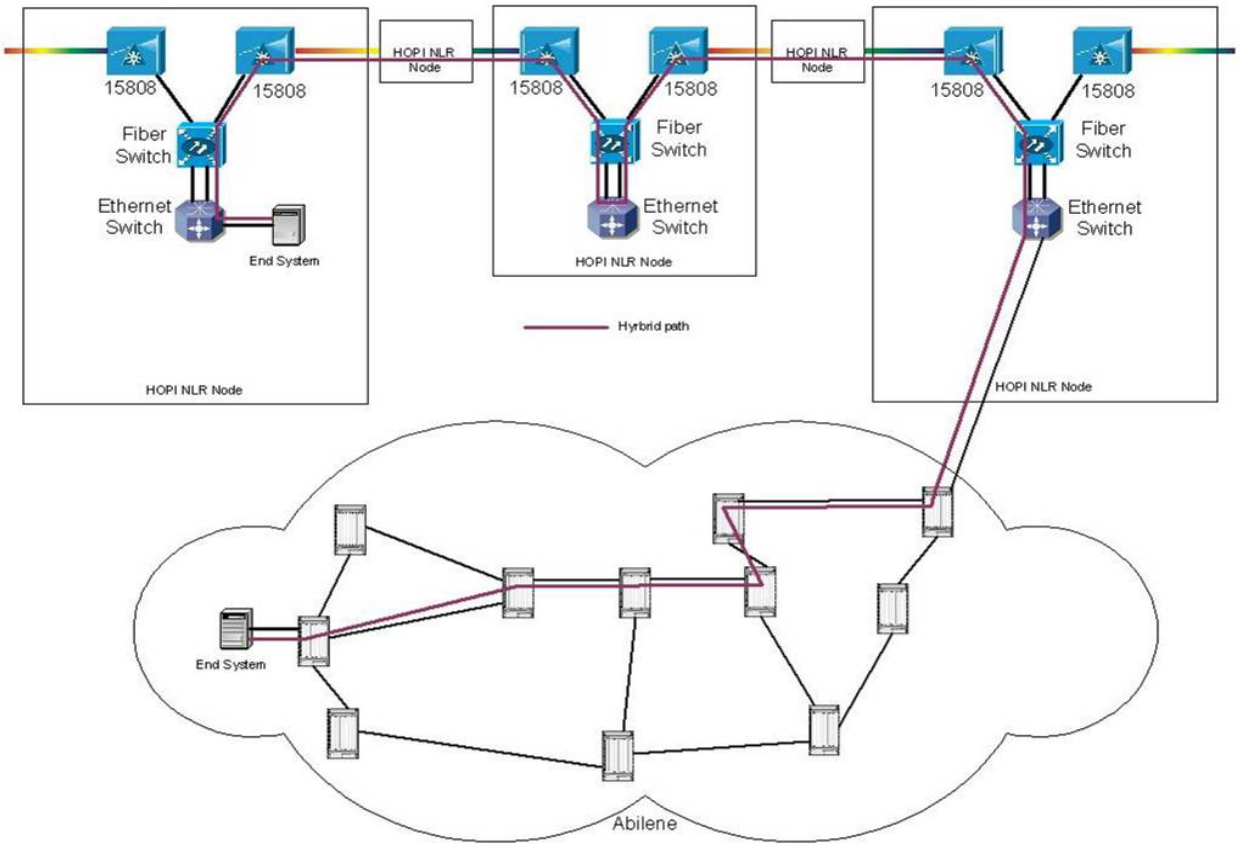
4. **Ethernet Switch on Path Behavior.** Observe 10 unidirectional 1 GigE flows on the 10 GigE HOPI wave. Begin with at least two Ethernet hops and increment by one the number of switches each time. Observe the additive affect that each switch has on the deterministic behavior. This assumes 10 GigE switches are deployed. Compare and contract the results to similar TDM measurements, which may be collected using HOPI infrastructure. This assumes SONET infrastructure is deployed in HOPI.
5. **An Ethernet Switch to TDM Device Comparison.** Evaluate Ethernet Switch Interleaved WFQ implementation to schedule traffic between ingress and egress queues to determine how close this implementation resembles deterministic TDM behavior. Imagine a 10 GigE link configured to support 8-10 1 GigE flows with WFQ sharing bandwidth between them.
6. **General Lightpath Testing and Characterization.** HOPI provides the opportunity to test the performance of several different circuit types of implementations, similar to the particular test cases above. Performance testing should include hardware based traffic generators/test equipment to document fine granularity values for bandwidth, delay, jitter, and loss under the following:

   - varying/multiple bandwidth conditions
   - varying/multiple packet size distributions
   - varying/multiple traffic timing profiles

   The following types of circuits could be constructed using the HOPI resources (See the figure below).

   - Ethernet Framed Wave
   - Ethernet Switched Wave
   - Router Label Switched Path
   - Hybrid Path

   The main objective of this test event would be to develop a detailed analysis and comparison of the different lightpath types. Analysis should include a qualitative and quantitative analysis of performance, instantiation requirements and difficulty, and variables affecting performance.



Miscellaneous Test Cases

1. **Computer Science Testbed.** Create a small test network for the computer science community to experiment with layer3 (or potentially layer2) protocols that is dynamically available under user software control. This is similar to the NOC/ITEC network described above, but focuses on a set of users not necessarily involved with the HOPI or Abilene projects. Determine the necessary user requirements to make this a useful testbed: deterministic paths, layer2 properties, etc.
2. **An IPv6 Multicast Testbed.** Establish a lightpath to carry IPv6 multicast traffic, to keep it isolated from the underlying routers while giving us enough room to run serious bandwidth. Install lightpaths to replace the IPv6-in-IPv4 multicast tunnels between routers, and trim back the endpoints of the paths as larger islands of native IPv6 multicast are created.
3. **Transport-level Gateway.** Create software to implement the notion of a transport-level gateway (TLG). Specifically, a TLG takes a pair of TCP connections, one from host A to the TLG and the other from the TLG to host B, and pipes the data so that the hosts see the appearance of a single TCP connection between host A and host B. The notion can be generalized to k TLGs taking k+1 TCP connections and giving the appearance of a single TCP connection.

   Test the ability of the TLG to 'mask' imperfections (with respect to MTU and loss) in one of the component connections provided that that component has very small latency. Thus, for example, if the first component has a 1-msec RTT but 1500-byte MTU, and small-but-non-miniscule loss and if the second

component has 70-msec RTT but 9000-byte MTU and miniscule loss, then we conjecture that (1) the very small RTT of the first component causes TCP throughput measured across the first component to be good despite its MTU/loss imperfections and (2) the TLG can operate efficiently enough that the performance of the concatenated TCP connection is (nearly) the minimum of the separate performance of the component TCP connections.

To operate at the 1 Gbps will require a 1 GigE path and each TLG to have two GigE interfaces (and the needed I/O and CPU horsepower).

Once the basic idea of the TLG's efficacy is confirmed, apply it to a three-component path, in which the first and last components are imperfect but short and in which the middle component is a clean lightpath. Compare this to a case where the middle component is a GigE-limited path across Abilene.

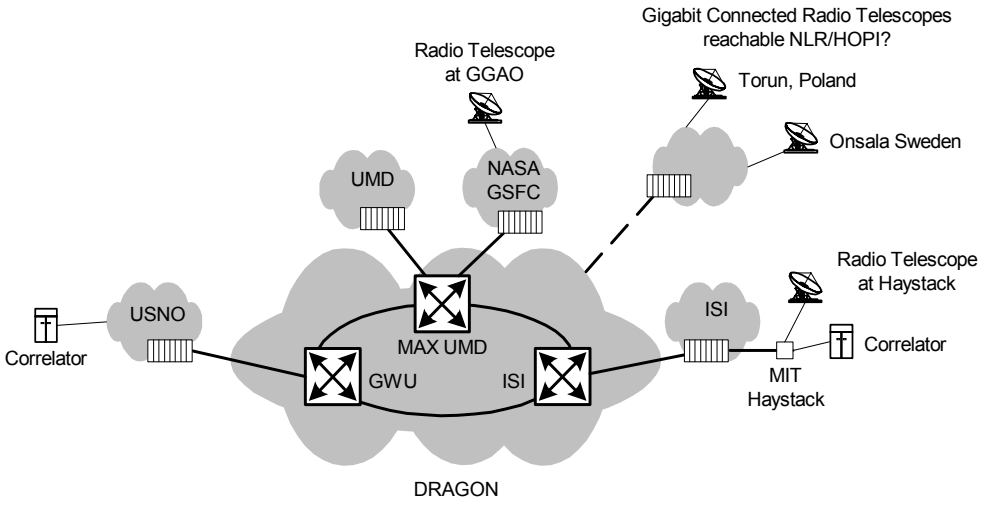
If tests across HOPI and/or Abilene are successful, apply it to a four-component path: <local imperfect> to <US continental HOPI terminating at StarLight> to <DataTAG> to <local imperfect at CERN>.

What makes this a HOPI test is that it tests the ability of very high quality wide-area components to mask imperfect components. This will be very important (at least) to supporting file transfers in the numerous transitions to the great wave future.

Application Based Test Cases

1. **Real-Time electronic Very Long Baseline Interferometry (eVLBI).** The eVLBI project provides an application based test. DRAGON will have Gbps rate connectivity to radio telescopes at NASA GSFC and MIT Haystack Observatory. In addition similar connectivity will be available to the correlator at MIT Haystack. There are also efforts underway to establish a fiber link to the correlator at USNO, and there is a radio telescope in Onsala Sweden connected at Gbps speeds. Another is due to come online in Torun, Poland.

   This test would utilize HOPI resources as part of an effort to be first to accomplish high-bandwidth real-time eVLBI correlation across large distances with multiple telescopes. This would require understanding what kind of connectivity exists between European telescopes and HOPI/NLR international peering points, and connecting DRAGON to HOPI/NLR. A very high level diagram might look something like this:



The test would involve understanding the requirements for deterministic paths associated with eVLBI and also understanding what levels of dynamic control are necessary.

2. **Model Large Hadron Collider (LHC) Networking Requirements.** The High Energy Nuclear Physics (HENP)community will have large bandwidth requirements starting in 2007 when the LHC comes online at CERN. The LHC presents unprecedented computing challenges: (1) massive, globally distributed datasets growing to the 100 petabyte level by the end of the decade; (2) petaflops of similarly distributed computing resources; and (3) collaborative data analysis by global communities of thousands of scientists. To meet these challenges, the LHC collaborations are developing and deploying Grid-based computing infrastructures that provide tremendous computing and storage resources, but rely on the network as an external, passive resource. An essential ingredient will be integrated operation of computational, storage and networking resources, using new tools such as ubiquitous monitoring, dynamic provisioning, optimization and management. It will make possible new discoveries by the global collaborations participating in the decades-long LHC research program. Furthermore, by integrating the results with emerging data-intensive Grid systems, it will help drive the next generation of Grid developments, worldwide monitoring systems and new modes of collaborative work. This test case would attempt to support the LHC project and understand both the deterministic and dynamic requirements needed to support the very large bandwidth needs of the HENP community.