DRAGON

Dynamic Resource Allocation via GMPLS Optical Networks

National Science Foundation

Jerry Sobieski
Mid-Atlantic Crossroads (MAX)

Tom Lehman
University of Southern California
Information Sciences Institute (USC ISI)

Bijan Jabbari
George Mason University (GMU)

Don Riley
University of Maryland (UMD)

DRAGON
Team Members

- Mid-Atlantic CrossRoads (MAX), University of Maryland (UMD)
- University of Southern California Information Sciences Institute (USC/ISI)
- George Mason University (GMU)
- Movaz Networks
- MIT Haystack Observatory
- NASA Goddard Space Flight Center (GSFC)
  - Visualization and Analysis Lab
  - Scientific Visualization Studio
  - Goddard Geophysical and Astronomical Observatory

DRAGON
Team Members

- US Naval Observatory (Wash., DC)
- University of Maryland College Park (UMD)
  - Visualization and Presentation Lab (VPL)
  - University of Maryland Institute for Advanced Computer Studies (UMIACS)
- NCSA (National Center for Supercomputing Applications) ACCESS Center

Funded by National Science Foundation (NSF)

➢Four year project, began in Fall 2003

➢Experimental Infrastructure Network (EIN) Program

DRAGON Mission

- Cyberinfrastructure Application Support
  - Experimental deployment of leading edge network infrastructure directly supporting Cyberinfrastructure e-Science applications
  - Enable new applications capabilities
- Advanced Network Services
  - Develop architectures, protocols and experimental implementations based on emerging standards and technology to provide “advanced” network services
  - Deploy on experimental infrastructure

DRAGON Project
“Advanced Services”

- Dynamic provisioning of deterministic guaranteed resource end-to-end paths
- Rapid provisioning of Application Specific Network topologies
- Reserve resources and topology in advance, instantiate when needed
- Do all this on an Inter-Domain basis with appropriate AAA
- Protocol, format, framing agnostic
  - Direct transmission of HDTV, ethernet, sonet, fibreChannel, or any optical signal

The DRAGON Project
Key Features / Objectives

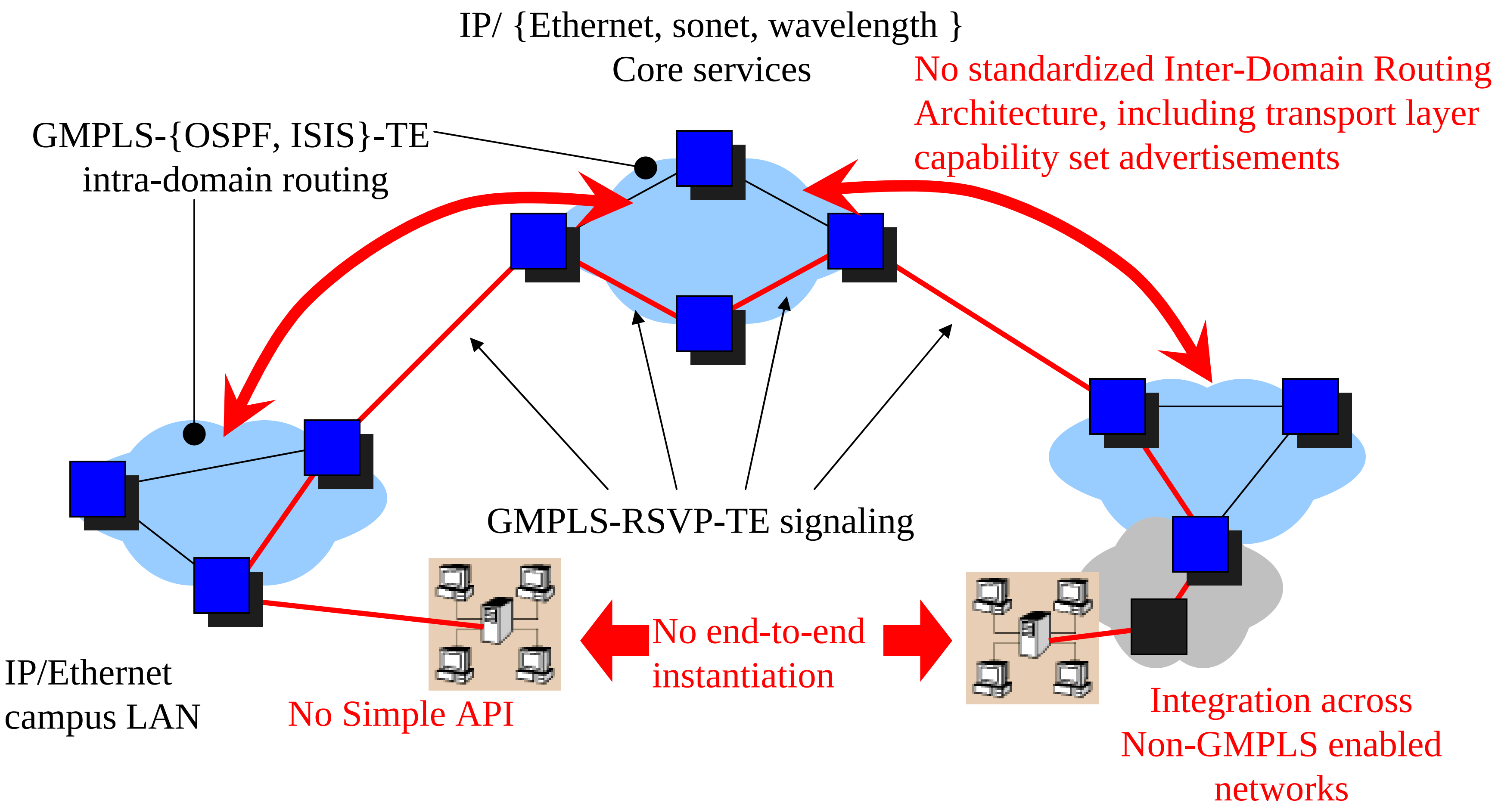
- Uses all optical transport in the metro core
  - Edge to edge Wavelength switching (2R OEO only for signal integrity)
  - Push OEO demarc to the edge, and increasingly out towards end user
- Standardized GMPLS protocols to dynamically provision intra-domain connections
  - GMPLS-OSPF-TE and GMPLS-RSVP-TE
- Develop the inter-domain protocol platform to
  - Distribute Transport Layer Capability Sets (TLCS) across multiple domains
  - Perform E2E path computation
  - Resource authorization, scheduling, and accounting
- Develop the “Virtual LSR”
  - Abstracts non-GMPLS network resources into a GMPLS “virtual LSR”.
- Simplified API
  - Application Specific Topology definition and instantiation
  - Resource resolution, proxy registration and signaling

Technical Issues

GMPLS High Level Overview

- Generalized Multiprotocol Label Switching
  - Evolved from MPLS
  - Defines a set of routing protocol and control plane standards/extensions to instantiate Label Switched Paths (LSPs, ~= “circuits”) through a network
    - GMPLS-{OSPF|ISIS}-TE, GMPLS-{RSVP|CRLDP}-TE
  - Works in conjunction with and complements existing IP network capabilities
- GMPLS defines a number of LSP/circuit types in a logical hierarchical fashion:
  - Fiber->Waves->{Sonet|Ethernet|FC}->...
  - PSC,L2SC,TDM,LSC,FSC label types
- Provides the network the capability to reconfigure topology dynamically
    - Or to address a similar topology requirement of the end systems(s)

End to End GMPLS Transport
What is missing?



IP/ {Ethernet, sonet, wavelength }
Core services

No standardized Inter-Domain Routing Architecture, including transport layer capability set advertisements

GMPLS-{OSPF, ISIS}-TE
intra-domain routing

GMPLS-RSVP-TE signaling

IP/Ethernet
campus LAN

No Simple API

No end-to-end instantiation

Integration across
Non-GMPLS enabled
networks

DRAGON Software Development Components

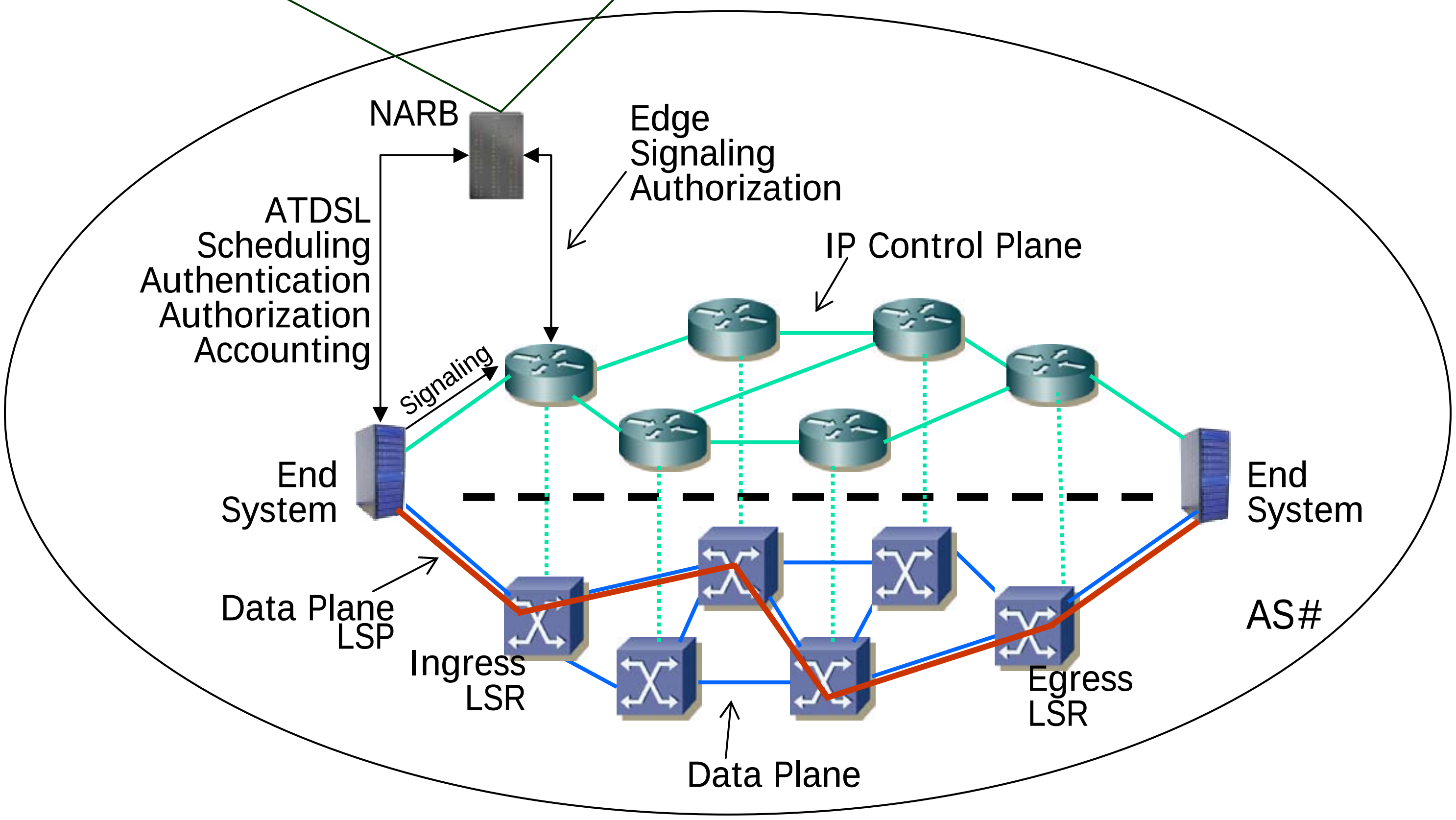
- Network Aware Resource Broker
  - Inter-domain routing platform to advertise Transport Layer Capability Sets (TLCS)
  - Dynamically monitors IGP and/or EGP for network topology changes
- Application Specific Topology Descriptions
  - Ability to request deterministic network resources
- Virtual Label Switched Routers
  - Migration path for non-GMPLS capable network components and proxies for “dumb” network attached devices (e.g. HDTV cameras)
- All Optical End-to-End Routing
  - Minimize OEO requirements for “light paths”

Network Aware Resource Broker (NARB)

- Each NARB instance represents a single Autonomous System (AS)
- Provides services and functions necessary to address many of the “missing capabilities” required for end-to-end GMPLS scheduling and provisioning
  - InterDomain Transport Layer Capability Set (TLCS) exchange and path computation
  - Processing of end system topology requests (based on ATDSL)
  - Authentication, Authorization, and Accounting (AAA)
  - Resource utilization scheduling, monitoring, and enforcement
  - Edge Signaling Authentication and Enforcement
  - Path Computation

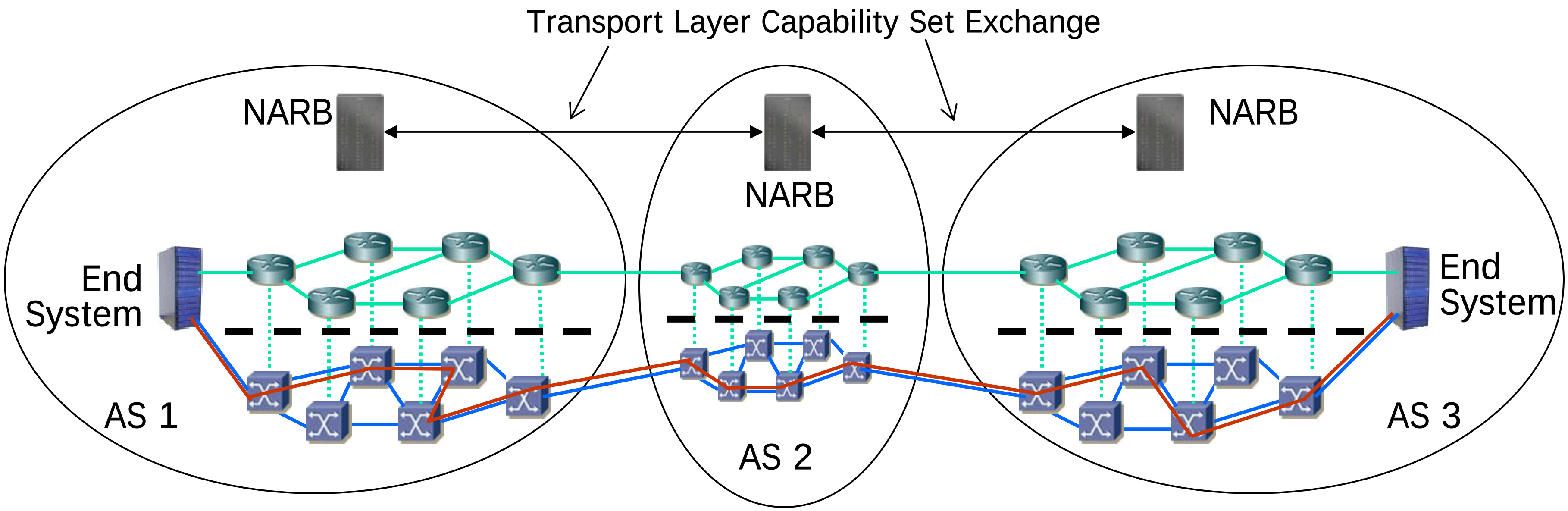
Network Aware Resource Broker (NARB) Functions – IntraDomain

- IGP Listener
- Path Computation
- Scheduling
- Edge Signaling Authentication
- Edge Signaling Enforcement
- ASTDL Induced Topology Computations
- Authorization (flexible policy based)
- Authentication
- Accounting



Network Aware Resource Broker (NARB) Functions - InterDomain

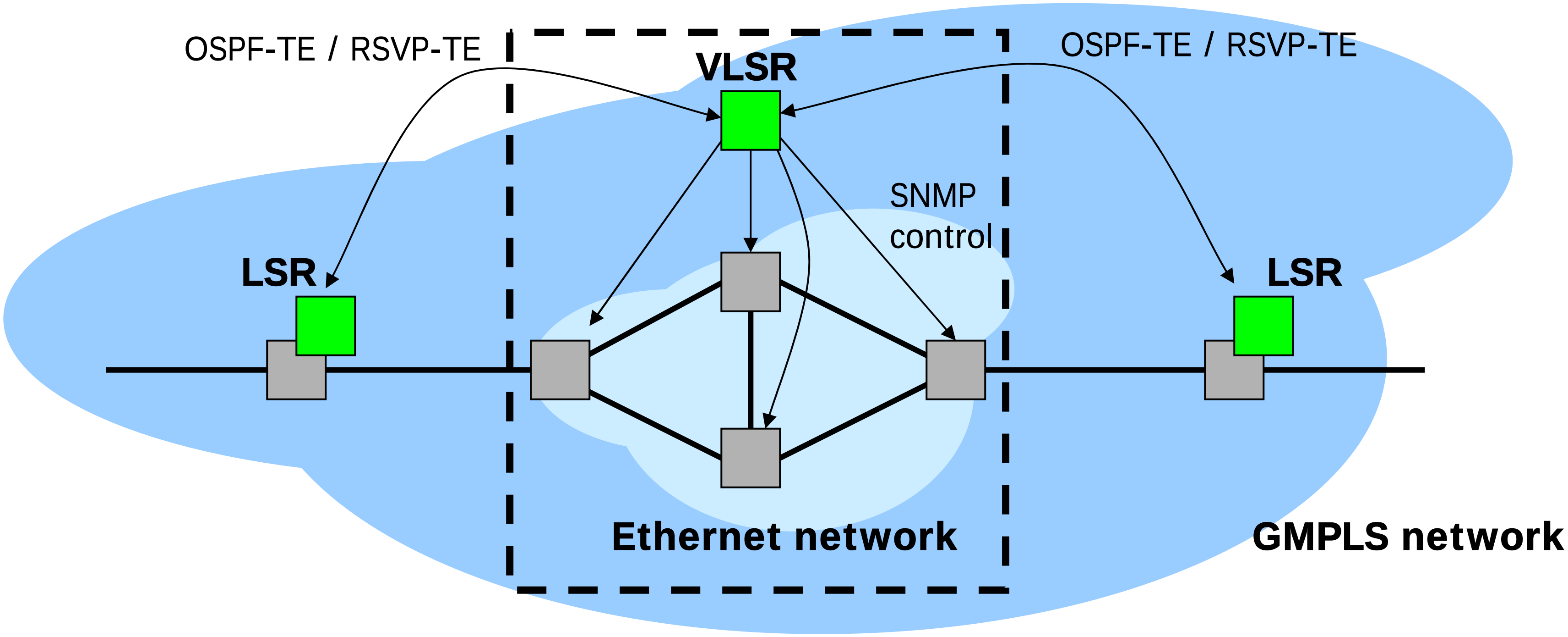
- InterDomain NARB must do all IntraDomain functions plus:
  - EGP Listener
  - Exchange of InterDomain transport layer capability sets
  - InterDomain path calculation
  - InterDomain AAA policy/capability/data exchange and execution



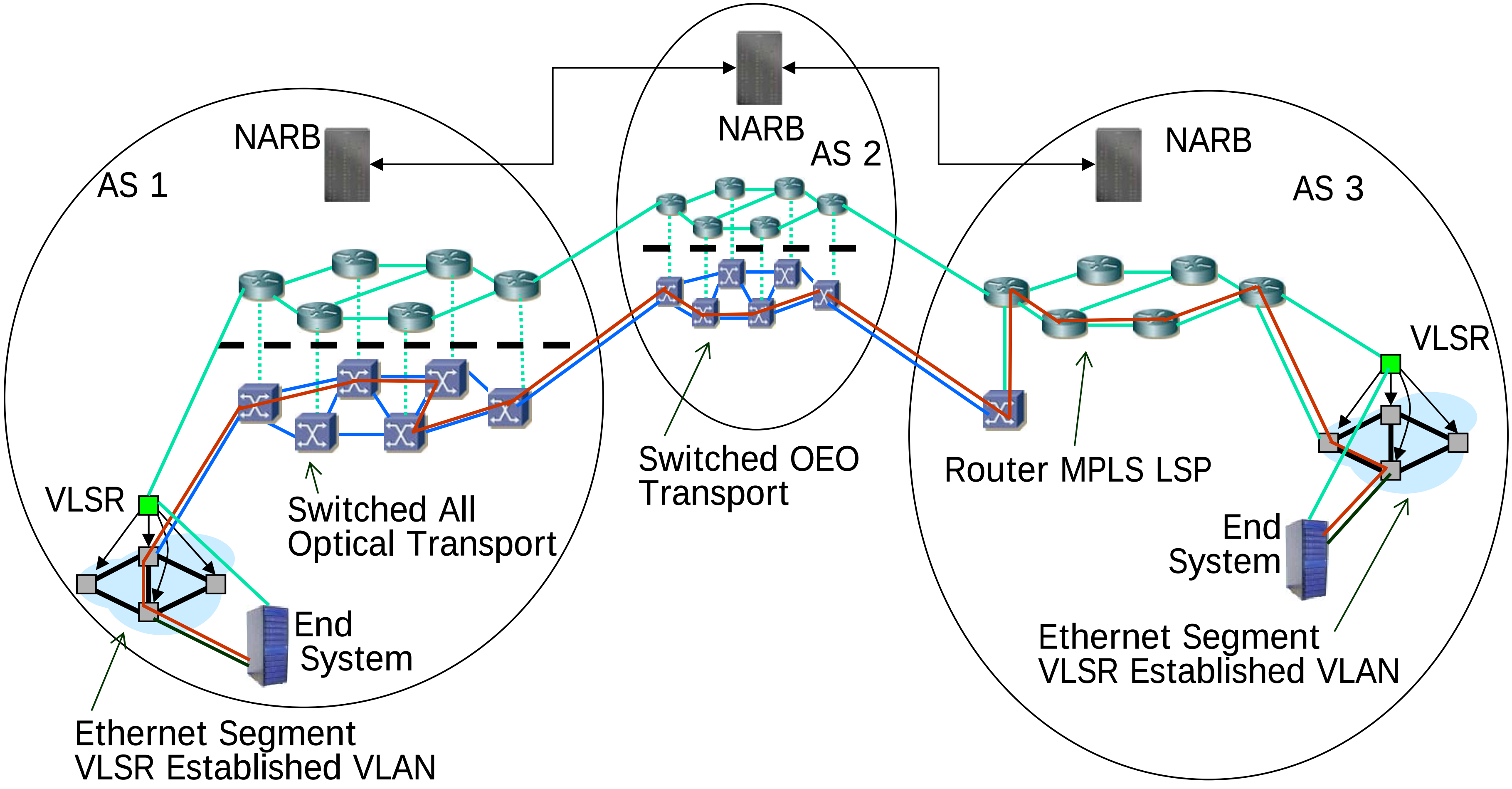
Virtual Label Switched Router - VLSR

- Many networks consist of switching components that do not speak GMPLS, e.g. current ethernet switches, fiber switches, etc
- Contiguous sets of such components can be abstracted into a Virtual Label Switched Router
  - A management agent (the VLSR) can be created that interacts with the DRAGON network via GMPLS protocols
  - The VLSR translates GMPLS resource requests into configuration commands to the covered switches via SNMP or a similar protocol.

VLSR Abstraction



Heterogeneous Network Technologies Complex End to End Paths



Application Specific Topologies

- A formalized definition language to describe and instantiate complex topologies
  - Complex topologies consisting of multiple LSPs must be instantiated as a whole.
  - Resource availability must be predictable, i.e. reservable in advance for utilization at some later time (when necessary)
  - By formally defining the application’s network requirements, service validation and performance verification can be performed (“wizard gap” issues)

Application Specific Topology

- End system facilities necessary to interface the application/user to the network services
- Must be conceptually straight forward for the research user to manipulate network resources
  - The application must be able to create necessary topology in a deterministic manner
  - Such resource provisioning must be compliant with AAA policy – end to end.
  - As much as possible, the api should complement existing standards – e.g. should not break co-resident networking capabilities

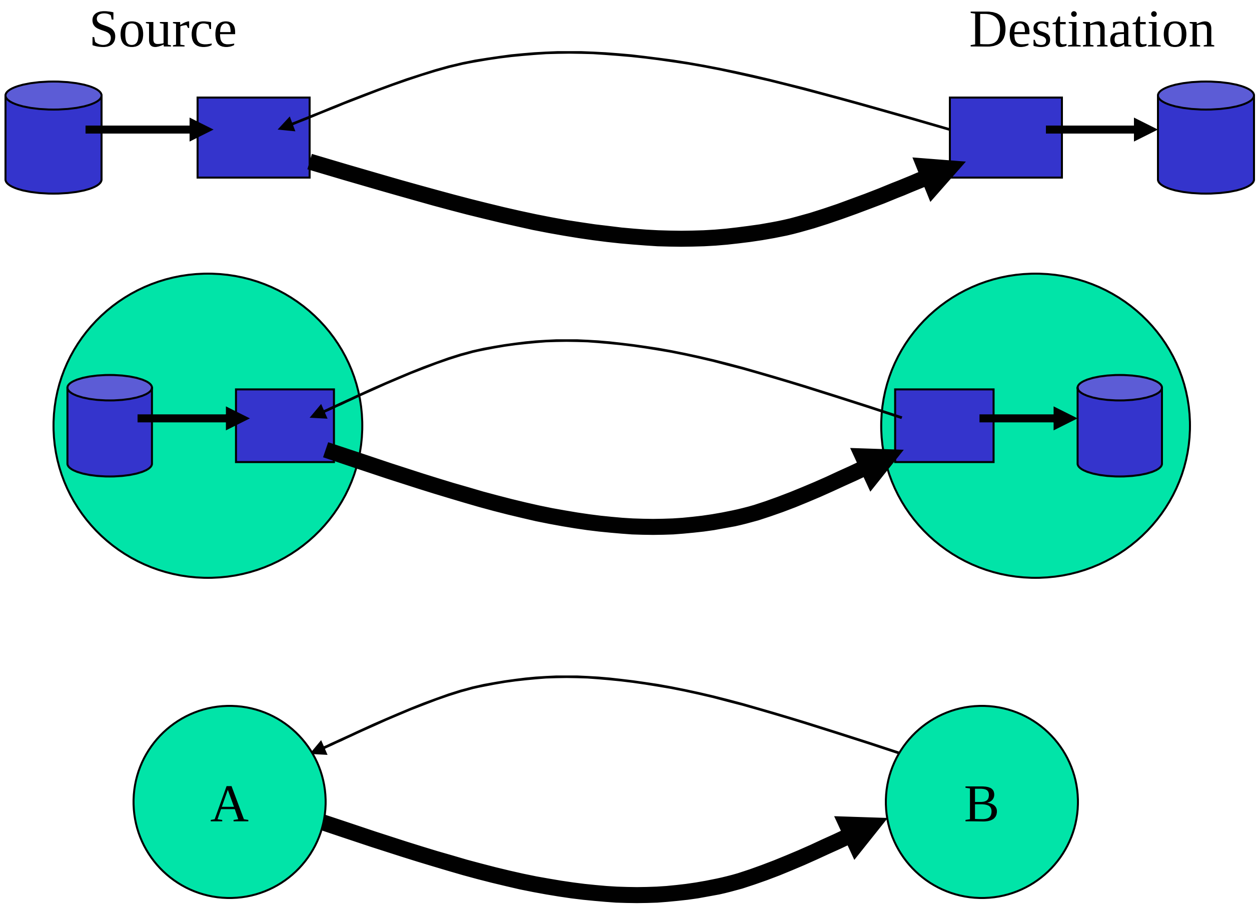
Capabilities

- Textual description of the application VPN
- Text compiled into a NetworkObject
- The NetworkObject is instantiated with specific addresses of nodes
- Nodes are contacted, NetObj instance is flooded, links are signaled.

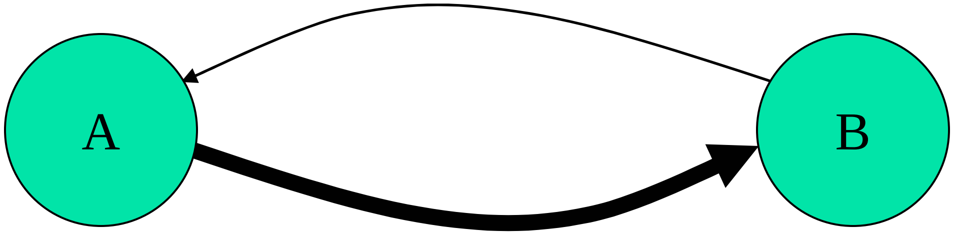
Application Specific Topology Description Language: Purpose

- Formalize the set of network resources, their characteristics and configuration, required by an application to communicate between it's distributed end systems.
  - **What** resources are required, **when** are they needed, **how** will they be accessed, **where** do they need to appear, **who** is authorized to use them
  - **Why** is left to the requester...
- Acquire those resources and present them to the application – simply.

Example 1 – a simple application
File Transfer



Example 1: simple File Transfer



Link from Node A to Node B is 1Gbs reserved
Link from Application B to Application A is 1Kbs best_effort

```
File_transfer_app := {
        node A:={module=/usr/bin/ftp; host=%1; }
        node B:={module=/usr/bin/ftpd; host=%2; }
        link fileflow:={ source=A; destination=B; bw=1gbs;
                         reserved; }
        link ackflow:={ source=B; dest=A; bw=100kbs; }
}
```

ASTDL: Capabilities

- User runtime library that converts a formal definition to internal representation.
  - Default friendly definition text
  - Telecom/system friendly internal topo object
  - CLI for scripting
- User API, which runs at the distributed application nodes, to process the AST Protocol
- Inter-system AST Protocol capabilities:
  - Exchange topology objects
  - Exchange topology management and status information

ASTDL Driver Example

```
#include "class_Topo.h++"
#define DRAGON_TCP_PORT 5555;
//
// User prime mover "ast_master" for AST miniond
//
using namespace std;
using namespace ASTDL;

int main(int argc, char *argv[])
{
  int stat;
  Topo *topo;
  topo = new Topo(argv[1]);
  if(topo == NULL) exit(1);
  stat = topo->Resolve();
  if(stat != 0) exit (2);
  stat = topo->Instantiate();
  if(stat != 0)
    { cout << "Error stat=" << stat << endl; exit(3);}
  stat = topo->UserHandoff();
  stat = topo->Release();
  exit(0)
}
```

Read topology definition source and create the Topo object

Resolve hostnames and other service specific data

Establish all the connections

exec() the user and pass off the connections.

All done. Tear it down.

Object Oriented

- The AST is an Object
  - Graph structure
    - Captures explicit node and link characteristics from the user’s formal definition
    - Unspecified characteristics are defaulted
  - Nodes are objects
    - Represent functional applications components
  - Links are objects
    - Represent connections between nodes

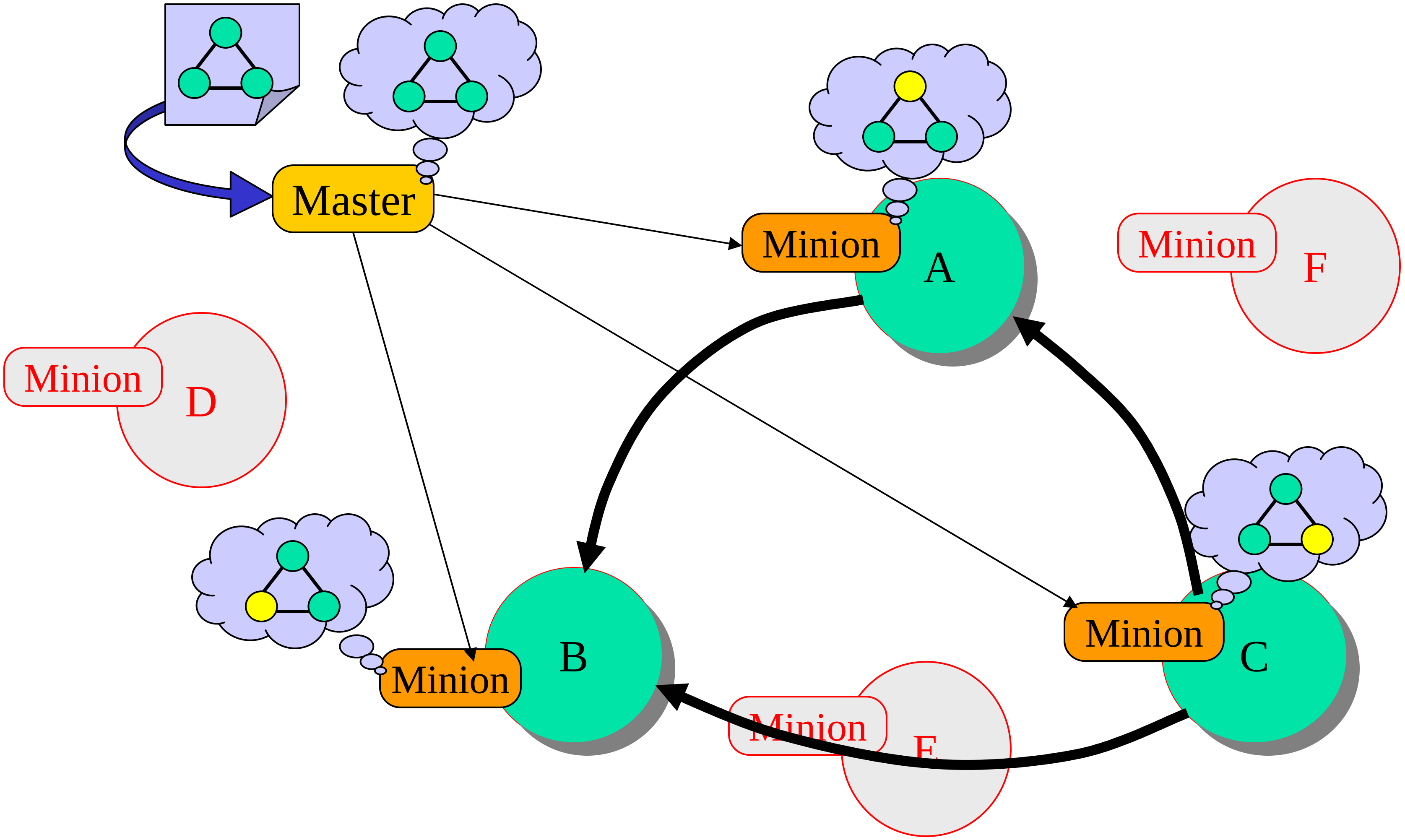
Other Objects

- A “connection” object
  - A connection object specifies two (or more end points)
    - Each endpoint identifies a destination address (IPv4), and a software function that will be the data plane service termination point (STP)
  - Also specifies link characteristics the network will use to instantiate the path
    - E.g. label type, required capacity, routing options (“diverse”), etc

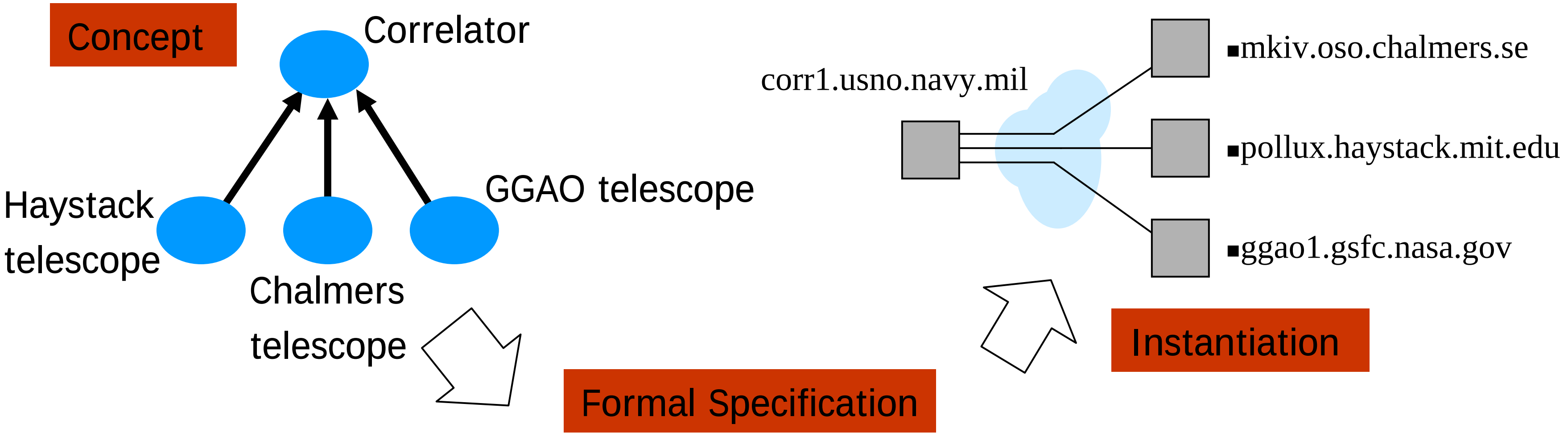
Connection object hierarchy

- Layer4 protocol
  - "stream" := TCP
  - "datagram" := UDP
- Layer3 protocol
  - "IPv4" := IPv4
  - "IPv6" := IPv6
- Layer2.x shim
  - "mpls" := mpls lsp
- Layer2 protocol
  - "ethernet", "sonet", "fiberchannel"
  - Hardware dependent due to layer 2 addressing requirements

The AST Process



Application Specific Topology Description Language - ASTDL



Concept

Correlator

Haystack telescope

Chalmers telescope

GGAO telescope

Formal Specification

corr1.usno.navy.mil

- mkiv.oso.chalmers.se
- pollux.haystack.mit.edu
- ggao1.gsfc.nasa.gov

Instantiation

```
Datalink:= { Type=Ethernet; bandwidth=1g;
                 SourceAddress=%1::vlbid;   DestinationAddress=%2; }
Topo_vlbi_200406 := {
         Correlator:=corr1.usno.navy.mil::vlbid;         // USNO
         DataLink( mkiv.oso.chalmers.se, Correlator );   // OSO Sweden
         DataLink( pollux.haystack.mit.edu, Correlator );// MIT Haystack
         DataLink( ggao1.gsfc.nasa.gov, Correlator );    // NASA Goddard
         }
```

C++ Code invocation example:

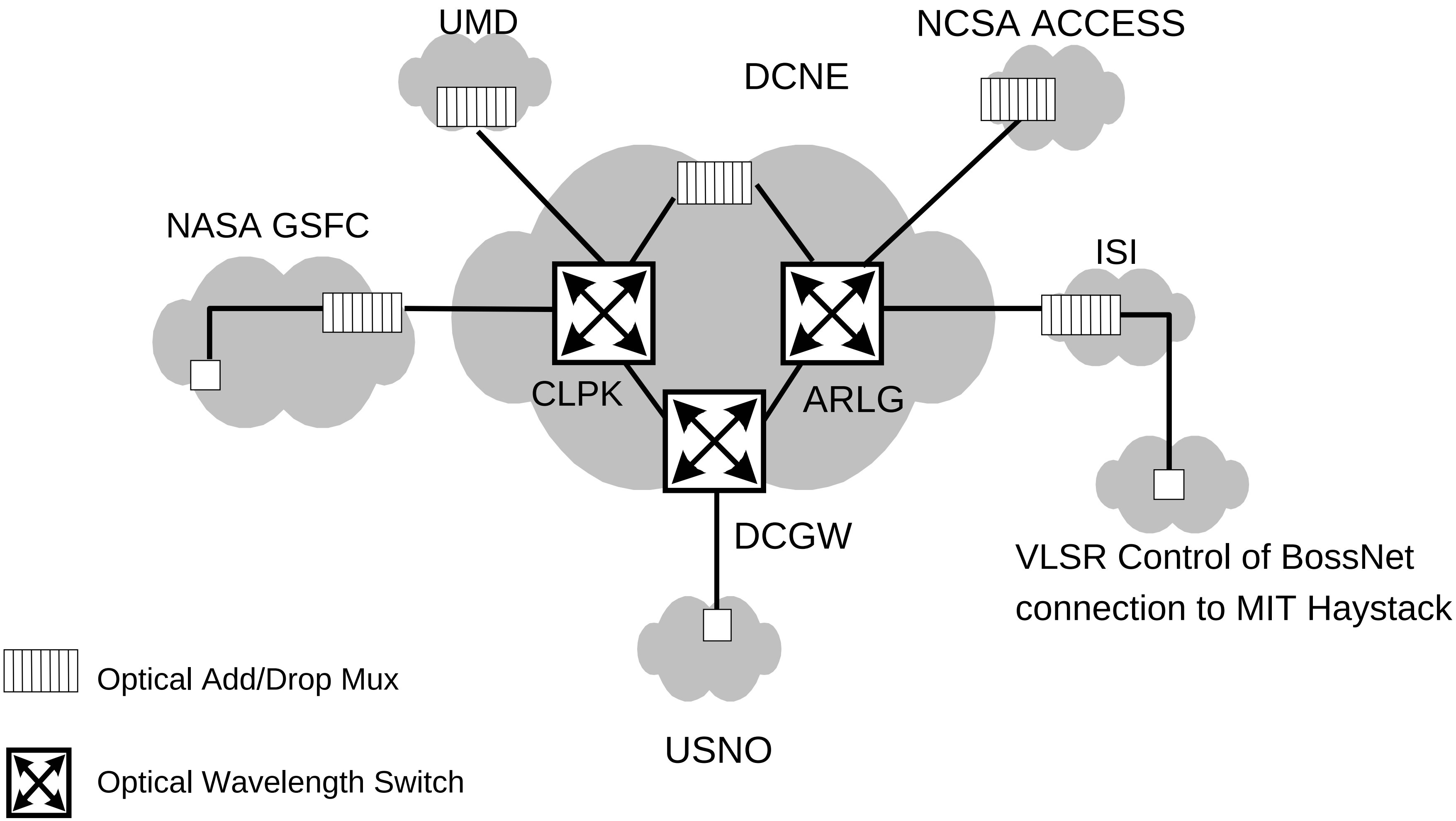
```
eVLBI = new ASTDL::Topo( “Topo_vlbi_200406”);  // Get the topology definition
Stat = eVLBI.Create();                         // Make it so!
```

ASTDL

- It is a Concept – many implementation details yet to be worked out
- ASTDL includes:
  - Interpreters to parse the definition language
  - Runtime libraries accessible by applications
    - Proxy agents to handle non-intelligent devices (e.g. video cameras)
  - Interface protocols to the NARB for ERO computation
  - Resource resolution and scheduling protocols/interfaces
  - Signaling triggers

DRAGON Network

DRAGON Network – Year 3



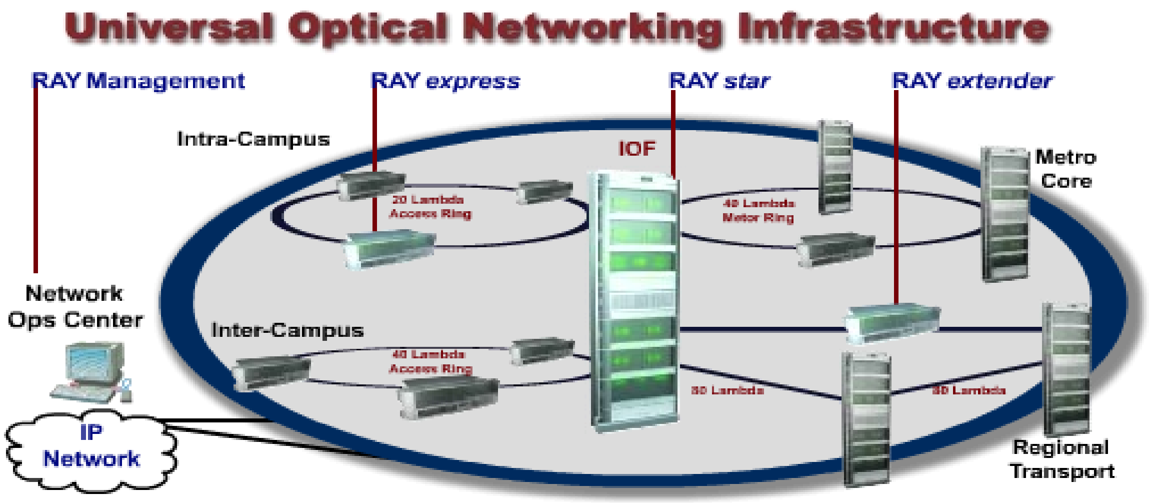
VLSR Control of BossNet
connection to MIT Haystack

▥ Optical Add/Drop Mux

☒ Optical Wavelength Switch

Commercial Partner
Movaz Networks

- Private sector partner for the DRAGON project proposal
  - Provide state of the art optical transport and switching technology
  - Major participant in IETF standards process
  - Software development group located in McLean Va (i.e. within MAX)
  - Demonstrated GMPLS conformance
- Advanced GMPLS optical switching technologies



Commercial Partner
Movaz Networks

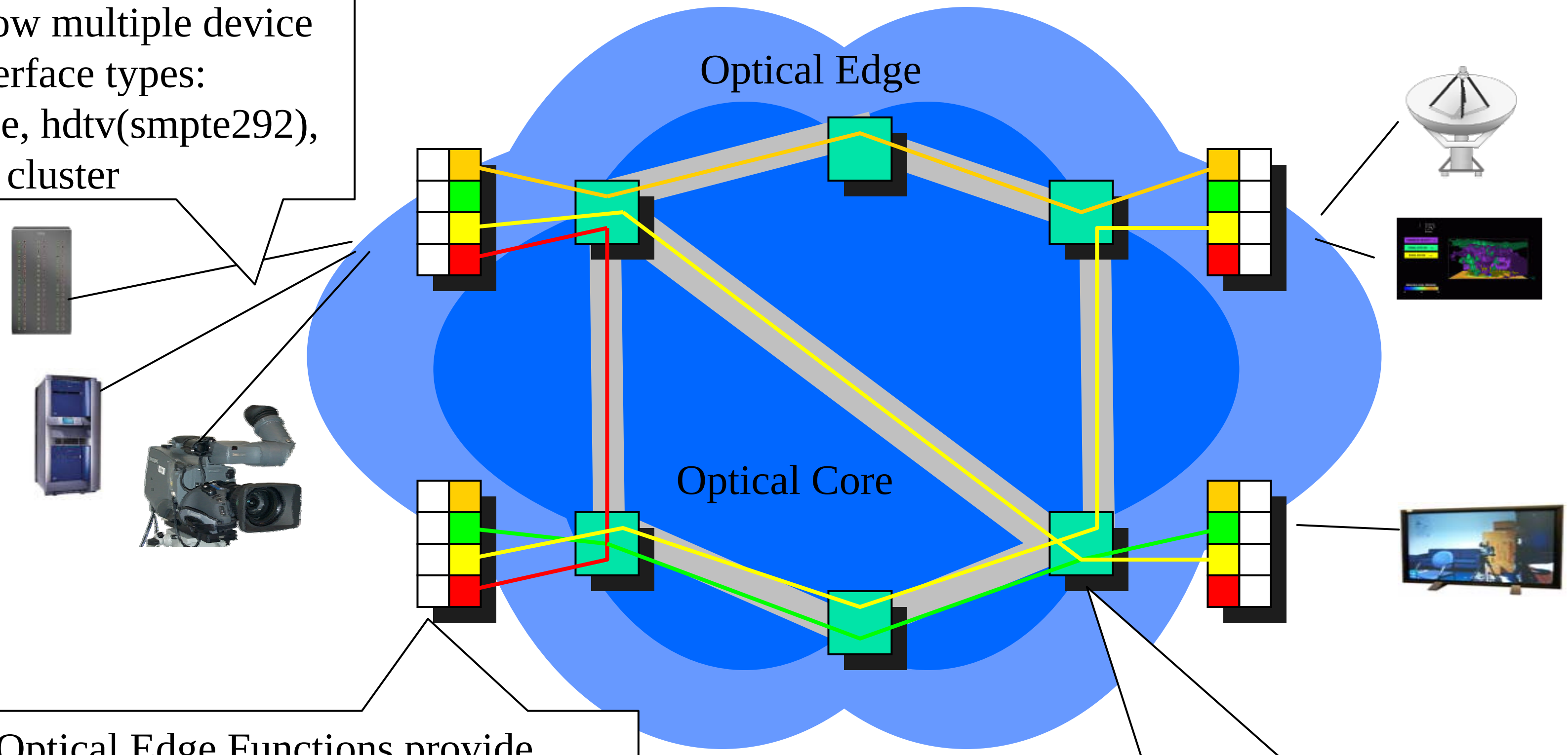
- MEMS-based switching fabric
- 400 x 400 wavelength switching, scalable to 1000s x 1000s
- 9.23"x7.47"x3.28" in size
- Integrated multiplexing and demultiplexing, eliminating the cost and challenge of complex fiber management

- Dynamic power equalization (<1 dB uniformity) eliminating the need for expensive external equalizers
- Ingress and egress fiber channel monitoring outputs to provide sub-microsecond monitoring of channel performance using the OPM
- Switch times < 5ms

New Technology Development and Deployment

- Movaz and DRAGON will be deploying early versions of new technology such as:
  - Tunable wavelength transponders
  - Alien wavelength conditioning
  - 40 gigabit wavelengths
  - RZ encoding (Ultra long haul)
  - Reconfigurable OADMs
- The development and deployment plans of selected technologies will be part of the annual review cycle

All Optical Core and Edge



2R OEO Transponders allow multiple device interface types:
gige, hdtv(smpte292), vis cluster

Optical Edge Functions provide demarc for entry into core
Aggregation
Regeneration
Optical Conditioning

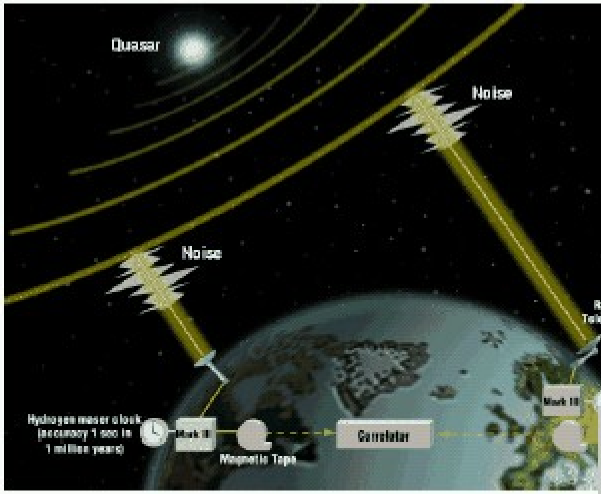
Wavelength switches functioning as optical Label Switched Routers (LSRs) in the core

Routing All Optical Lambdas

- Advantages of “all optical” waves:
  - Framing agnostic: the format of the data modulated onto the wavelength is of no concern (or little concern) to the network
  - Reduced Optical-Electrical-Optical conversion components reducing the cost and complexity of the core
- Challenges
  - Good Optical SNR (I.e. low BER) requires careful attention to fiber engineering, amplification systems, wave band equalization, dispersion management
    - All of these vary with wave path, modulation rates, wavelength, in-path components, etc.
  - Computing optimal paths locally is hard (where you have full visibility of the network characterisitcs)
    - Computing optimal paths across multiple domains is more challenging

Initial Applications

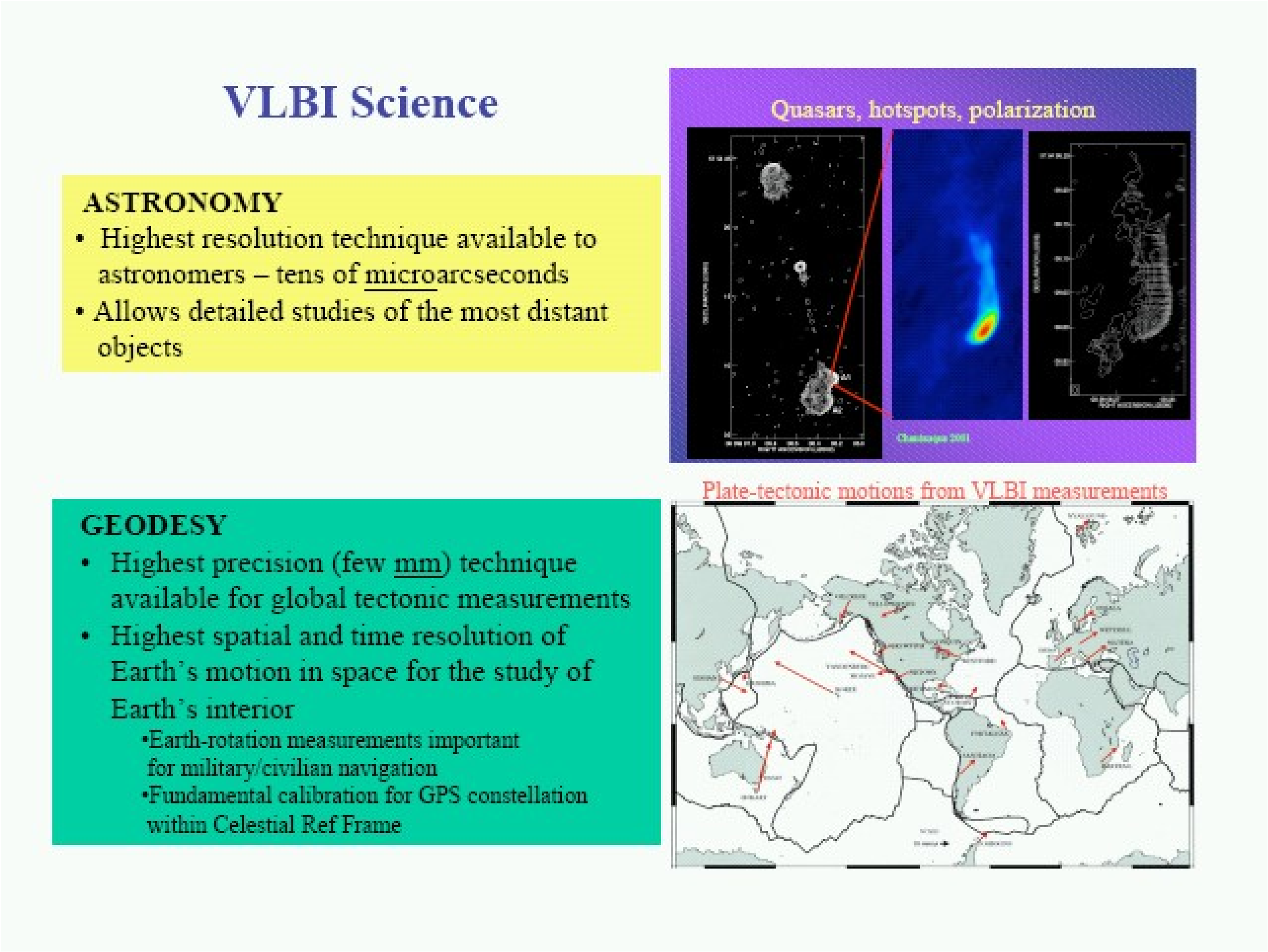
Very Long Baseline Interferometry (VLBI)



The Very-Long Baseline Interferometry (VLBI) Technique
(with traditional data recording on magnetic tape or disk)

The Global VLBI Array
(up to ~20 stations can be used simultaneously)

Very Long Baseline Interferometry (VLBI)



VLBI Science

ASTRONOMY

- Highest resolution technique available to astronomers – tens of microarcseconds
- Allows detailed studies of the most distant objects

GEODESY

- Highest precision (few mm) technique available for global tectonic measurements
- Highest spatial and time resolution of Earth's motion in space for the study of Earth's interior
  - Earth-rotation measurements important for military/civilian navigation
  - Fundamental calibration for GPS constellation within Celestial Ref Frame

Quasars, hotspots, polarization

Plate-tectonic motions from VLBI measurements

electronic-Very Long Baseline Interferometry (e-VLBI)

- What is it?
  - Radio time series are captured simultaneously by several telescopes around the world (the “Very Long” part)
  - These time series are correlated pairwise (the “Baseline”) to identify events occuring within the time series (the “Interferometry”) thus allowing the scientist to calculate very accurately where the event occurred.
  - The traditional method for moving the time series data to the correlator sites has been via tapes and jets.
  - Current methods still incur generally two days to capture and move the data to the correlator – that’s too long.
  - Why are <Realtime | Near RT | on-demand> resources so important to this application?

electronic-Very Long Baseline Interferometry (e-VLBI)

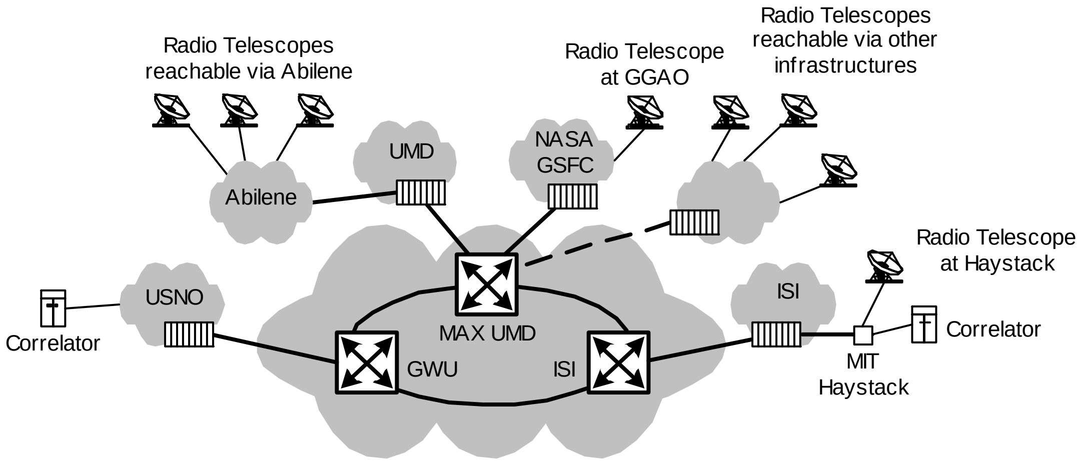
- Because systems such as GPS depend on integrated weather models and VLBI runs, the delay experienced getting the data to the correlators significantly impacts the accuracy and longevity of the predictions:
  - Weather forecast models are typically 5 days
  - Minus 2 days for data transfer = 3 days prediction
  - NRT data transfer will improve predictions by ~40%
- Other celestial events may be transient as well on a scale of minutes to days or weeks
  - Steering the observation NRT allows dramatically more effective use of time on instruments
  - And greatly improved opportunities to acquire useful observations on unpredictable transient events
  - True real-time correlation

electronic-Very Long Baseline Interferometry (e-VLBI)

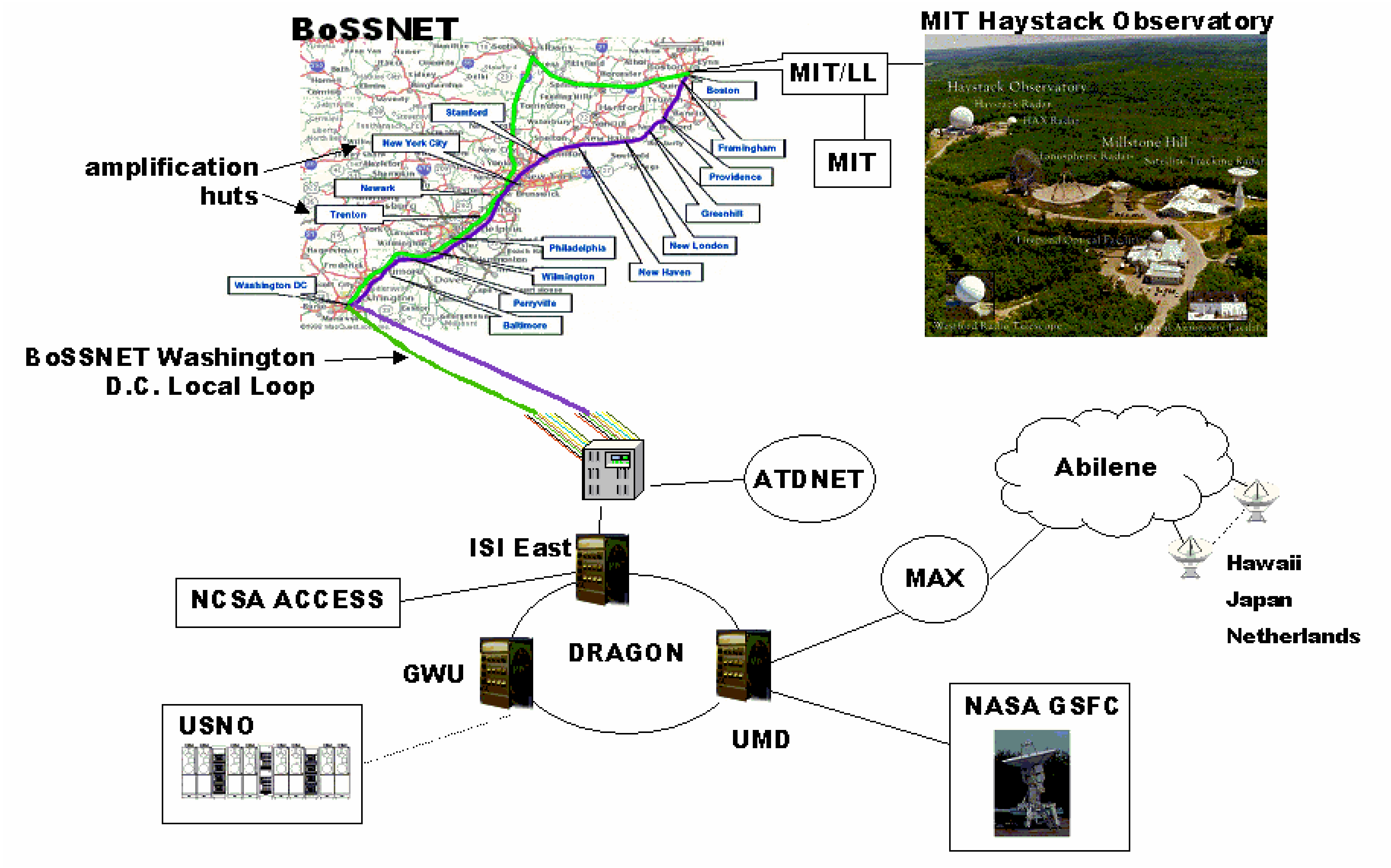
- Why is this such an interesting application to demonstrate dedicated, predictable, high performance network topologies?
  - The VLBI process can be used to study the earth.
    - By focusing on very distant very accurately placed objects, scientists can study the changes in the telescope baselines
    - This can provide very accurate information regarding tectonic plate movement, changes in the earth's shape due to glacial rebound from the ice age, etc.
  - Such changes in the Earth's shape change things like the gravitational effects
  - Most notably, VLBI's ability to detect geodetic wobbles in the earth, allow it to predict small but important perturbations in the inertial frame of reference experienced by satellites
    - The Global Positioning System uses VLBI "intensives" to continually recalculate the satellite orbital positions.
  - Interestingly, these geodetic wobbles can be affected by events such as major atmospheric storms such as typhoons or hurricanes.

electronic-Very Long Baseline Interferometry (e-VLBI)

- electronic-Very Long Baseline Interferometry (e-VLBI)
  - MIT Haystack
  - NASA GSFC (GGAO)
  - USNO
  - Radio Telescopes reachable via Abilene
- eVLBI Experiment Configuration


Radio Telescopes reachable via Abilene
Radio Telescope at GGAO
Radio Telescopes reachable via other infrastructures
UMD
NASA GSFC
Abilene
Radio Telescope at Haystack
USNO
ISI
Correlator
MAX UMD
GWU
ISI
Correlator
MIT Haystack

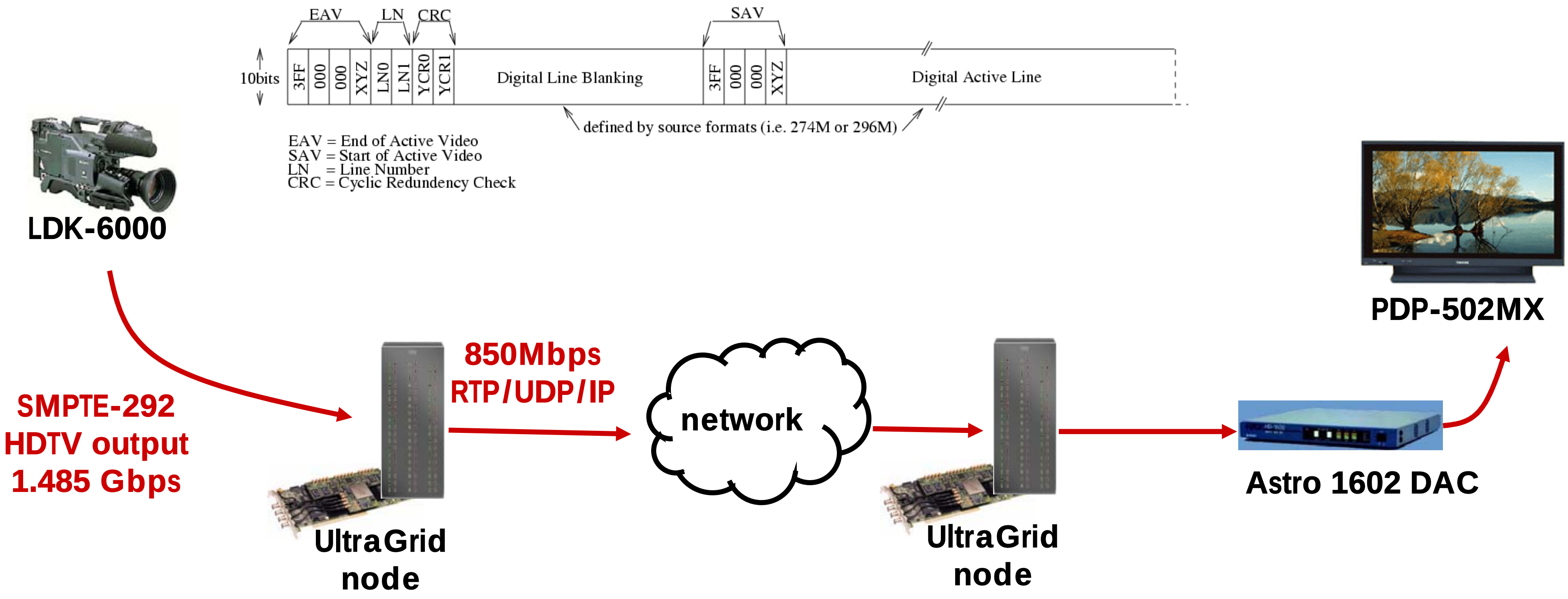
DRAGON eVLBI Experiment Configuration



High Definition Collaboration and Visual Area Networking (HD-CVAN)

- HD-CVAN Collaborators
  - UMD VPL
  - NASA GSFC (VAL and SVS)
  - USC/ISI (UltraGrid Multimedia Laboratory)
  - NCSA ACCESS
- Dragon dynamic resource reservation will be used to instantiate an application specific topology
  - Video directly from HDTV cameras and 3D visualization clusters will be natively distributed across network
- Integration of 3D visualization remote viewing and steering into HD collaboration environments

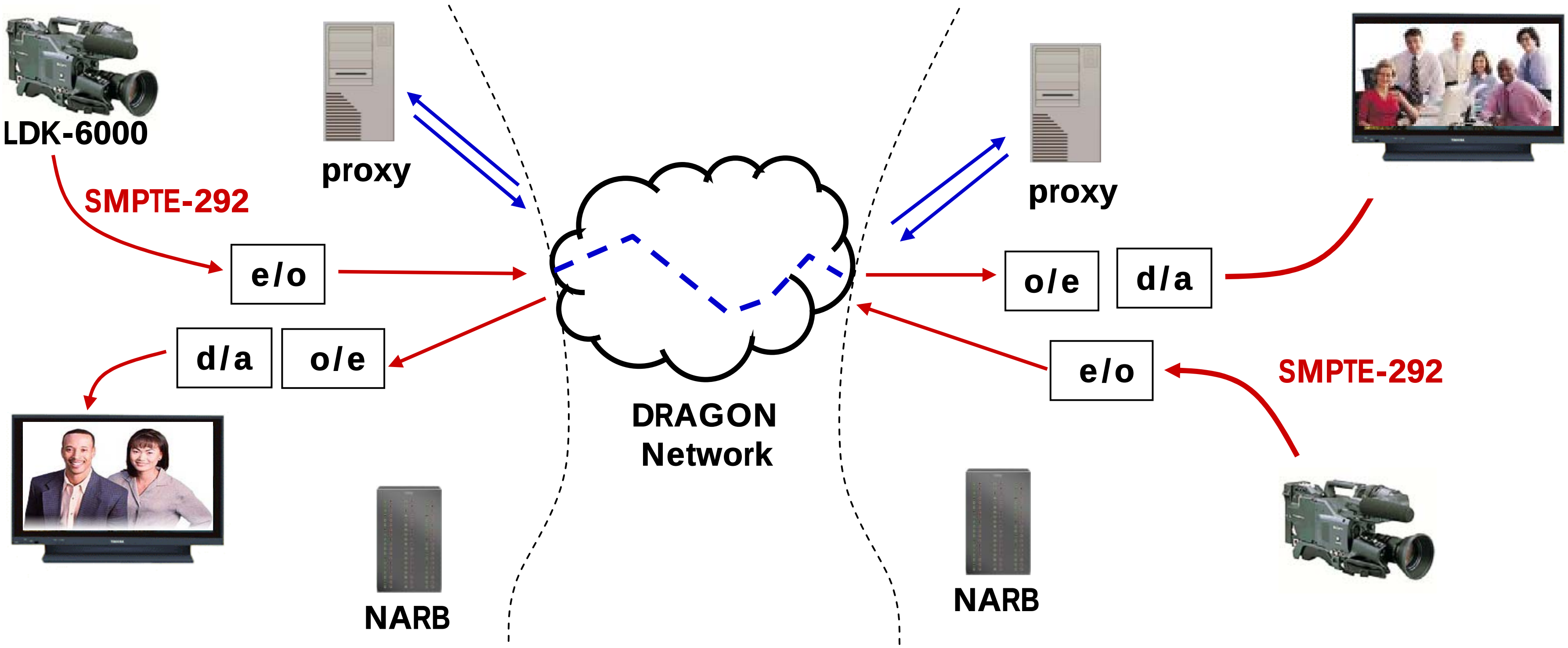
Uncompressed HDTV-over-IP Current Method



EAV = End of Active Video
SAV = Start of Active Video
LN = Line Number
CRC = Cyclic Redundency Check

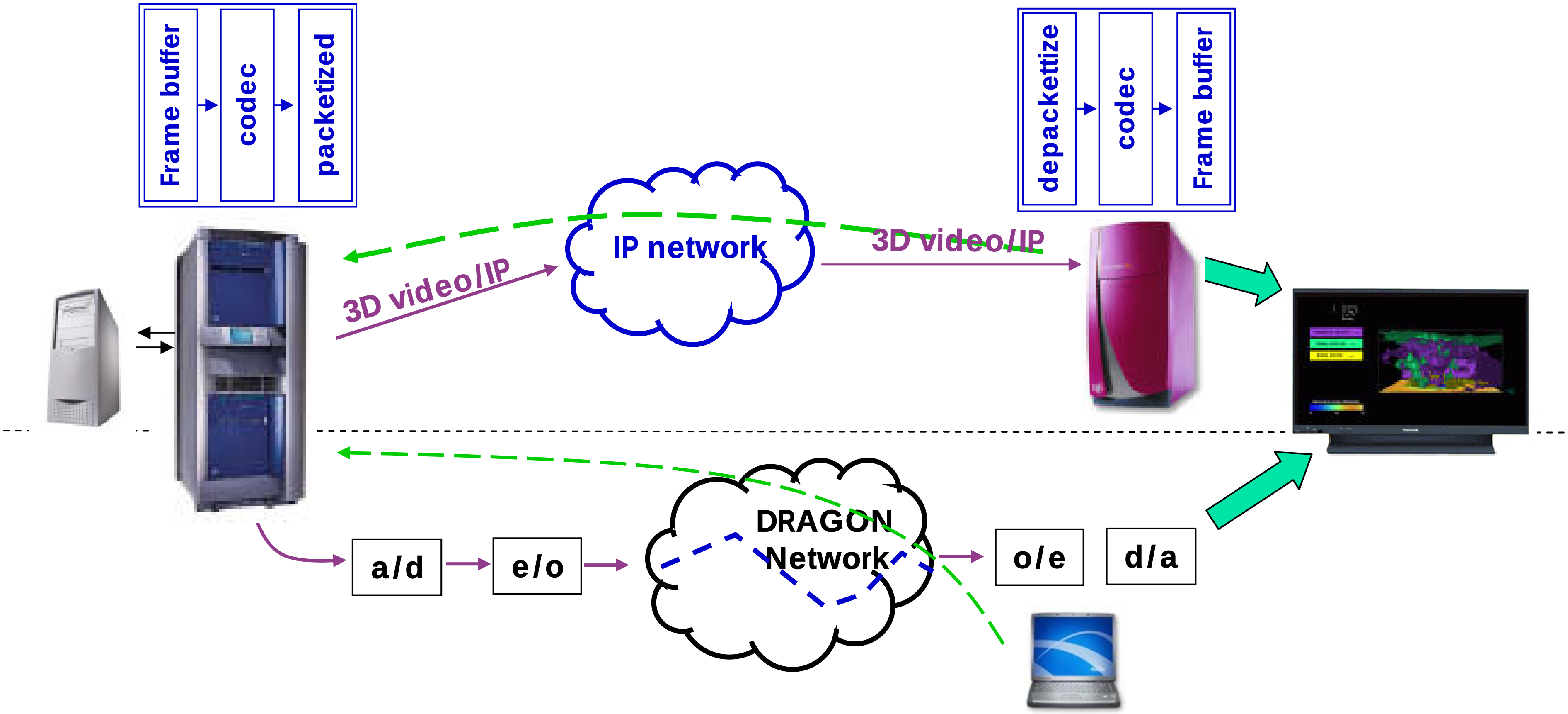
- Not truly HDTV --> color is subsampled to 8bits
- Performance is at the mercy of best-effort IP network
- UltraGrid processing introduces some latency

Low latency High Definition Collaboration DRAGON Enabled



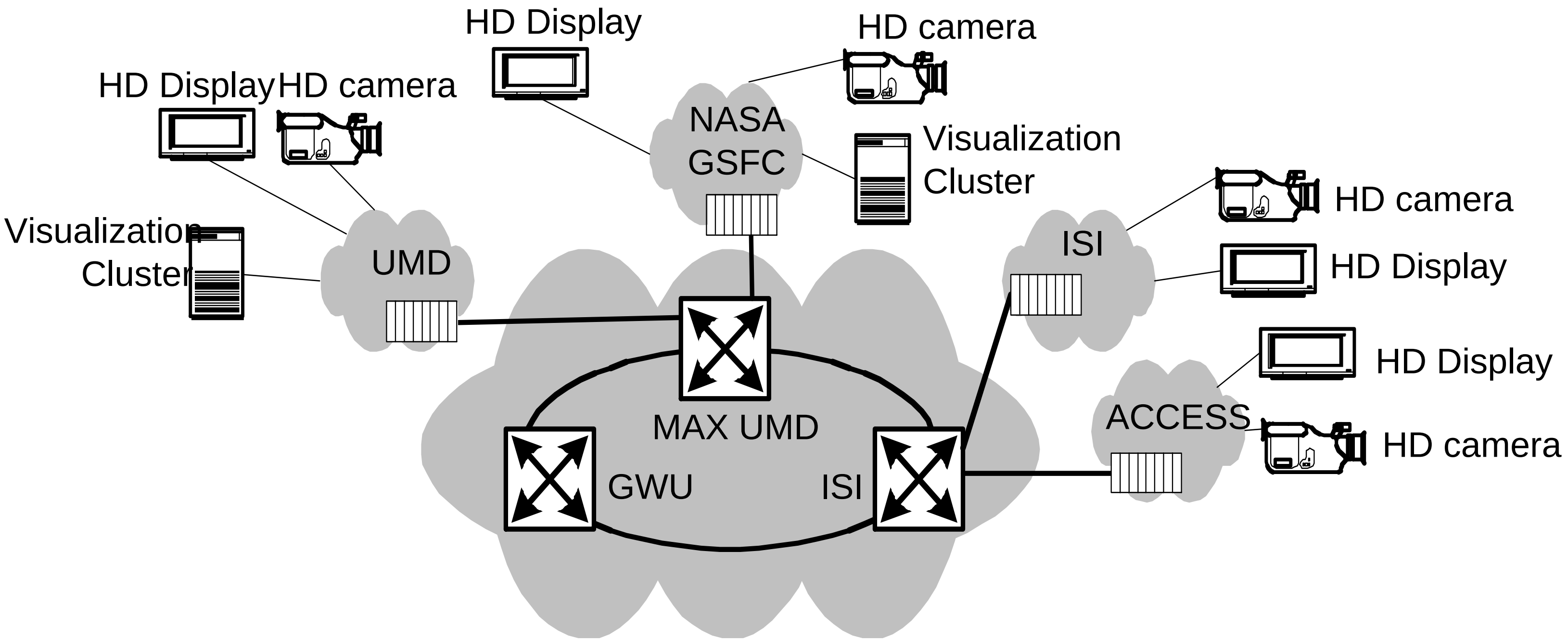
- End-to-end native SMPTE 292M transport
- Media devices are directly integrated into the DRAGON environment via proxy hosts
  - Register the media device (camera, display, …)
  - Sink and source signaling protocols
  - Provide Authentication, authorization and accounting.

Low Latency Visual Area Networking



- Directly share output of visualization systems across high performance networks.
- DRAGON allows elimination of latencies associated with IP transport.

Local HD-CVAN Configuration



- Local HD-CVAN instantiation in the DC metropolitan area:
  - UMD, NASA, ISI and ACCESS

- Create integrated Access Grid style environment with HD conferencing and remote 2D/3D visualization.