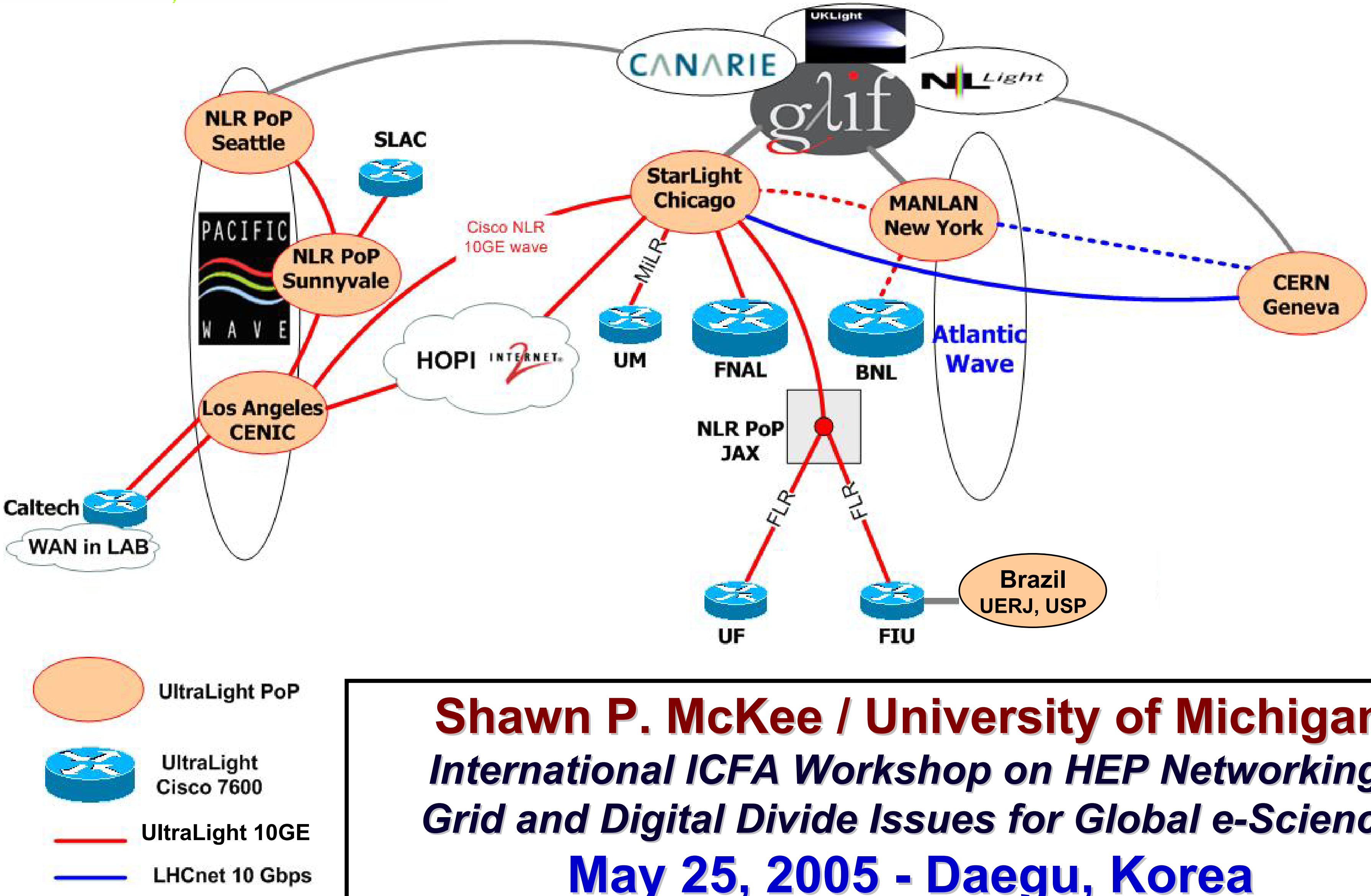
UltraLight: Overview and Status



Shawn P. McKee / University of Michigan

International ICFA Workshop on HEP Networking, Grid and Digital Divide Issues for Global e-Science

May 25, 2005 - Daegu, Korea

The UltraLight Project

UltraLight is

- A four year $2M NSF ITR funded by MPS
- Application driven Network R&D
- A collaboration of BNL, Caltech, CERN, Florida, FIU, FNAL, Internet2, Michigan, MIT, SLAC

Two Primary, Synergistic Activities

- Network "Backbone": Perform network R&D / engineering
- Applications "Driver": System Services R&D / engineering

Goal: Enable the network as a managed resource

Meta-Goal: *Enable physics analysis and discoveries which could not otherwise be achieved*

Impact on the Science

http://www.ultralight.org

- **Flagship Applications: HENP, eVLBI, Biomedical "Burst" Imaging**
- **Use the new capability to create prototype computing infrastructures for CMS and ATLAS**
  - **Extend and enhance their Grid infrastructures, promoting the network to an active, managed element**
  - **Exploiting the MonALISA, GEMS, GAE & Sphinx Infrastr.**
  - **Enable Massive Data Transfers to Coexist with Production Traffic and Real-Time Collaborative Streams**
  - **Support "Terabyte-scale" Data Transactions (Min. not Hrs.)**
- **Extend Real-Time eVLBI to the 10 – 100 Gbps Range**
  - **Higher Resolution (Greater Physics Reach), Faster Turnaround**
  - **New RT Correlation Approach for Important Astrophysical Events, Such As Gamma Ray Bursts or Supernovae**

Project Coordination Plan

- H. Newman, Director & PI
- Steering Group: Avery, Bunn, McKee, Summerhill, Ibarra, Whitney
- TC Group: HN + TC Leaders: Cavanaugh, McKee, Kramer, Van Lingen; Summerhill, Ravot, Legrand

PROJECT Coordination

1. Rick Cavanaugh, Project Coordinator
   - Overall Coordination and Deliverables.
2. Shawn McKee, Network Technical Coordinator
   - Coord. of Building the UltraLight Network Infrastructure (with Ravot)
3. Frank Van Lingen, Application Services Technical Coordinator
   - Coord. Building the System of Grid- and Web-Services for applications (with Legrand)
4. Laird Kramer, Education and Outreach Coordinator
   - Develop learning sessions and collaborative class and research projects for undergraduates and high school students

- EXTERNAL ADVISORY BOARD: Being formed.

UltraLight Backbone

- UltraLight has a non-standard core network with dynamic links and varying bandwidth inter-connecting our nodes.
  - Optical Hybrid Global Network
- The core of UltraLight will dynamically evolve as function of available resources on other backbones such as NLR, HOPI, Abilene or ESnet.
- The main resources for UltraLight:
  - LHCnet (IP, L2VPN, CCC)
  - Abilene (IP, L2VPN)
  - ESnet (IP, L2VPN)
  - Cisco NLR wave (Ethernet)
  - HOPI NLR waves (Ethernet; provisioned on demand)
  - UltraLight nodes: Caltech, SLAC, FNAL, UF, UM, StarLight, CENIC PoP at LA, CERN

International Partners

One of the UltraLight program's strengths is the large number of important international partners we have:

- AMPATH http://www.ampath.net
- AARNet http://www.aarnet.edu.au
- Brazil/UERJ/USP http://www.hepgridbrazil.uerj.br/
- CA*net4 http://www.canarie.ca/canet4
- GLORIAD http://www.gloriad.org/
- IEEAF http://www.ieeaf.org/
- Korea/KOREN http://www.koren21.net/eng/network/domestic.php
- NetherLight http://www.surfnet.nl/info/innovatie/netherlight/home.jsp
- UKLight http://www.uklight.ac.uk/

As well as collaborators from China, Japan and Taiwan.

UltraLight is well positioned to develop and coordinate global advances to networks for LHC Physics

UltraLight Sites

UltraLight currently has 10 participating core sites (shown alphabetically)

The table provides a quick summary of the near term connectivity plans

Details and diagrams for each site and its regional networks are shown in our technical report

Site	Date	Type	Storage	Out of Band
BNL	March	OC48	TBD	TBD
Caltech	January	10 GE	1 TB May	Y
CERN	January	OC192	9 TB July	Y
FIU	January	OC12	TBD	Y
FNAL	March	10 GE	TBD	TBD
I2	March	MPLS L2VPN	TBD	TBD
MIT	May	OC48	TBD	TBD
SLAC	July	10 GE	TBD	TBD
UF	February	10 GE	1 TB May	Y
UM	April	10 GE	9 TB July	Y

Workplan/Phased Deployment

UltraLight envisions a 4 year program to deliver a new, high-performance, network-integrated infrastructure:

Phase I will last 12 months and focus on *deploying the initial network infrastructure* and *bringing up first services* (Note: we are well on our way, the network is almost up and the first services are being deployed)

Phase II will last 18 months and concentrate on *implementing all the needed services* and *extending the infrastructure to additional sites* (We are entering this phase now)

Phase III will complete UltraLight and last 18 months. The focus will be on a *transition to production in support of LHC Physics; + eVLBI Astronomy*

UltraLight Network: PHASE I

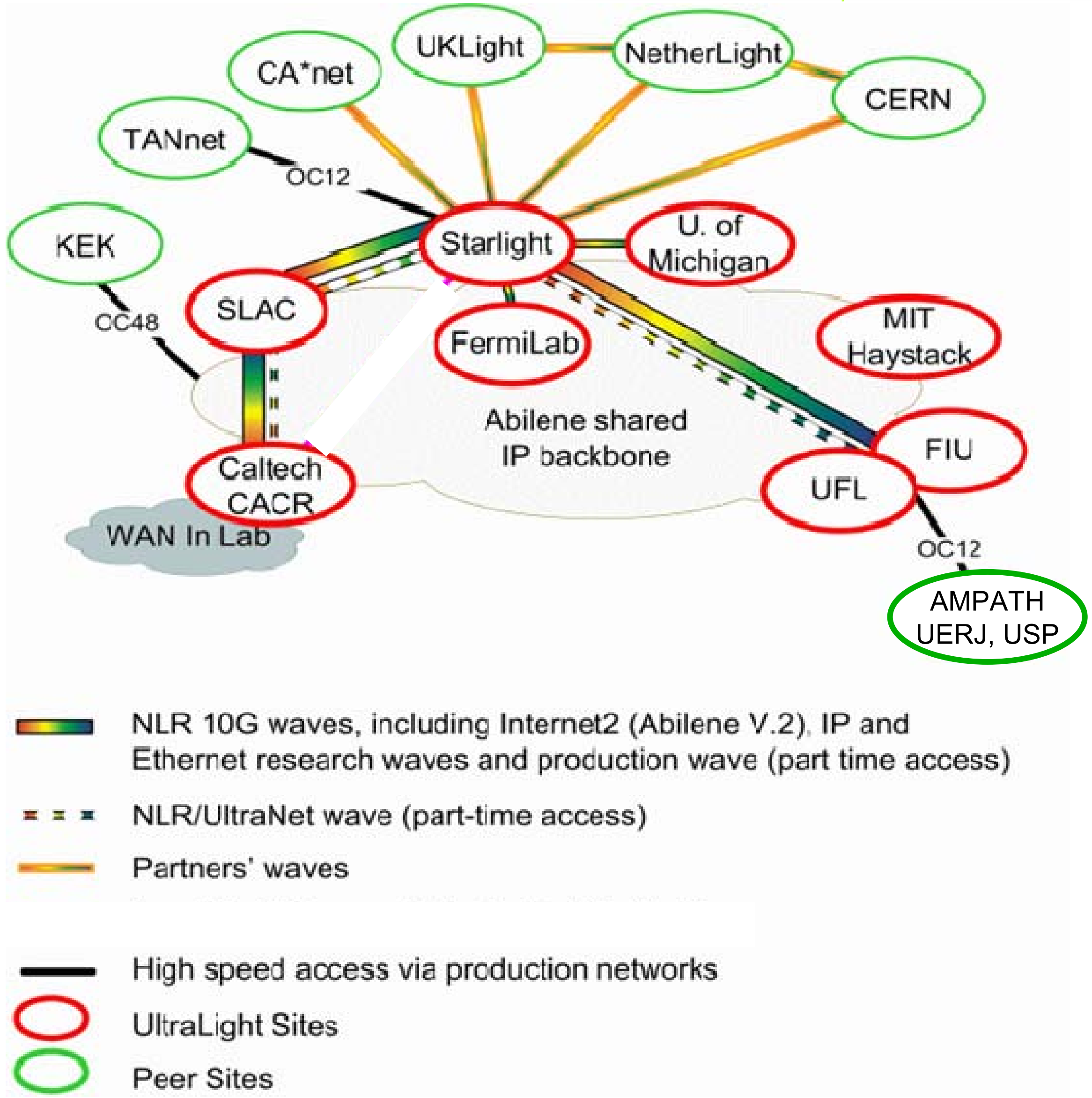
Implementation via “sharing” with HOPI/NLR

Also LA-CHI Cisco/NLR Research Wave

DOE UltraScienceNet Wave SNV-CHI (LambdaStation)

Connectivity to FLR to be determined

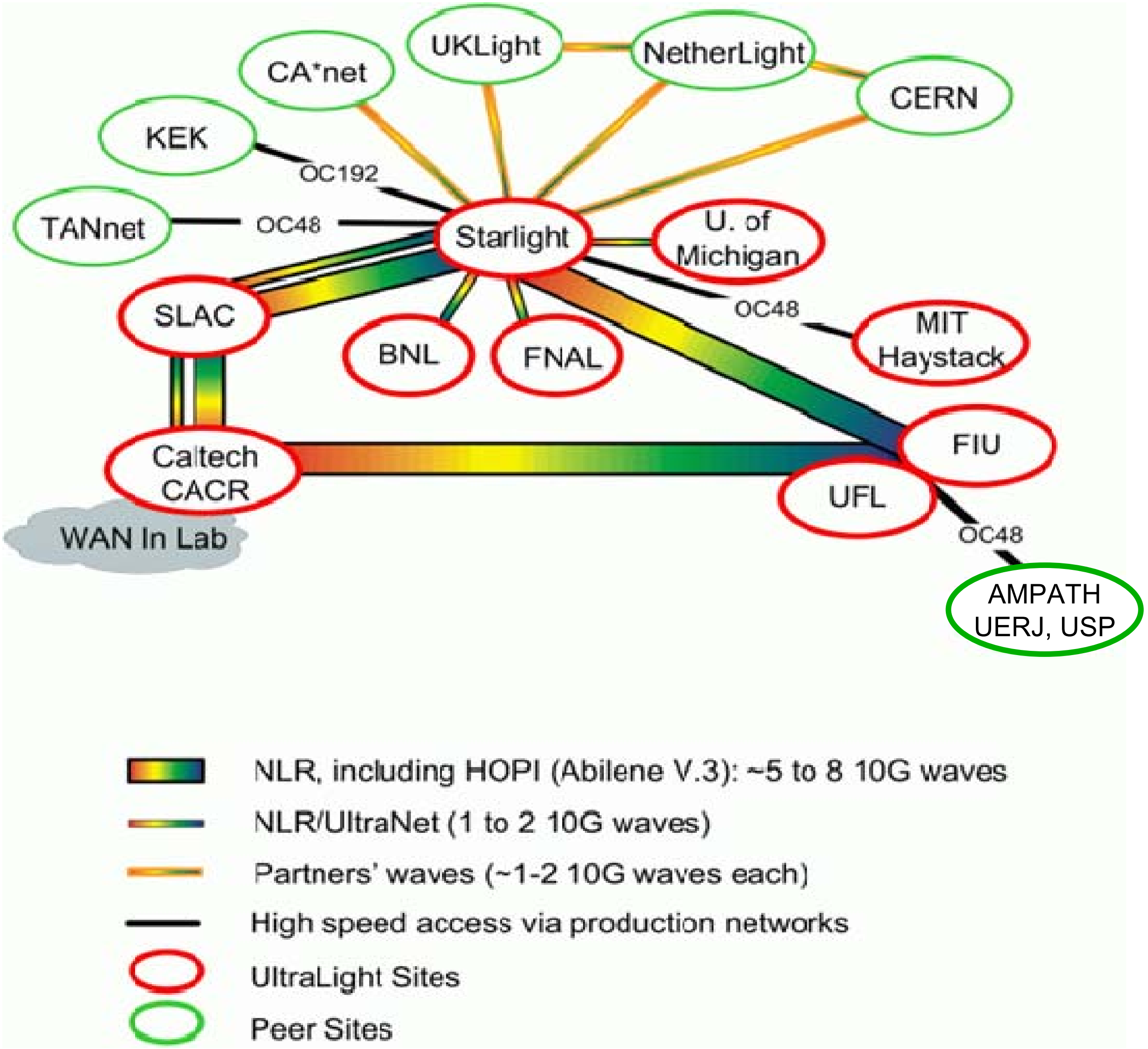
MIT involvement welcome, but unfunded



UltraLight Network: PHASE II

Move toward multiple "lambdas"

Bring in FLR, as well as BNL (and MIT)

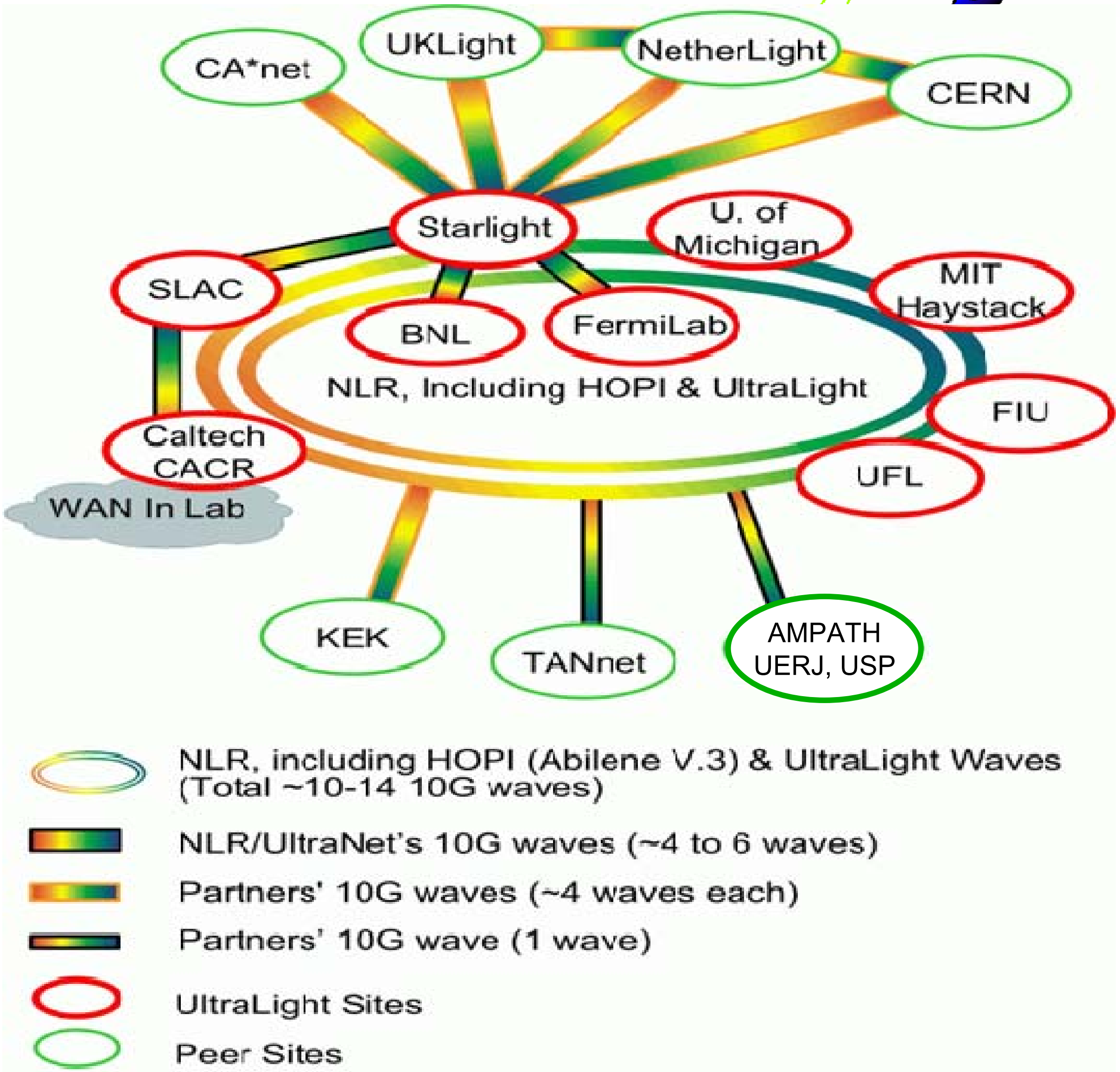


UltraLight Network: PHASE III

Move into production

Optical switching fully enabled amongst primary sites

Integrated international infrastructure



UltraLight Network Engineering

- ***GOAL:*** **Determine an effective mix of bandwidth-management techniques for this application-space, particularly:**
  - **Best-effort/"scavenger" using effective ultrascale protocols**
  - **MPLS with QOS-enabled packet switching**
  - **Dedicated paths arranged with TL1 commands, GMPLS**
- ***PLAN: Develop, Test the most cost-effective integrated combination of network technologies on our unique testbed:***

1. **Exercise UltraLight applications on NLR, Abilene and campus networks, as well as LHCNet, and our international partners**
   - ***Progressively enhance Abilene with QOS support to protect production traffic***
   - ***Incorporate emerging NLR and RON-based lightpath and lambda facilities***
2. **Deploy and systematically study ultrascale protocol stacks (such as FAST) addressing issues of performance & fairness**
3. **Use MPLS/QoS and other forms of BW management, and adjustments of optical paths, to optimize end-to-end performance among a set of virtualized disk servers**

UltraLight: Effective Protocols

The protocols used to reliably move data are a critical component of Physics "end-to-end" use of the network

TCP is the most widely used protocol for reliable data transport, but is becoming ever more ineffective for higher and higher bandwidth-delay networks.

UltraLight will explore extensions to TCP (HSTCP, Westwood+, HTCP, FAST) designed to maintain fair-sharing of networks and, at the same time, to allow efficient, effective use of these networks.

UltraLight plans to identify the most effective fair protocol and implement it in support of our "Best Effort" network components.

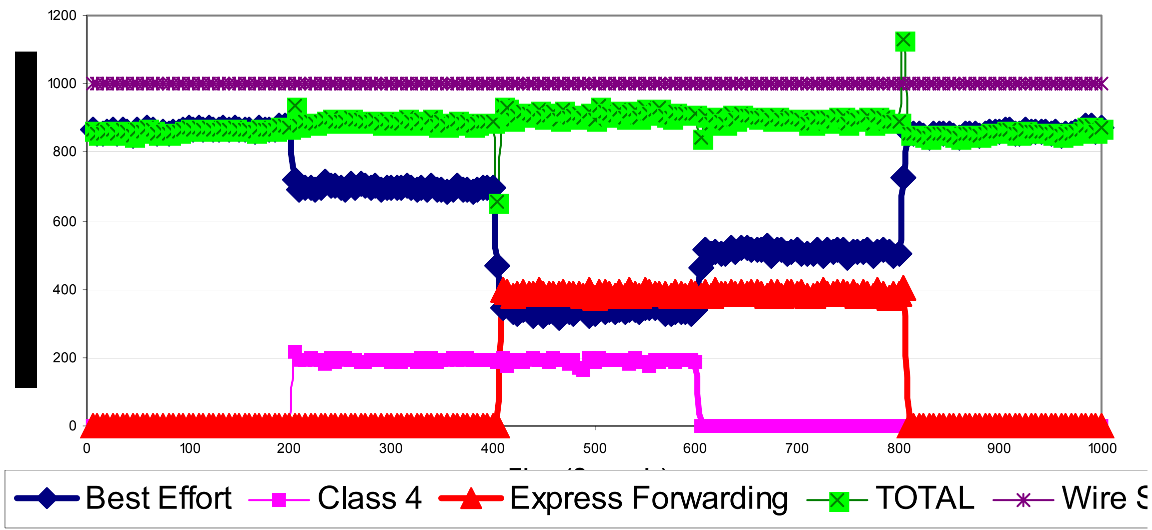
MPLS/QoS for UltraLight

UltraLight plans to explore the full range of end-to-end connections across the network, from best-effort, packet-switched through dedicated end-to-end light-paths.

MPLS paths with QoS attributes fill a middle ground in this network space and allow fine-grained allocation of virtual pipes, sized to the needs of the application or user.

UltraLight, in conjunction with the DoE/MICS funded TeraPaths effort, is working toward extensible solutions for implementing such capabilities in next generation networks

TeraPaths Initial QoS test at BNL



Optical Path Plans

Emerging "light path" technologies are becoming popular in the Grid community:

- They can extend and augment existing grid computing infrastructures, currently focused on CPU/storage, to include the network as an integral Grid component.
- Those technologies seem to be the most effective way to offer network resource provisioning on-demand between end-systems.

A major capability we are developing in Ultralight is the ability to dynamically switch optical paths across the node, bypassing electronic equipment via a fiber cross connect.

The ability to switch dynamically provides additional functionality and also models the more abstract case where switching is done between colors (ITU grid lambdas).

MonaLisa to Manage LightPaths

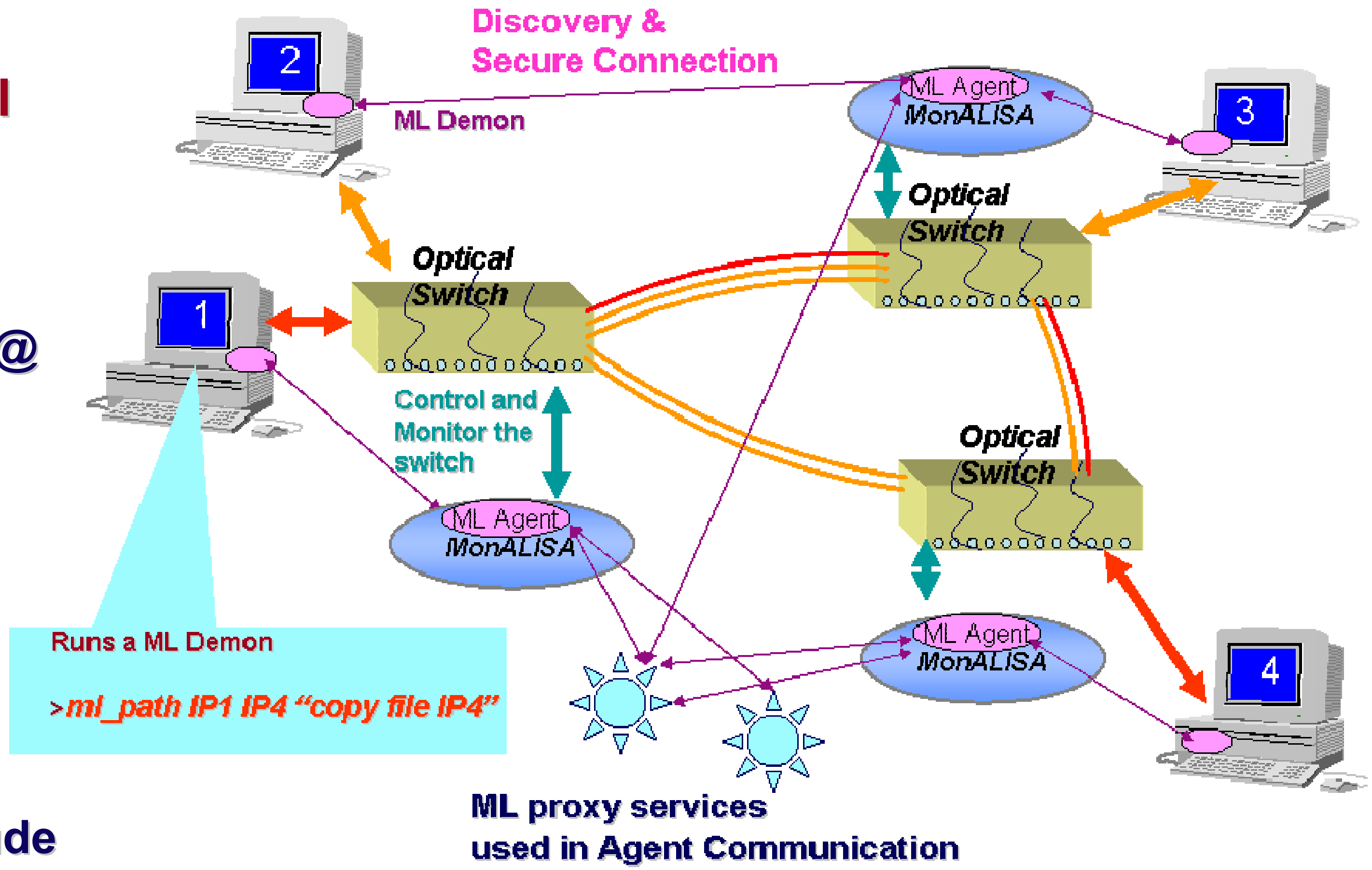
Dedicated modules to monitor and control optical switches

Used to control

- CALIENT switch @ CIT
- GLIMMERGLASS switch @ CERN

ML agent system

- Used to create global path
- Algorithm can be extended to include prioritisation and pre-allocation



Monitoring for UltraLight

Network monitoring is *essential* for UltraLight.

We need to understand our network infrastructure and track its performance both historically and in real-time to enable the network as a managed robust component of our overall infrastructure.

There are two ongoing efforts we are leveraging to help provide us with the monitoring capability required:

IEPM http://www-iepm.slac.stanford.edu/bw/

MonALISA http://monalisa.cern.ch

Both efforts have already made significant progress within UltraLight. We are working on the level of detail to track, as well as determining the most effect user interface and presentation.

End-Systems performance

Latest disk to disk over 10Gbps WAN: 4.3 Gbits/sec (536 MB/sec) - 8 TCP streams from CERN to Caltech; windows, 1TB file

Quad Opteron AMD848 2.2GHz processors with 3 AMD-8131 chipsets: 4 64-bit/133MHz PCI-X slots.

3 Supermicro Marvell SATA disk controllers + 24 SATA 7200rpm SATA disks

- Local Disk IO – 9.6 Gbits/sec (1.2 GBytes/sec read/write, with <20% CPU utilization)

10GE NIC

- 10 GE NIC – 7.5 Gbits/sec (memory-to-memory, with 52% CPU utilization)
- 2*10 GE NIC (802.3ad link aggregation) – 11.1 Gbits/sec (memory-to-memory)
- Need PCI-Express, TCP offload engines

A 4U server with 24 disks (9 TB) and a 10 GbE NIC is capable of 700 MBytes/sec in the LAN and ~500 MBytes/sec in a WAN is $ 25k today.

Small server with a few disks (1.2 TB) capable of 120 Mbytes/sec (matching a GbE port) is $ 4K.

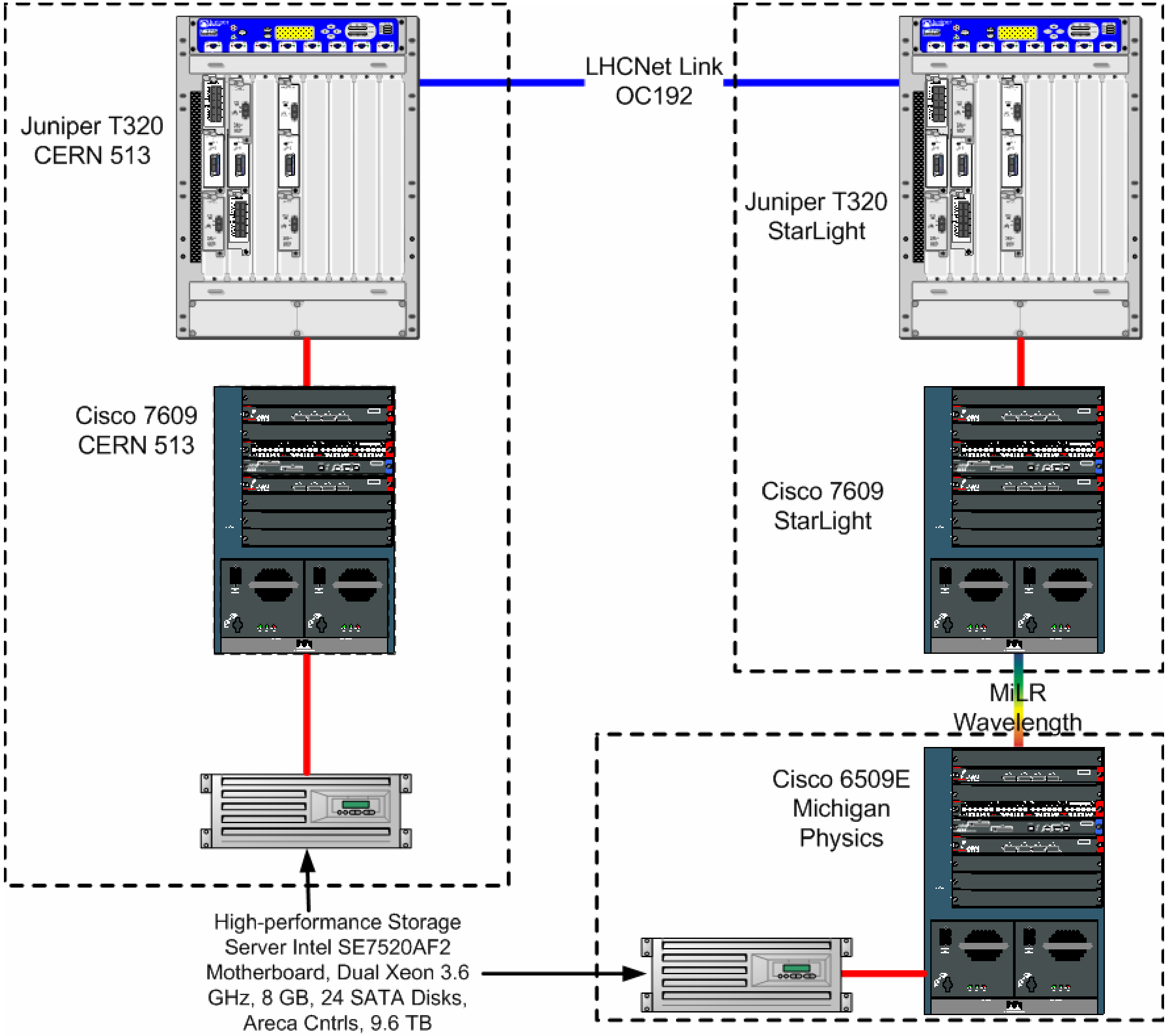
UltraLight/ATLAS Data Transfer Test

UltraLight is interested in disk-to-disk systems capable of utilizing 10 Gbit networks

We plan to begin testing by July

Starting goal is to match the ~500 Mbytes/sec already achieved

Target is to reach 1 GByte/sec by the end of the year

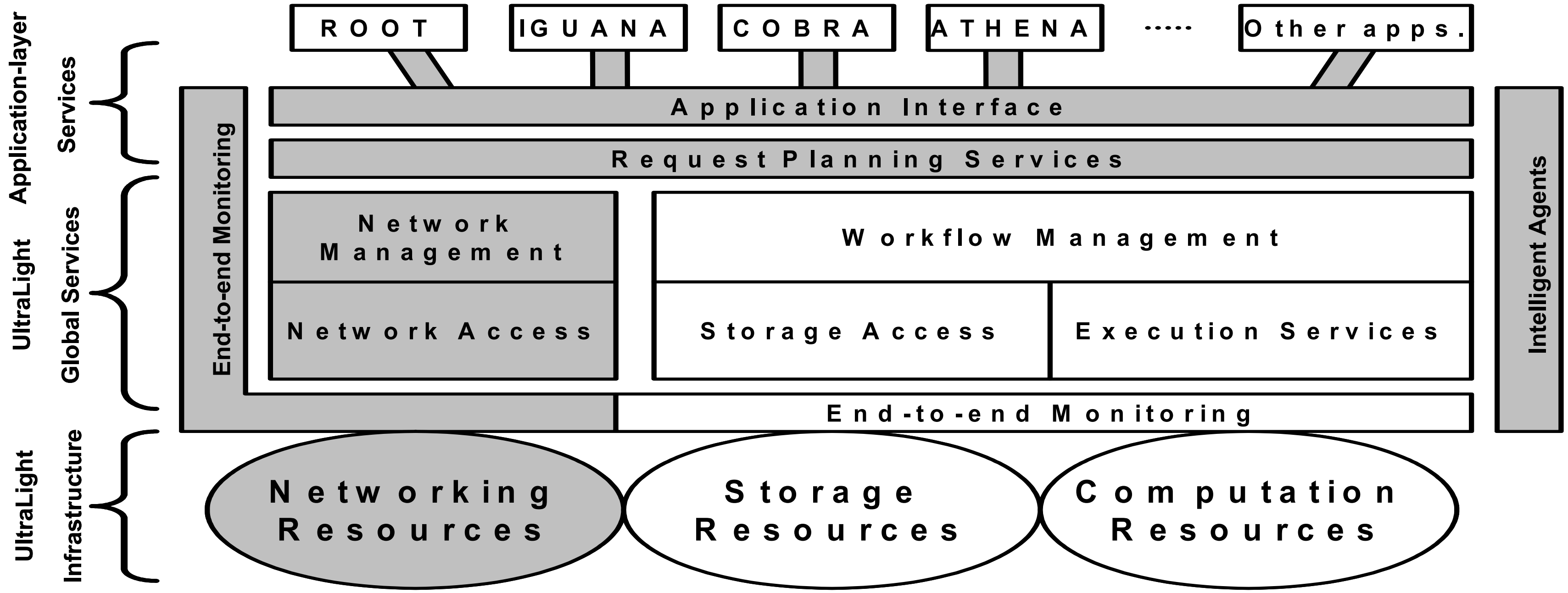


UltraLight Global Services

- ***Global Services*** support management and co-scheduling of multiple resource types, and provide strategic recovery mechanisms from system failures
- Schedule decisions based on CPU, I/O, Network capability and End-to-end task performance estimates, incl. loading effects
- Decisions are constrained by local and global policies
- Implementation: Autodiscovering, multithreaded services, service-engines to schedule threads, making the system scalable and robust
- Global Services Consist of:
  - *Network and System Resource Monitoring*, to provide pervasive end-to-end resource monitoring info. to HLS
  - *Policy Based Job Planning Services,* balancing policy, efficient resource use and acceptable turnaround time
  - *Task Execution Services,* with job tracking user interfaces, incremental replanning in case of partial incompletion
- Exploit the *SPHINX (UFI) framework for Grid-wide policy-based scheduling*; extend its capability to the network domain

UltraLight Application Services

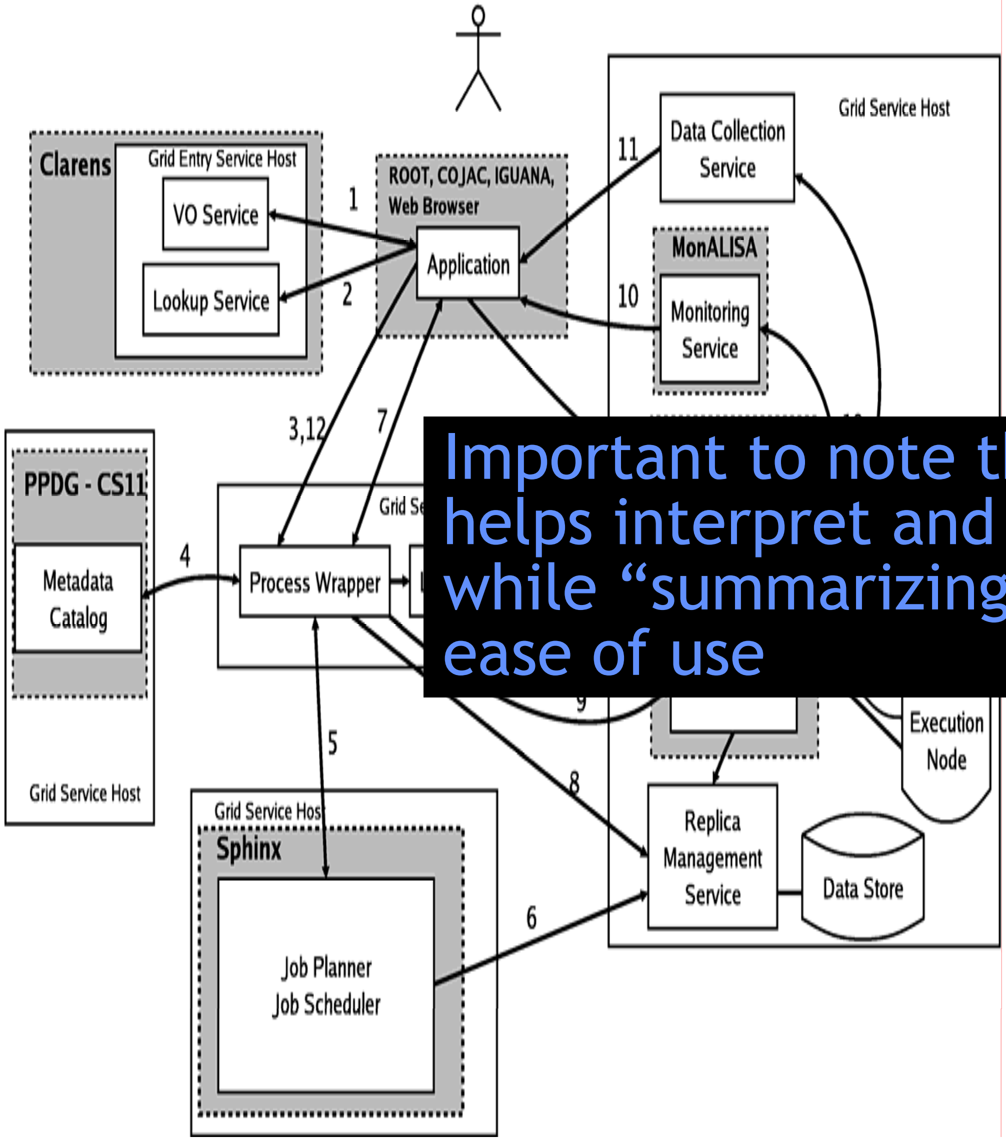
Make UltraLight Functionality Available to the Physics Applications, & Their Grid Production & Analysis Environments



- *Application Frameworks Augmented to Interact Effectively with the Global Services (GS)*
  - *GS Interact in Turn with the Storage Access & Local Execution Service Layers*

- *Apps. Provide Hints to High-Level Services About Requirements*
  - *Interfaced also to managed Net and Storage services*
  - *Allows effective caching, pre-fetching; opportunities for global and local optimization of thru-put*

A User's Perspective: GAE In the UltraLight Context



Important to note that the system helps interpret and optimize itself while "summarizing" the details for ease of use

1. **Client Authenticates Against VO**
2. **Uses Lookup Service to Discover Available Grid Services**
3. **& 4. Contacts Data Location & Job Planner/Scheduler Services**
5. **& 6.-8. Generates a Job Plan**
9. **Starts Plan Execution by Sending ... eduled Execution**

... ervice *Sends* ... *k* to the Client Application, While

(b) Steering Service is Used by Clients or Agents to Modify Plan When Needed

11. **Data is Returned to the Client Through Data Collection Service**
12. ***Iterate in Next Analysis Cycle***

UltraLight Educational Outreach

- **Based at FIU, leveraging its CHEPREO and CIARA activities to provide students with opportunities in physics, astronomy and network research**
  - ➢ **Alvarez, Kramer will help bridge and consolidate activities with GriPhyN, iVDGL and eVLBI**
- ***GOAL:* *To inspire and train the next generation of physicists, astronomers and network scientists***
- ✓ ***PLAN:***
  1. **Integrate students in the core research and application integration activities at participating universities**
  2. **Use the UltraLight testbed for student-defined network projects**
  3. **Opportunities for (minority) FIU students to participate**
  4. **Student & teacher involvement through REU, CIARA, Quarknet**
  5. **Use UltraLight's international reach to allow US students to participate in research experiences at int'l labs and accelerators, from their home institutions.**
- **A Three Day Workshop has been organized for June 6-10 to launch these efforts**

UltraLight Near-term Milestones

UltraLight

Protocols:
- Integration and of FAST TCP (V.1) (July 2005)
- New MPLS and optical path-based provisioning (Aug 2005)
- New TCP implementations testing (Aug-Sep, 2005)

Optical Switching:
- Commission optical switch at the LA CENIC/NLR and CERN (May 2005)
- Develop dynamic server connections on a path for TB transactions (Sep 2005)

Storage and Application Services:
- Evaluate/optimize drivers/parameters for I/O filesystems (April 2005)
- Evaluate/optimize drivers/parameters for 10 GbE server NICs (June 2005)
- Select hardware for 10 GE NICs, Buses and RAID controllers (June-Sep 2005)
- Breaking the 1 GByte/s barrier (Sep 2005)

Monitoring and Simulation:
- Deployment of end-to-end monitoring framework. (Aug 2005)
- Integration of tools/models to build a simulator for network fabric. (Dec 2005)

Agents:
- Start development of Agents for resource scheduling (June 2005)
- Match scheduling allocations to usage policies (Sep 2005)

WanInLab:
- Connect Caltech WanInLab to testbed (June 2005)
- Procedure to move new protocol stacks into field trials (June 2005)

Summary: Network Progress

For many years the Wide Area Network has been the bottleneck; this is no longer the case in many countries thus making deployment of a data intensive Grid infrastructure possible!

- Recent I2LSR records show for the first time ever that the network can be truly transparent; throughputs are limited by end-hosts
- Challenge shifted from getting adequate bandwidth to deploying adequate infrastructure to make ***effective*** use of it!

Some transport protocol issues still need to be resolved; however there are many encouraging signs that practical solutions may now be in sight.

1GByte/sec disk to disk challenge.
Today: 1 TB at 536 MB/sec from CERN to Caltech

- Still in Early Stages; Expect Substantial Improvements

Next generation network and Grid system

- Extend and augment existing grid computing infrastructures (currently focused on CPU/storage) to include the network as an integral component.

Conclusions

UltraLight promises to deliver a critical missing component for future eScience: *the integrated, managed network*

We have a strong team in place, as well as a plan, to provide the needed infrastructure and services for production use by LHC turn-on at the end of 2007

We look forward to a busy productive year working on UltraLight!

Questions?