The UltraLight Project and the Challenges of Data Storage and Networking for the CMS Experiment

UltraLight Collaboration

- Caltech (lead inst.)
- BNL
- Michigan
- MIT
- Florida
- Florida International
- FNAL
- San Diego
- SLAC
- Vanderbilt

R. Cavanaugh

University of Florida

ICFA06

Krakow, Poland

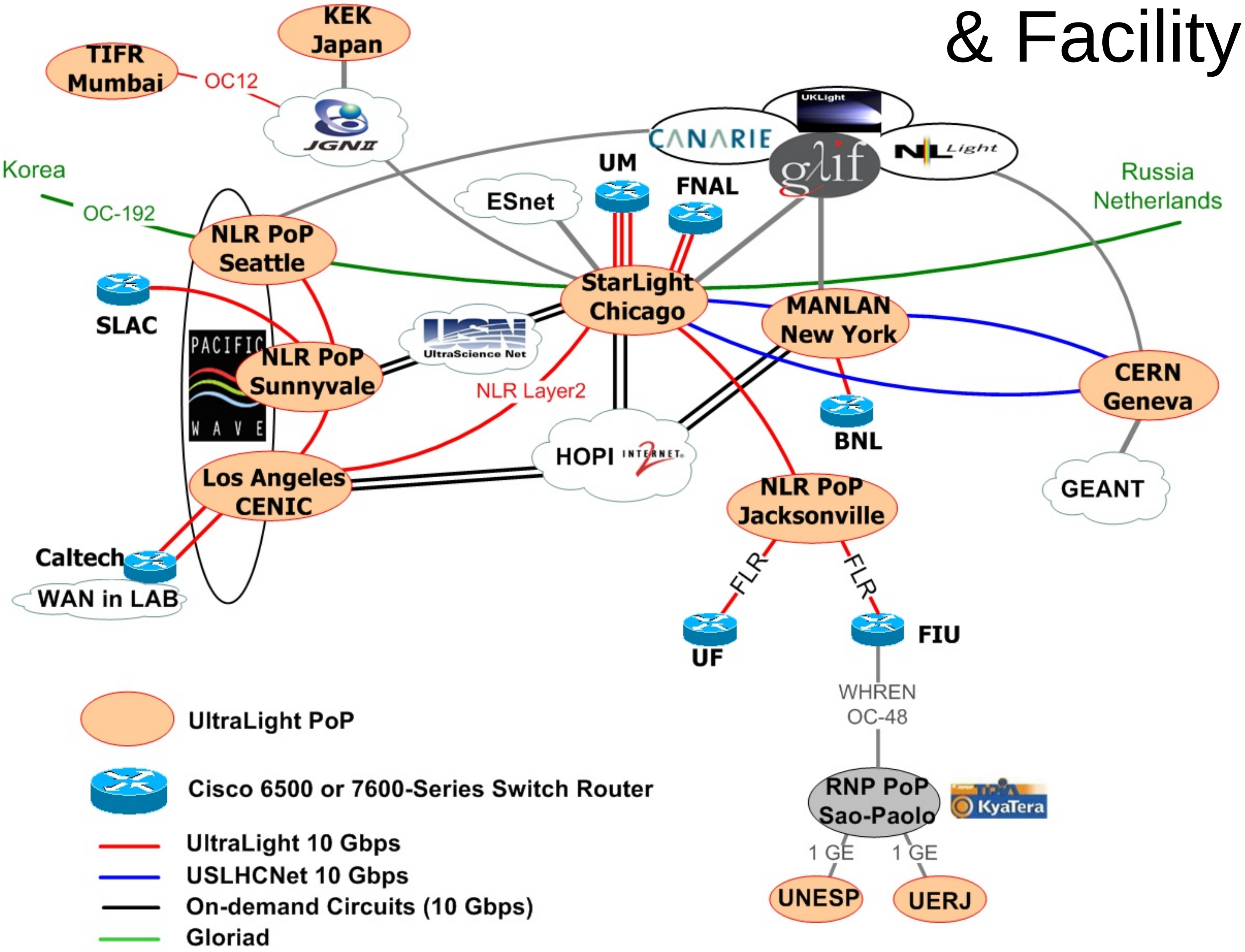
UltraLight: A New Class of Integrated Information Systems

- Expose the Network as an Actively Managed Resource

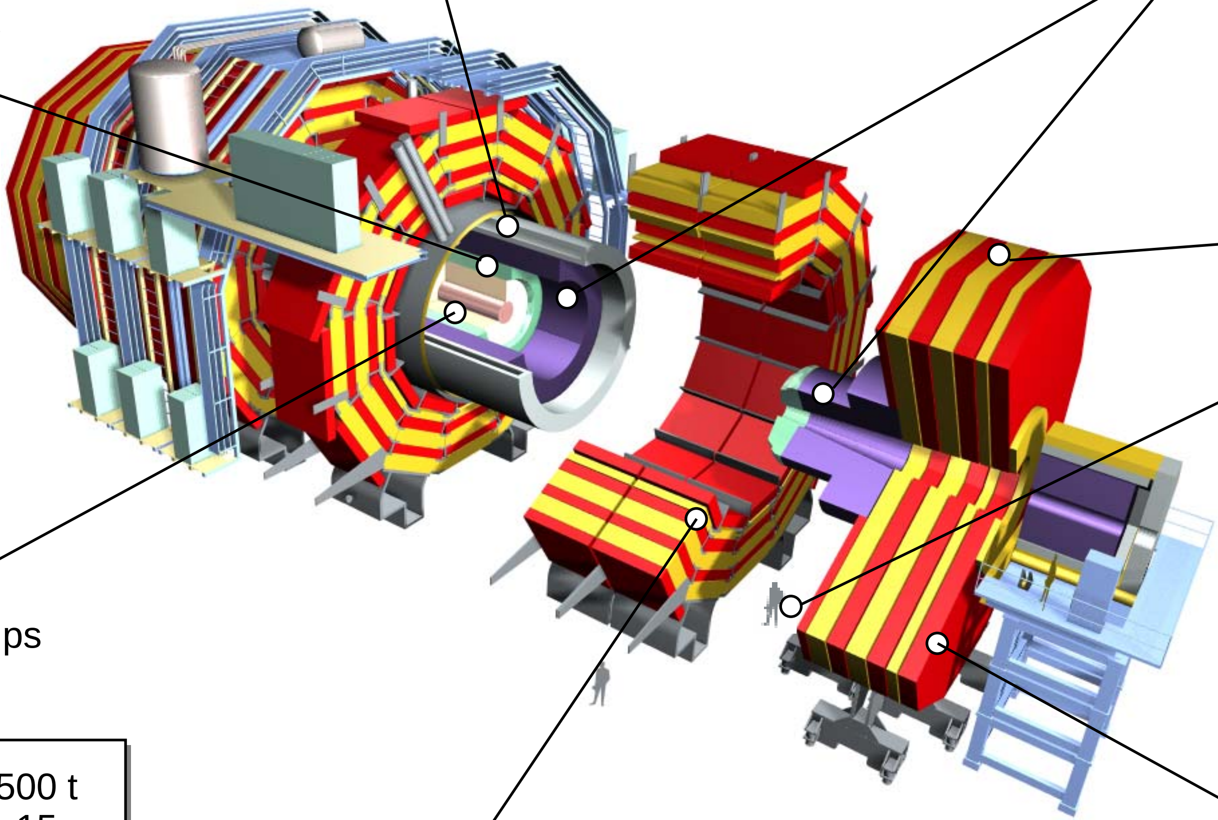
- Based on a "Hybrid" packet- and circuit-switched optical network infrastructure
  - Ultrascale Protocols (e.g. FAST) and Dynamic Optical Paths
- Monitor, Manage and Optimize resources in real-time
  - Using a set of Agent-Based Intelligent Global Services
- Leverages already-existing, developing software infrastructure in round-the-clock operation:
  - MonALISA, GAE/Clarens, OSG

- Exceptional Support from
  - Industry: Cisco & Calient
  - Research community: NLR, CENIC, Internet2/Abilene, ESnet

UltraLight Network Laboratory Testbed & Facility

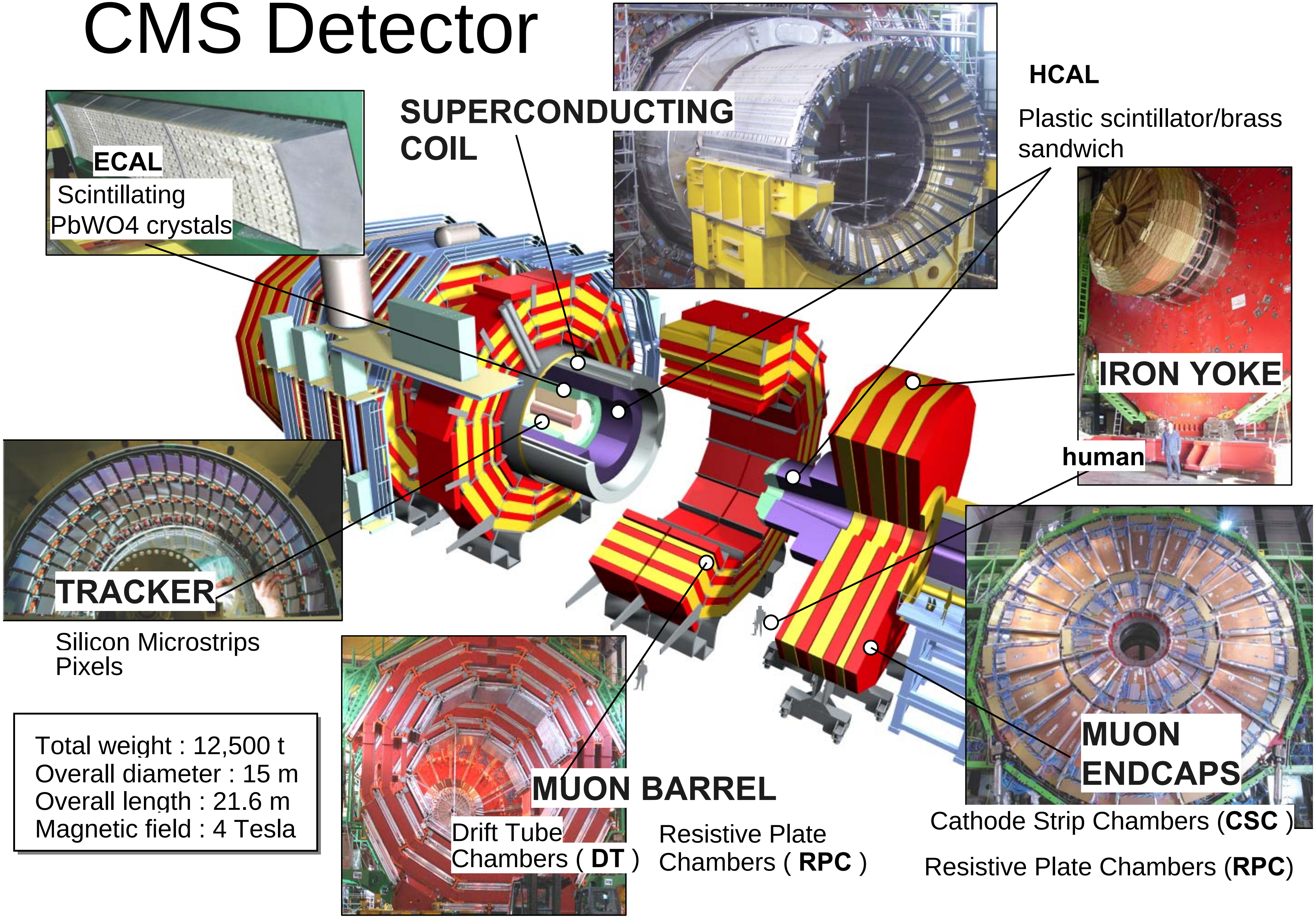


CMS Detector



Total weight : 12,500 t
Overall diameter : 15 m
Overall length : 21.6 m
Magnetic field : 4 Tesla

CMS Detector



CMS Detector

SUPERCONDUCTING COIL

HCAL

Plastic scintillator/brass sandwich

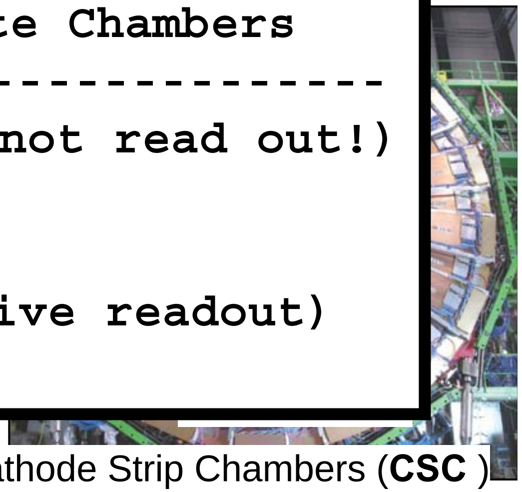
Silic
Pixel

Detector	No. Channels	Sensors
Vertex	80 000 000	Pixels
Tracker	16 000 000	Silicon Microstrips
Preshower	512 000	Silicon
Calorimeters	125 000	Scintillating Crystals
		Scintillator / Brass sandwich
Muons	1 000 000	Drift/Strip/Plate Chambers
Total	~100 000 000	Channels (most not read out!)
Event Size	~1-2	MB (with selective readout)
DAQ Bandwidth	~200	MB / sec

Total w
Overal
Overall length : 21.6 m
Magnetic field : 4 Tesla

MUON BARREL

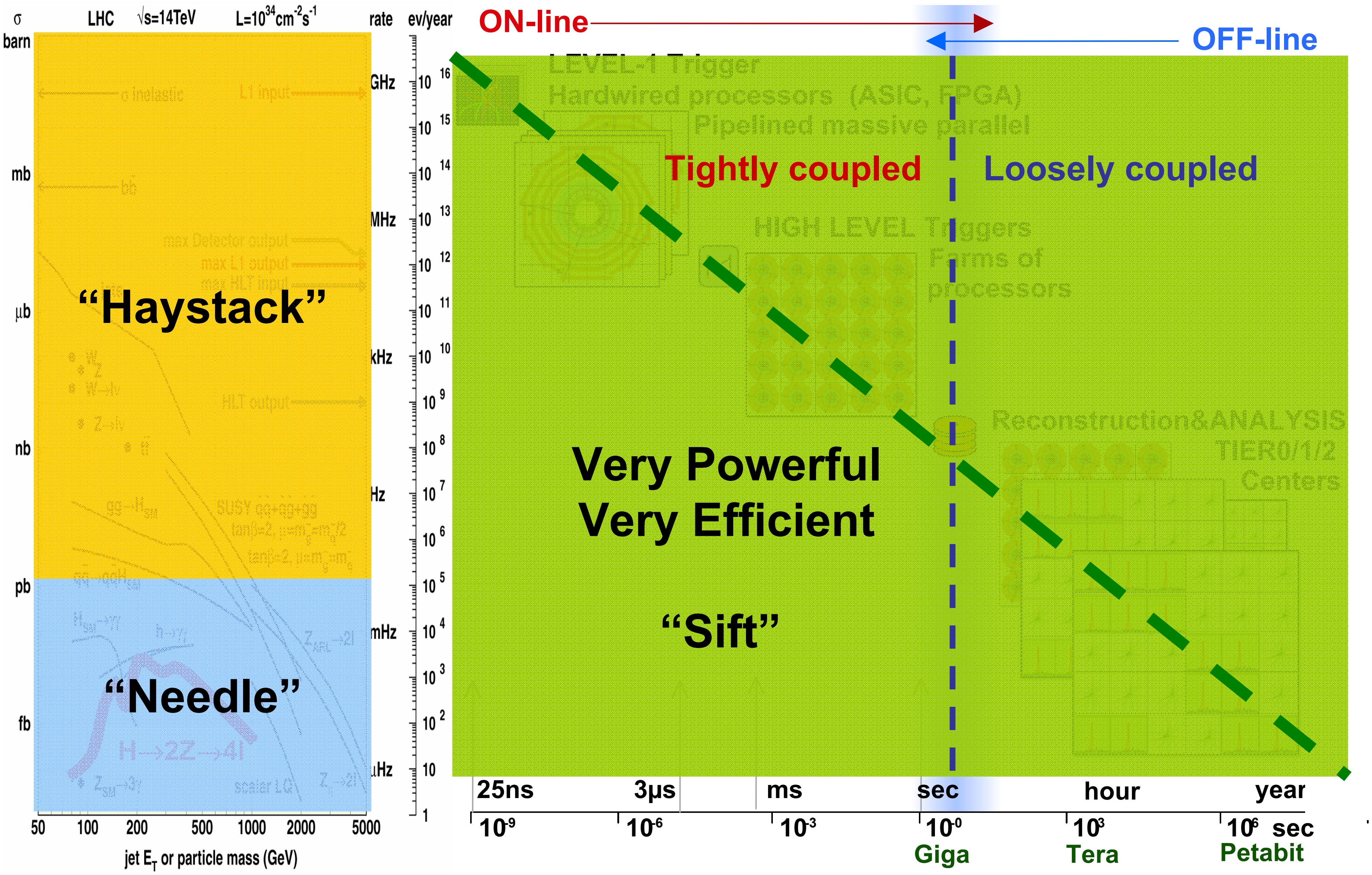
Resistive Plate Chambers (**RPC**)



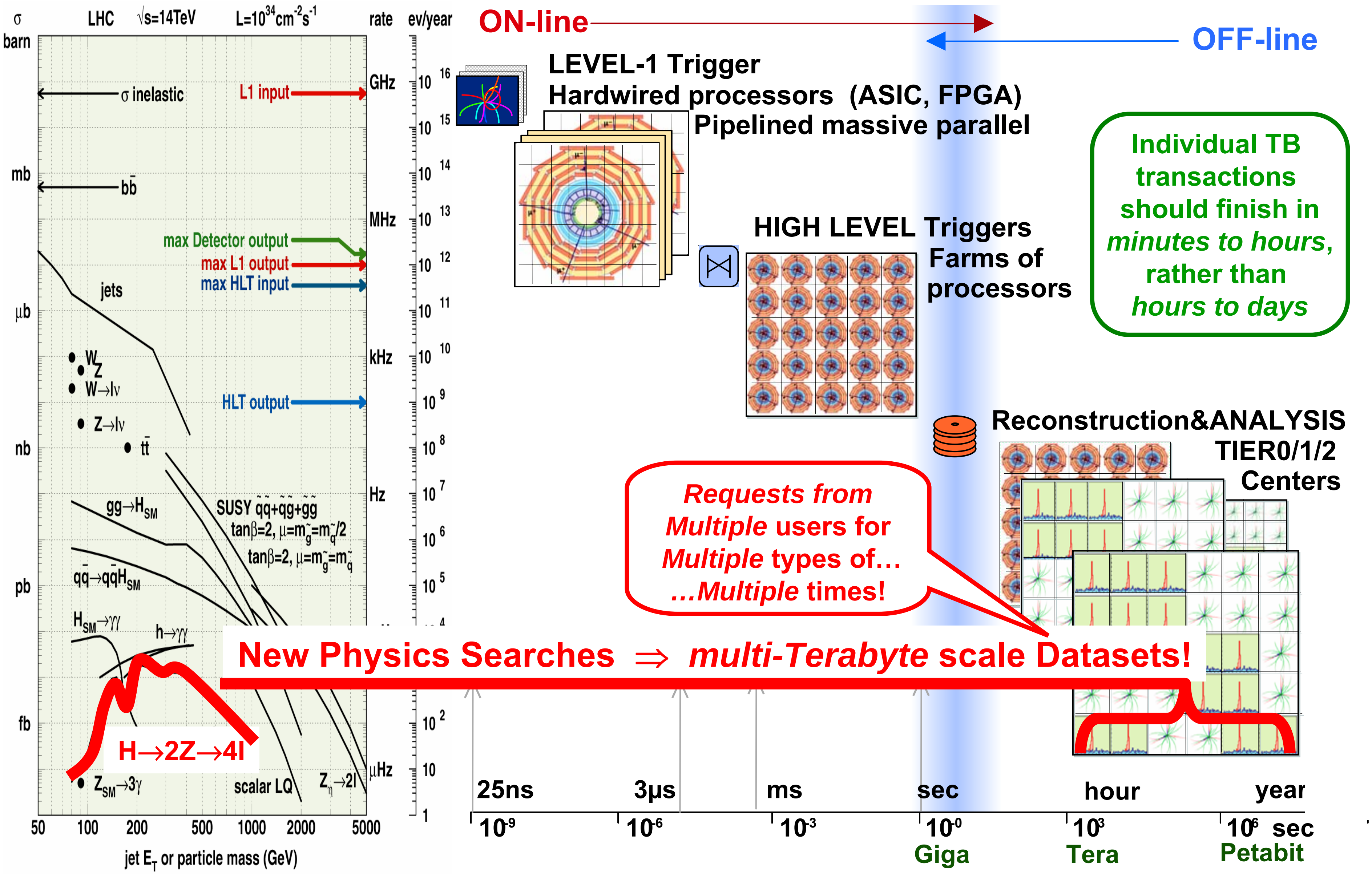
Cathode Strip Chambers (**CSC**)

Resistive Plate Chambers (**RPC**)

Main Science Problem



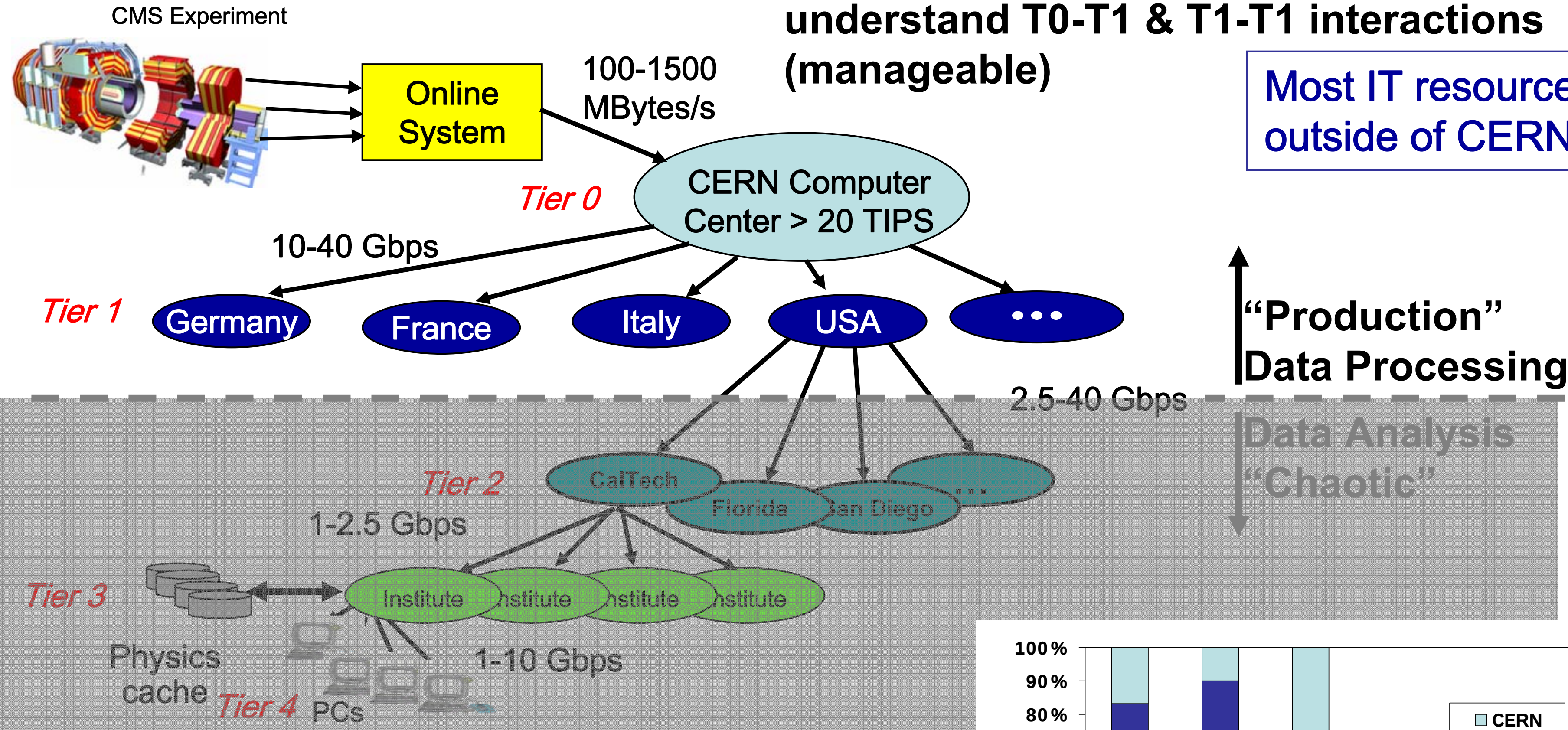
Main Science Problem



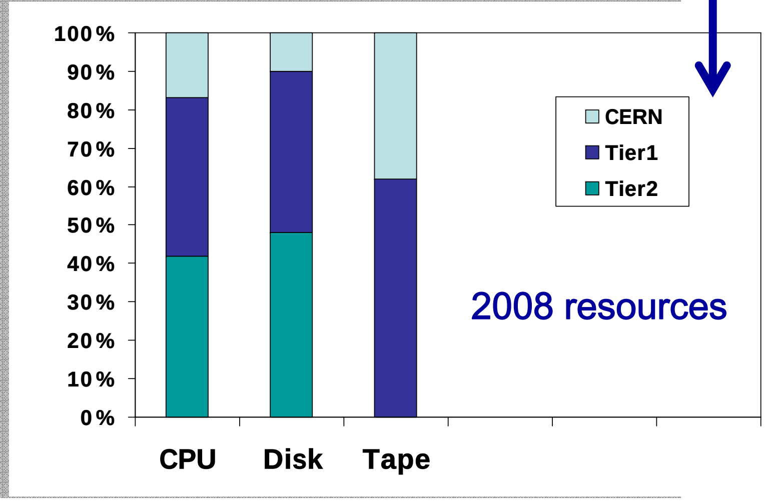
LHC Computing Grid Hierarchy

A lot of effort has been devoted to understand T0-T1 & T1-T1 interactions (manageable)

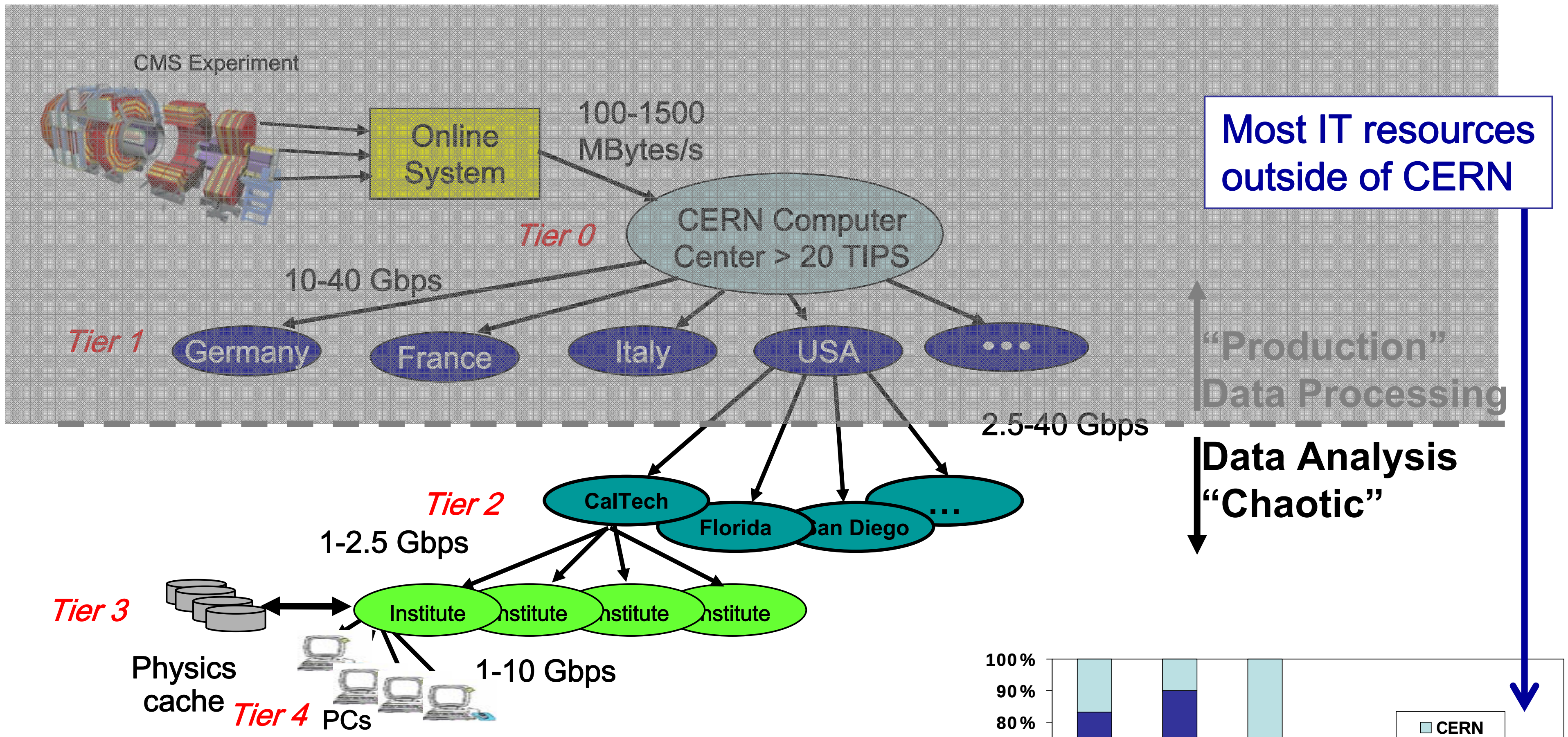
Most IT resources outside of CERN

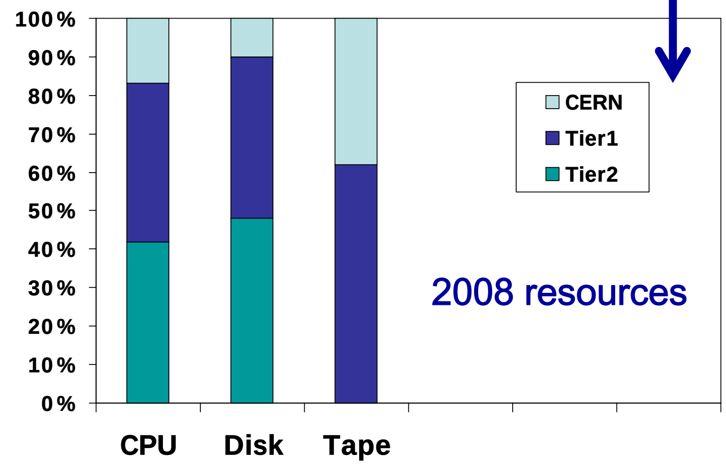


- ~10s of Petabytes/yr by 2007-8
- ~1000 Petabytes in < 10 yrs



LHC Computing Grid Hierarchy

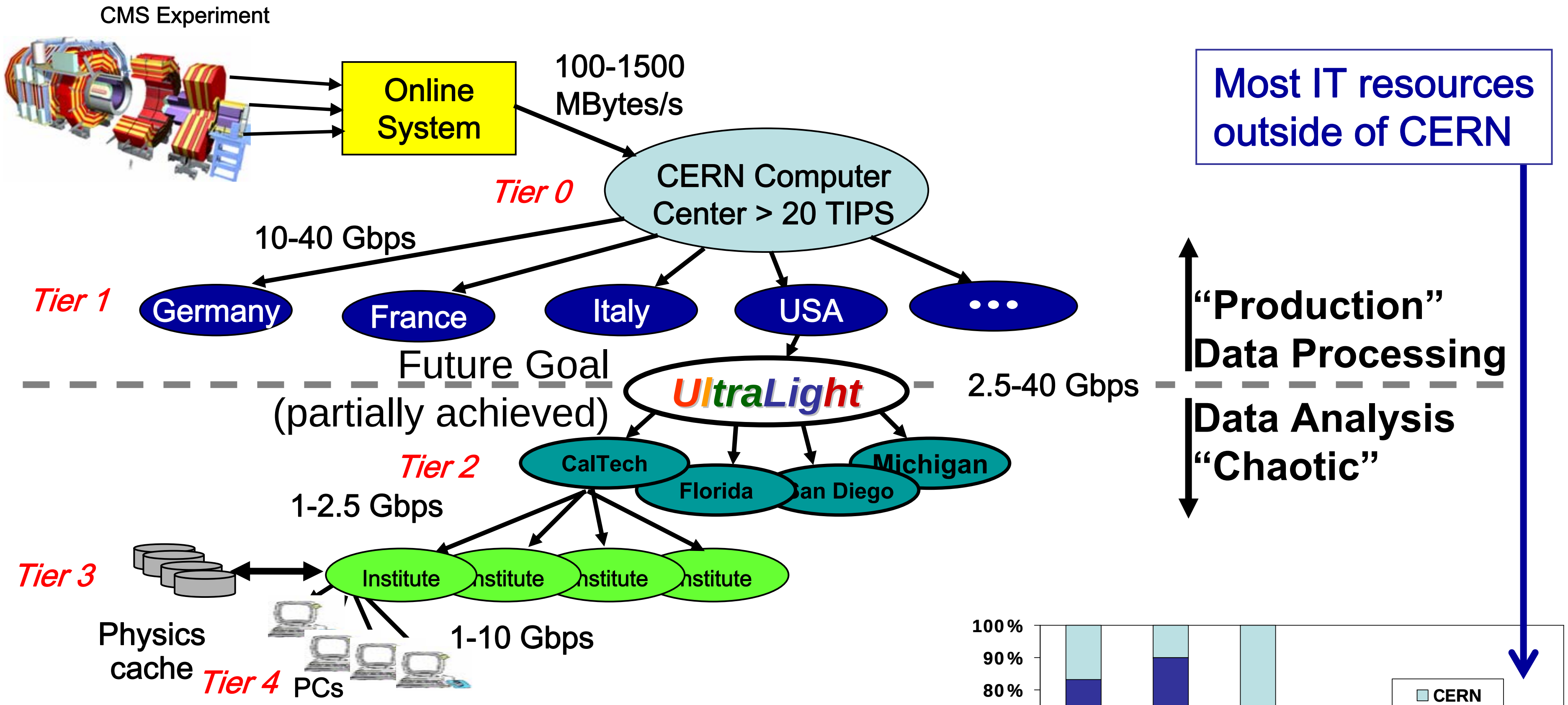




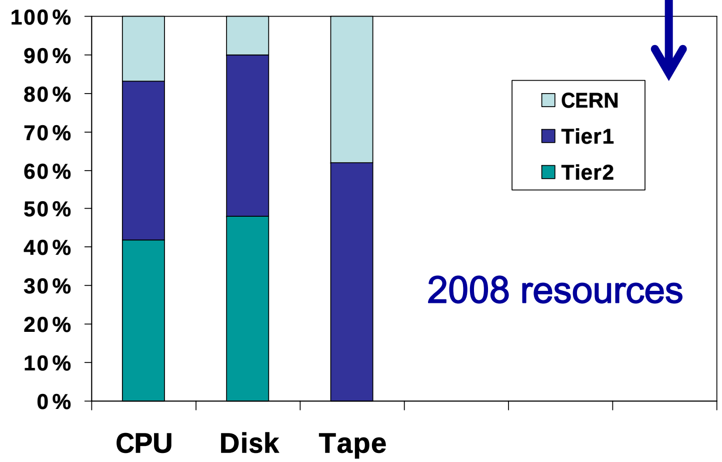
- ~10s of Petabytes/yr by 2007-8
- ~1000 Petabytes in < 10 yrs

Much less effort has been devoted to understand T1-T2 & T2-T2 interactions (less manageable & more complex)

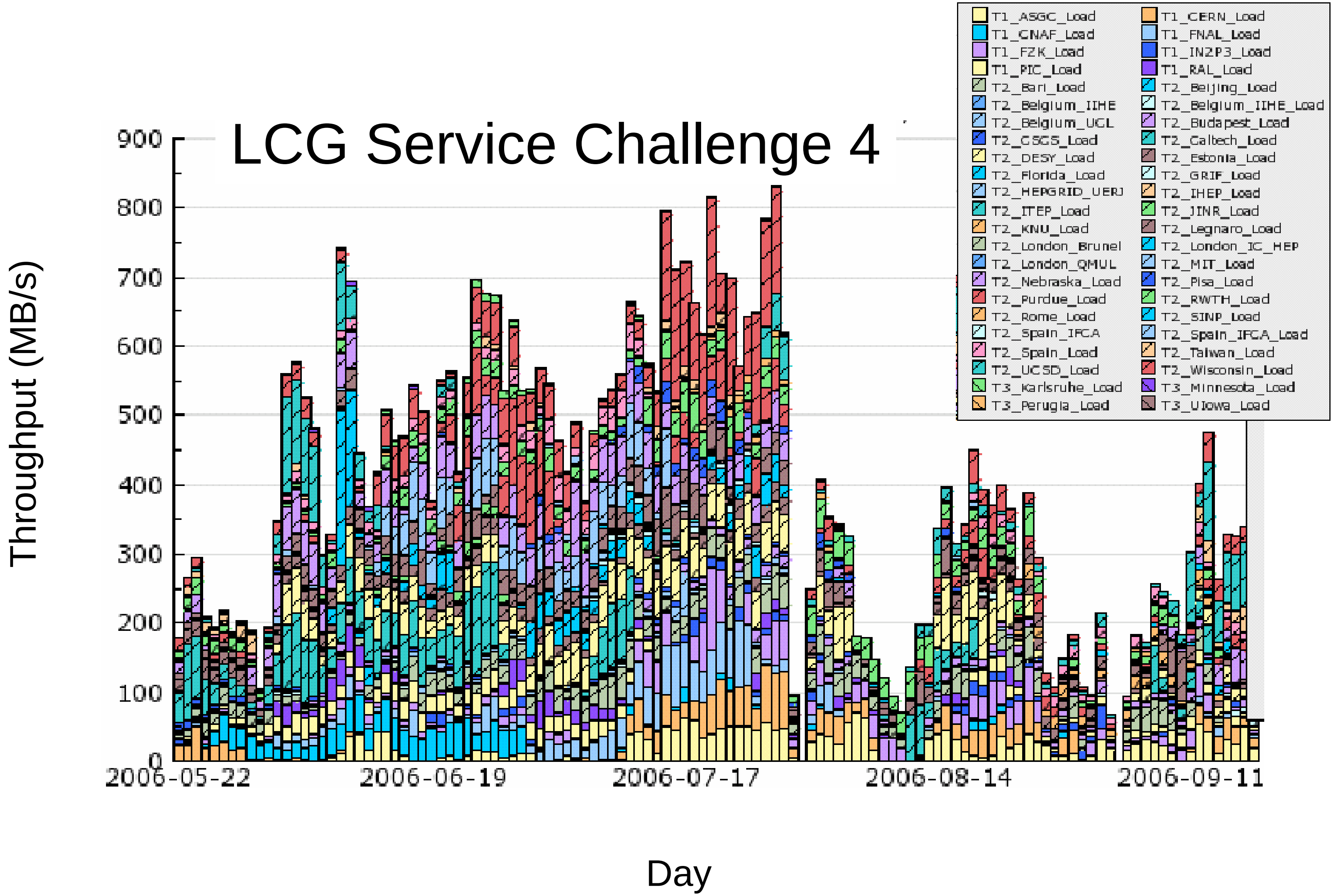
LHC Computing Grid Hierarchy



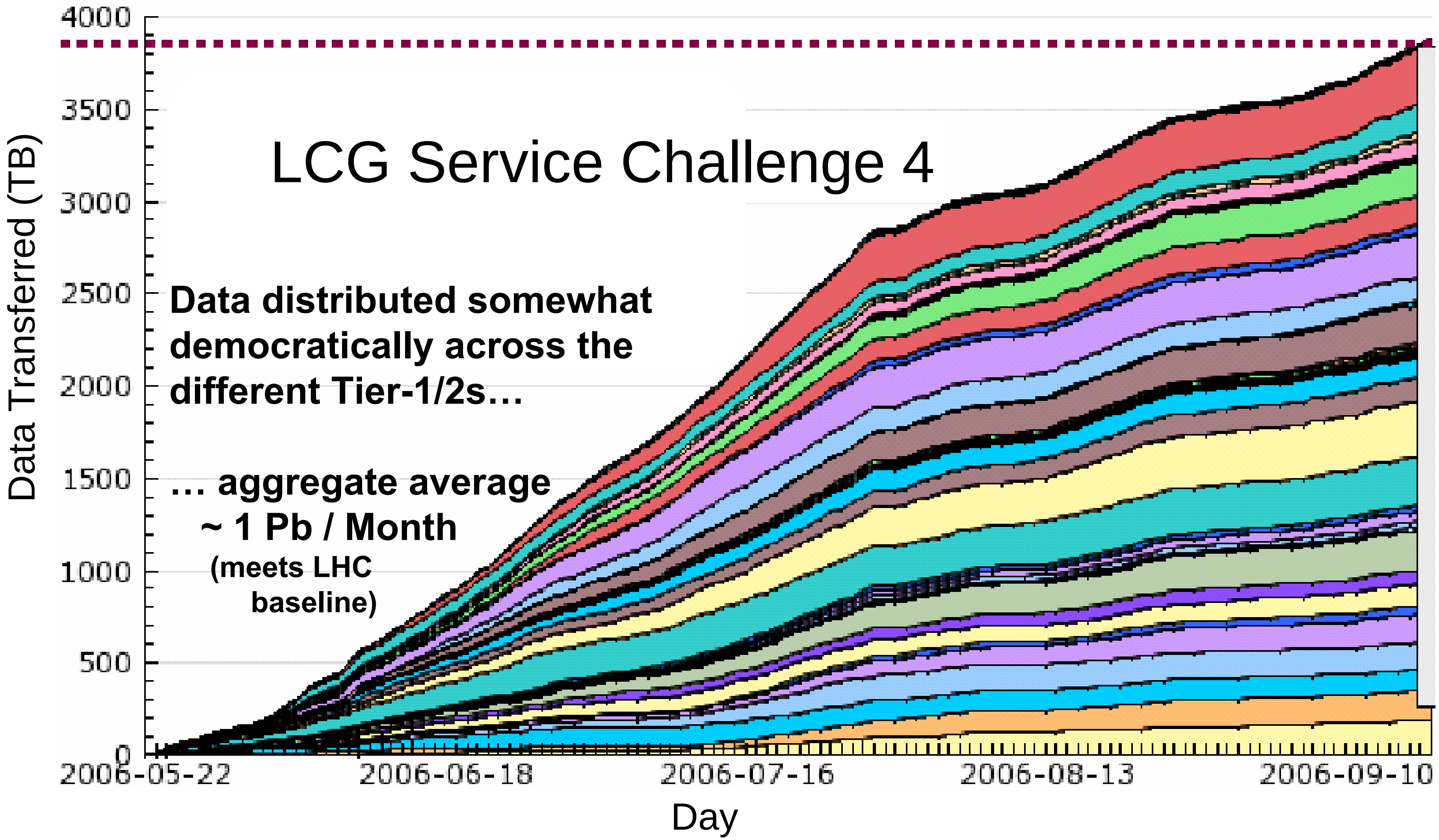
- ~10s of Petabytes/yr by 2007-8
- ~1000 Petabytes in < 10 yrs
- UltraLight Network Provides a Data "Bus"
  - T1-T2 & T2-T2 Non-hierarchical data flows
  - Natural from Data Analysis point of view!



Recent Experience with CMS Transfers across the LHC Grid Hierarchy



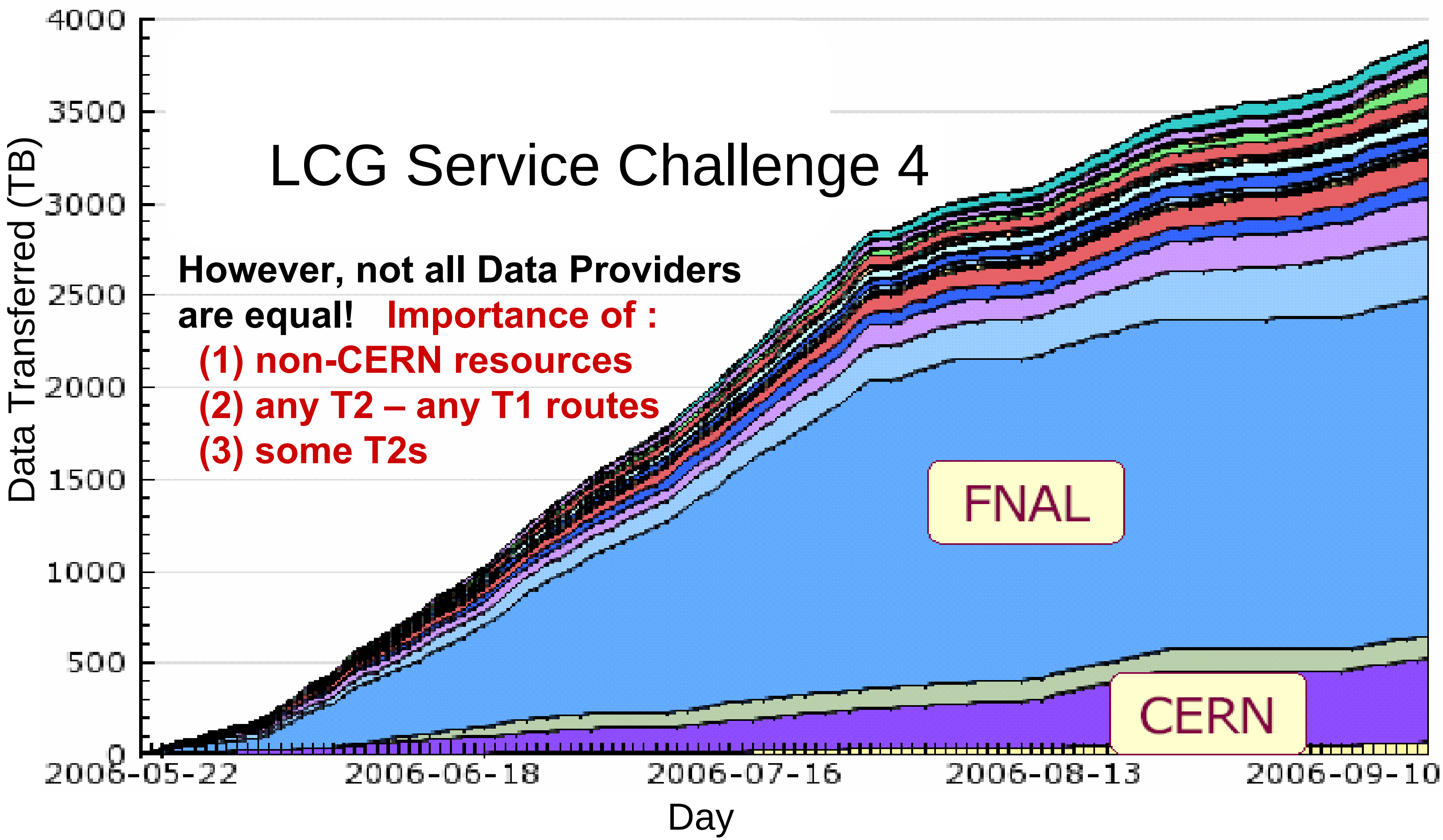
CMS Transfer Sinks



"not bad" : many (most?) Tier2 analysis centres not starved…

…Critical for delivering Physics to the Collaboration!!

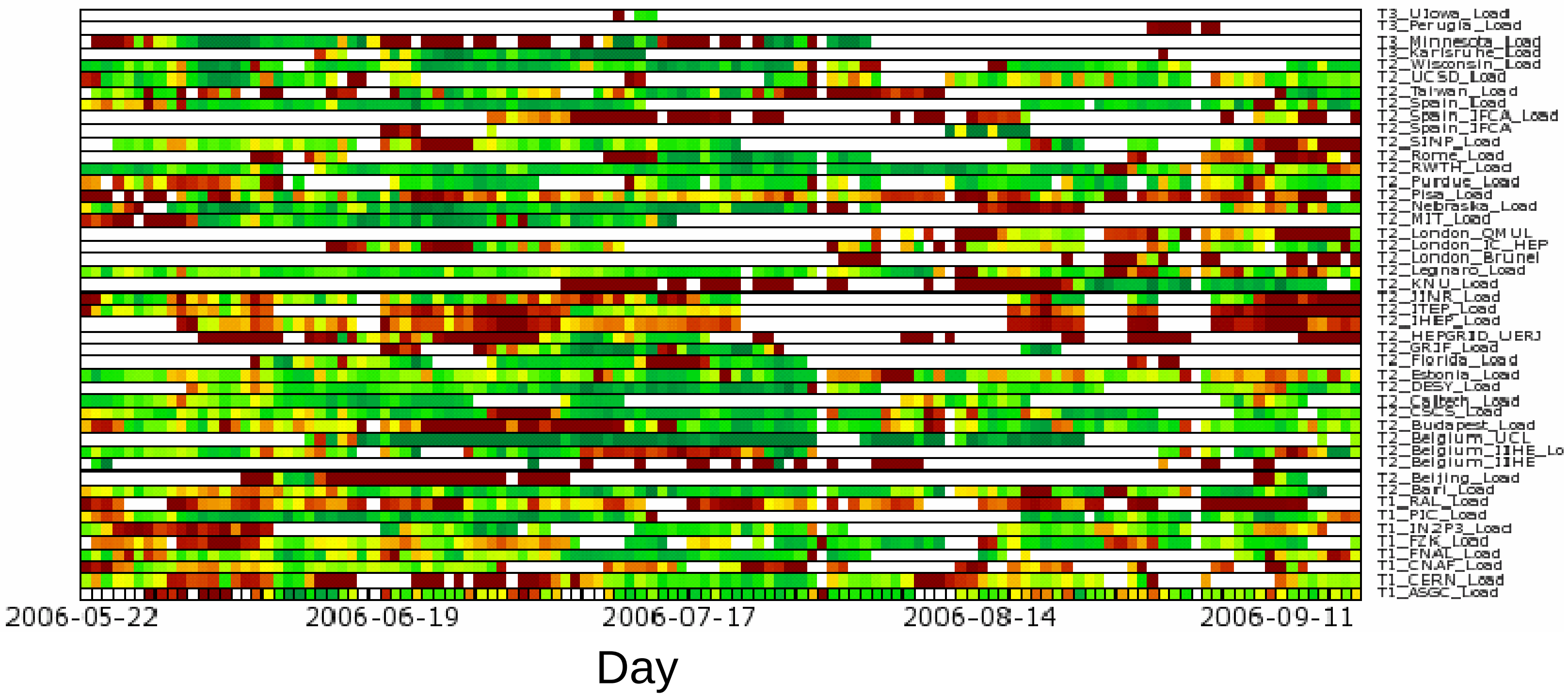
CMS Transfer Sources



Fault tolerance suggests usefulness of dynamically (not statically) routed flows

CMS Transfer Quality of Service

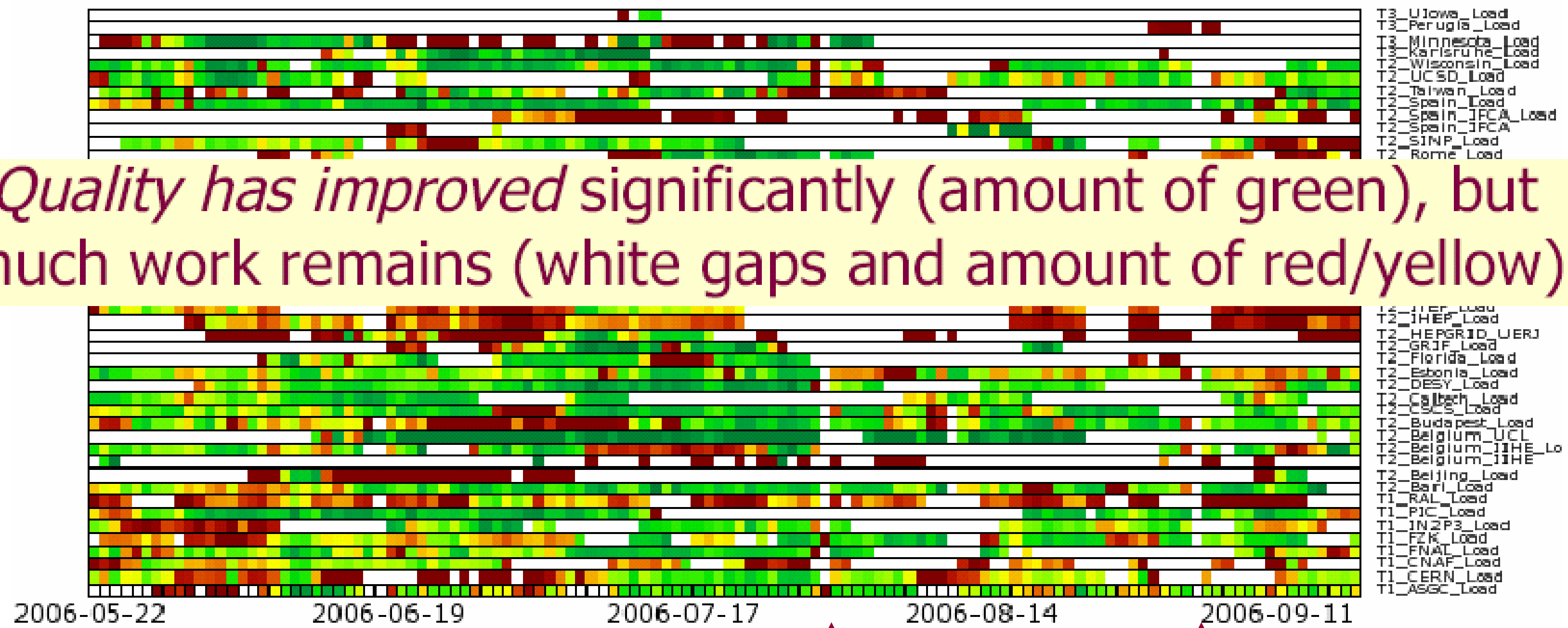
LCG Service Challenge 4



Fraction of successful Transfers

CMS Transfer Quality of Service

LCG Service Challenge 4



Quality has improved significantly (amount of green), but much work remains (white gaps and amount of red/yellow)

Day

End-host readiness is the biggest problem!
Very labour intensive…

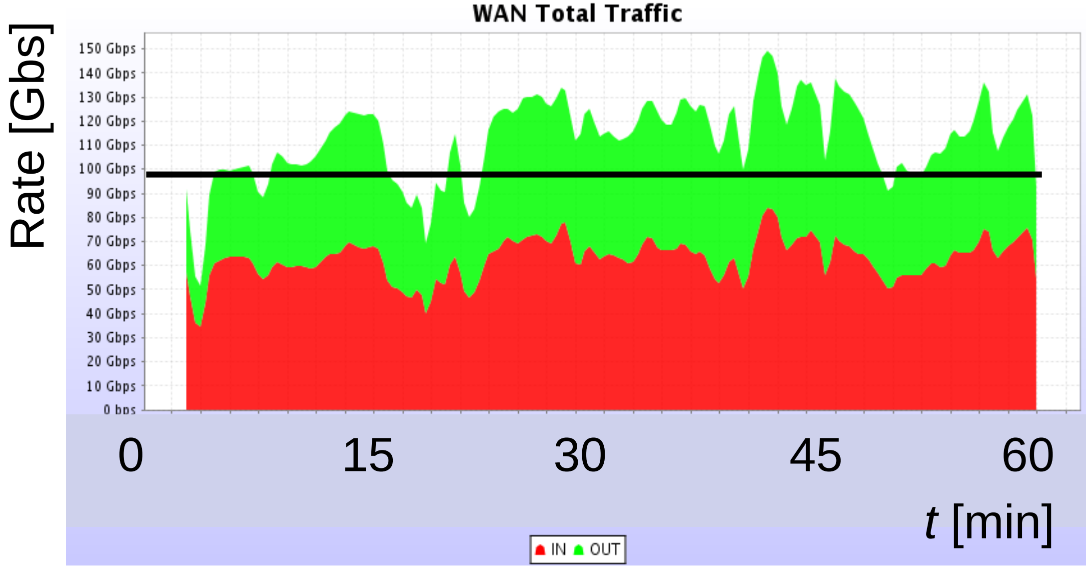
Service outages:
2 days out of 120

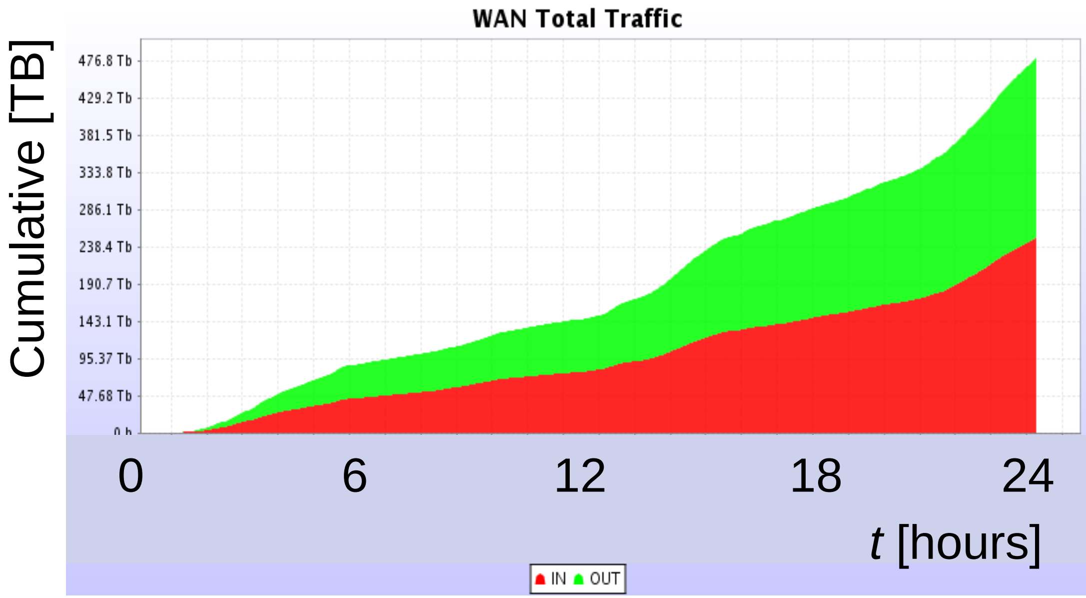
0-10% | 10-20% | 20-30% | 30-40% | 40-50% | 50-60% | 60-70% | 70-80% | 80-90% | 90-100% | 100+%

Fraction of successful Transfers

What is Achievable in the UltraLight Lab?

- Supercomputing 2005
- 151 Gbps peak rate
- 100+ Gbps sustained throughput for hours
- 475 Terabytes of physics data transported in less than 24 hours
- Sustained rate of 100+ Gb/s translates to > 1 Petabyte per day





Not a production-level exercise…

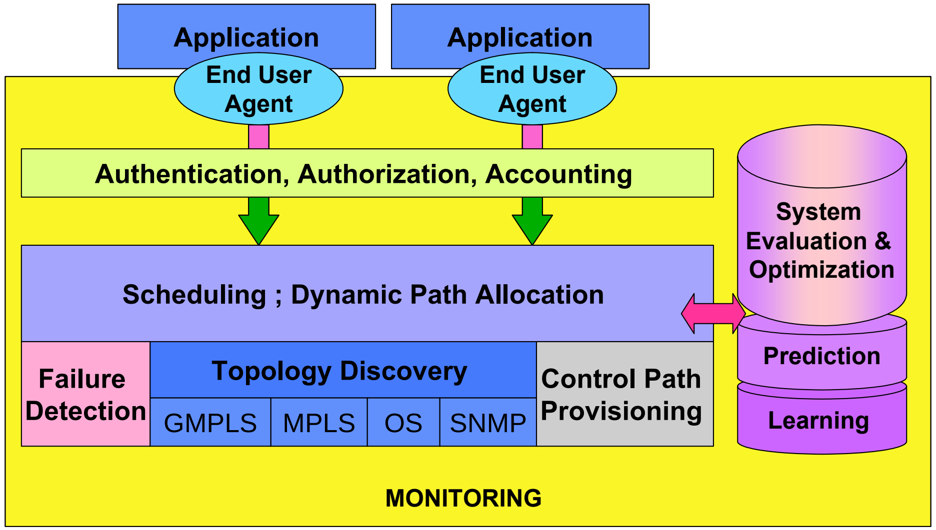
…Required tremendous amount of continuous manual attention! (~1 /λ)

UltraLight & CMS Experiences Expose The End-host Challenge

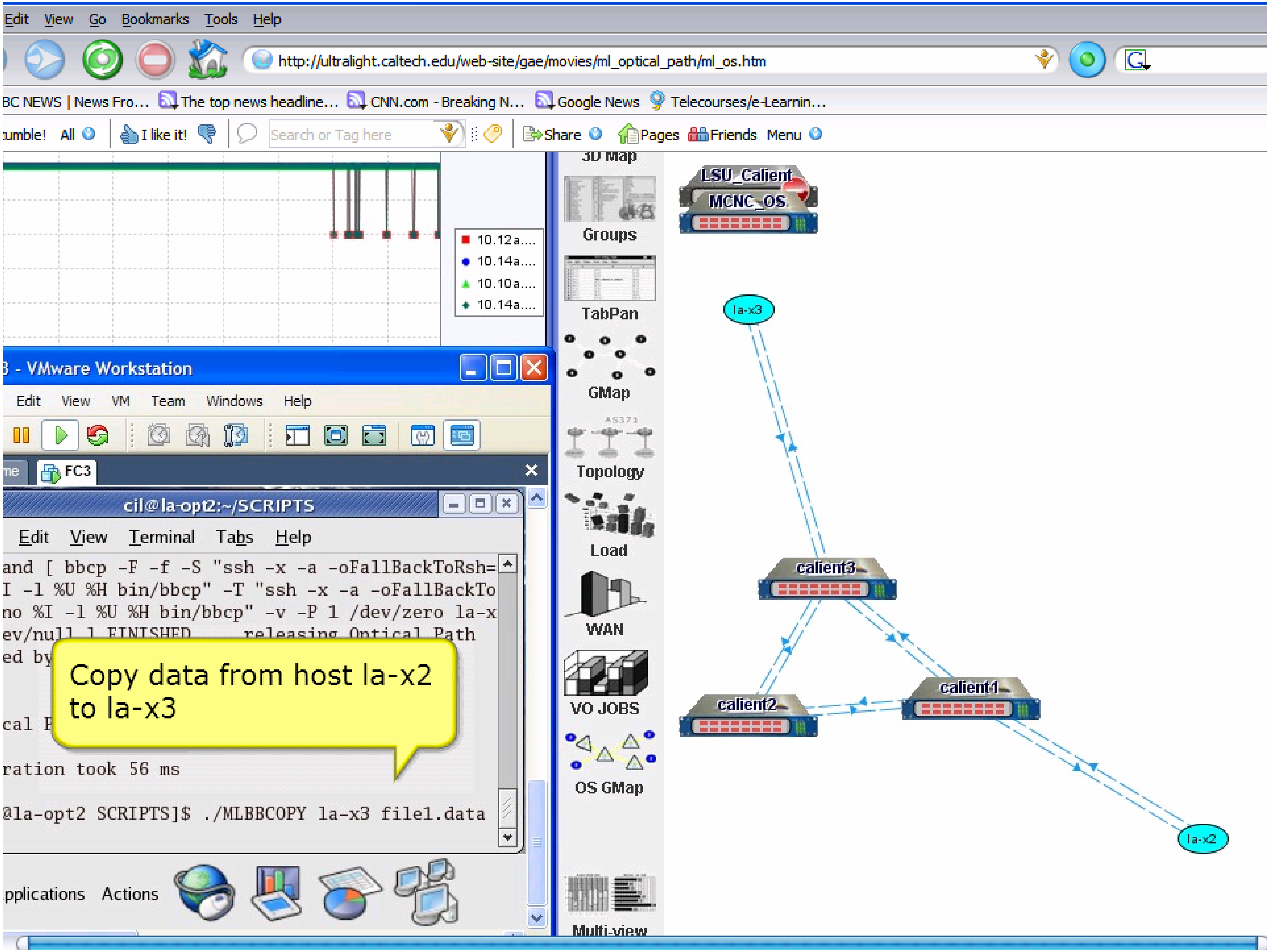
- Intense level of manual interventions (current reality)
  - The plague of data transfer performance!!
  - Requires distributed knowledge (expensive & continually out-of-date)
  - Introduces long failure and upgrade response-times
- Capable End-Host hardware often not deployed!!
- Storage technologies improving – still relatively young
  - SRM (WLCG, EGEE, OSG)
    - standard interface (only)
  - CASTOR (CERN)
    - Large scale (complex) storage (disk+tape) solution
    - Fully managed (queues, priorities, etc) data management
  - DPM (EGEE), dCache (DESY / FNAL) & LSTORE (Vanderbilt)
    - Smaller scale (simpler, but still complex), aggregate disk management
    - No queues or prioritized data management
- Clear need for E2E management, automation and self-healing!!
  - Has proven to be difficult in a loosely coupled "social" environment

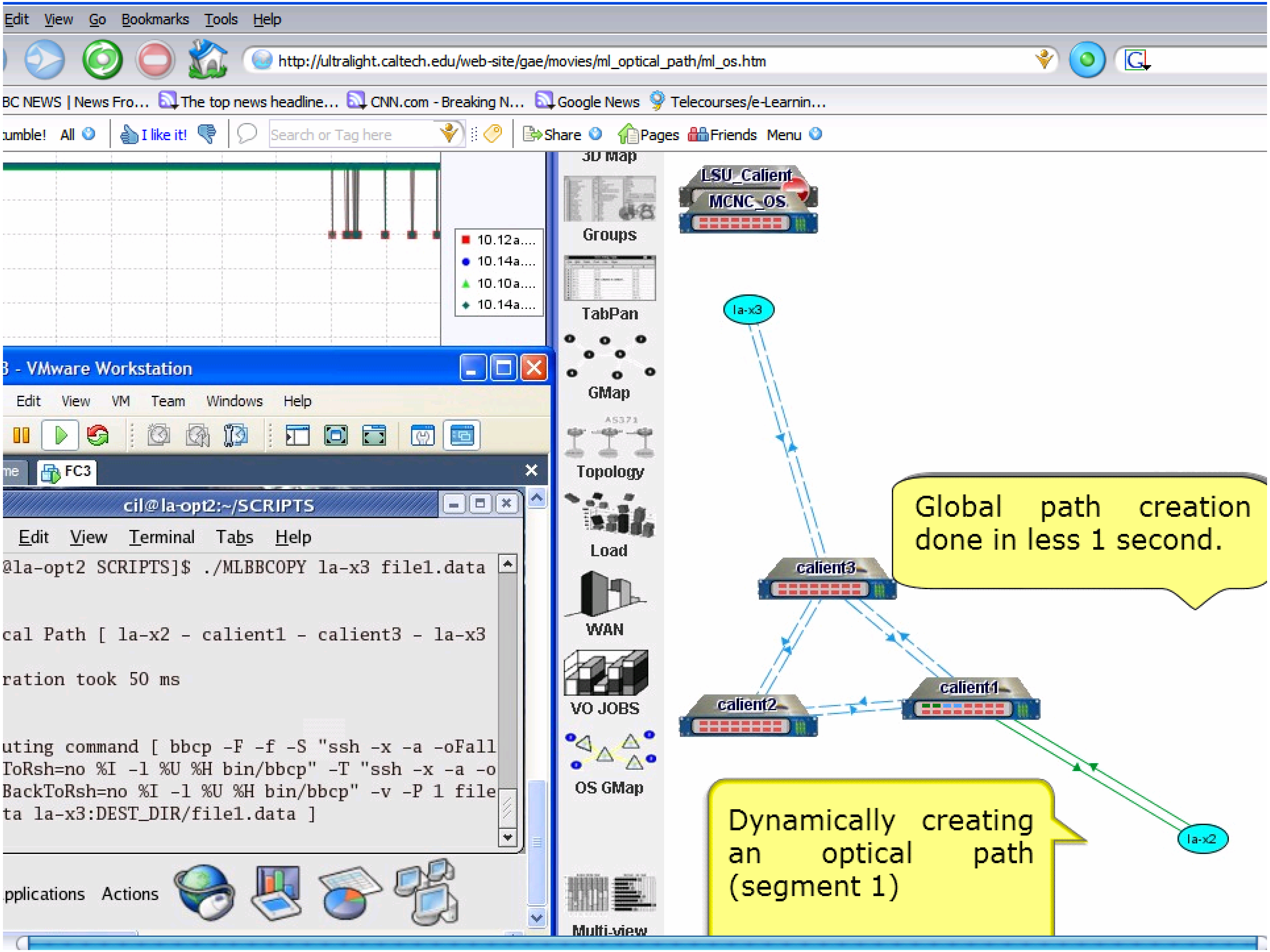
UltraLight Network R&D Global Planning Services

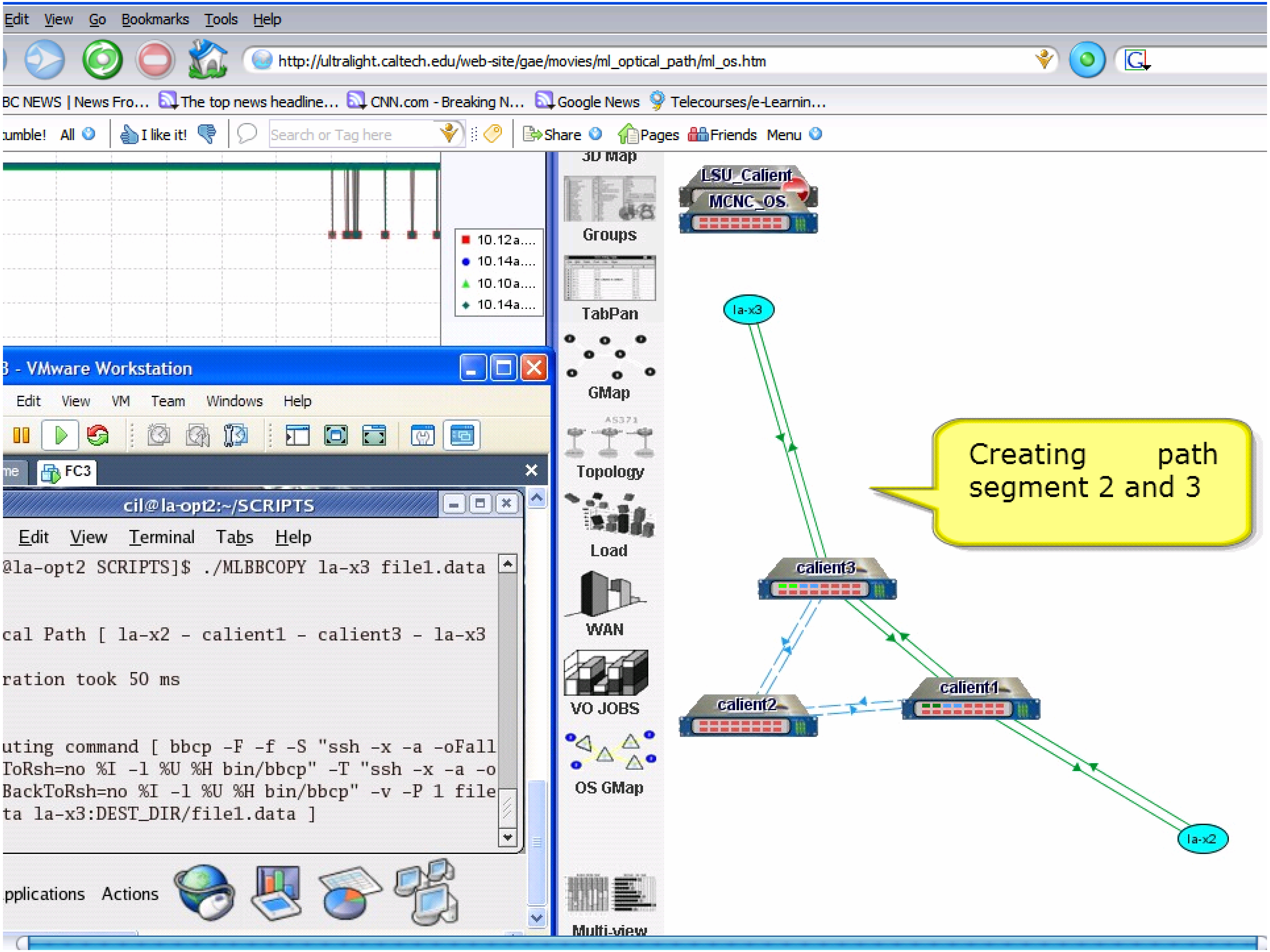
- VINCI :
  - Virtual Intelligent Networks for Computing Infrastructures
  - Based on existing MonALISA framework
  - What does vinci do?
- LISA :
  - Localhost Information Service Agent
  - Monitors end-systems
    - User
    - servers

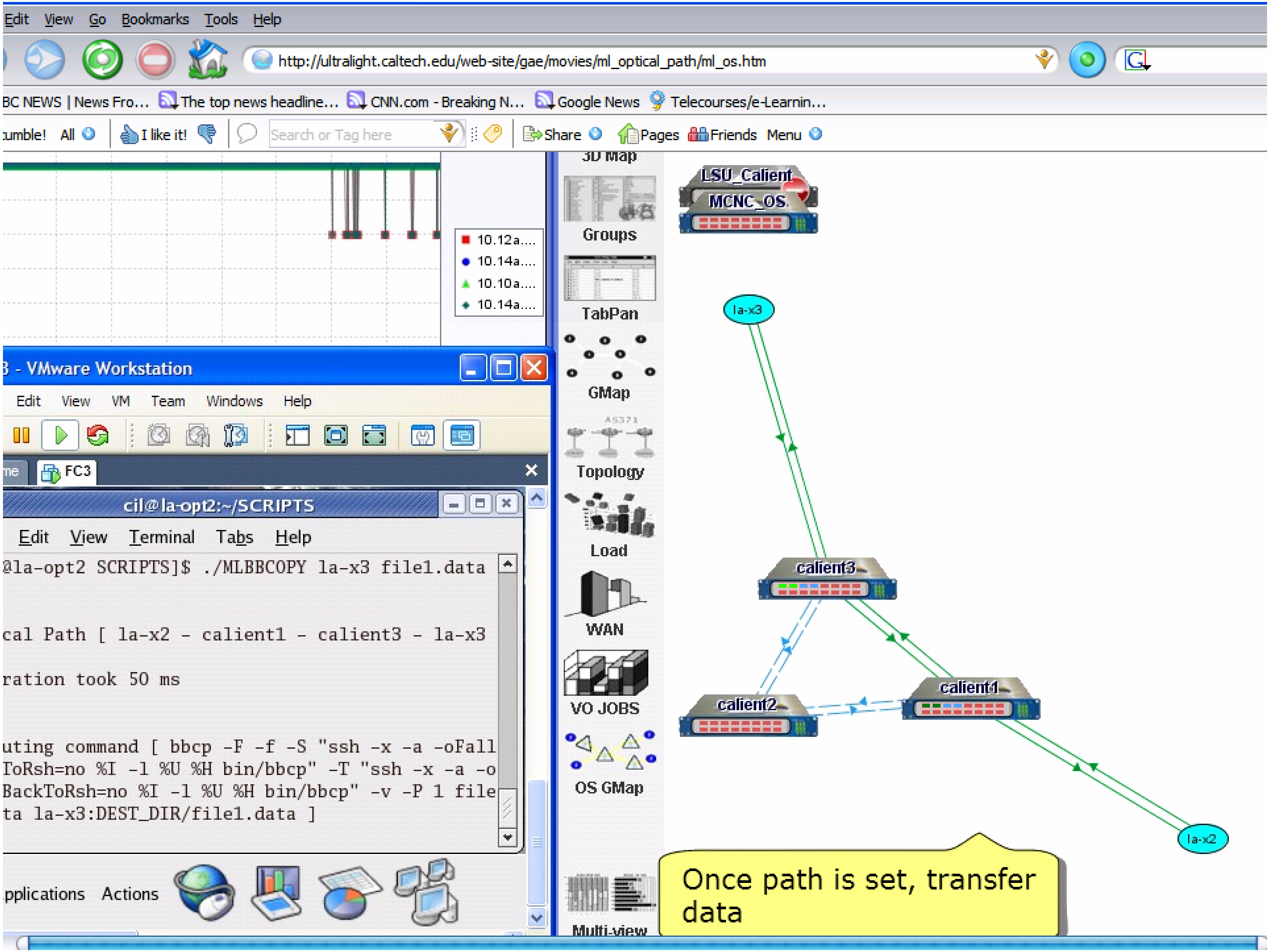


- UltraLight contributing to a VINCI prototype
  - production level system will come from elsewhere

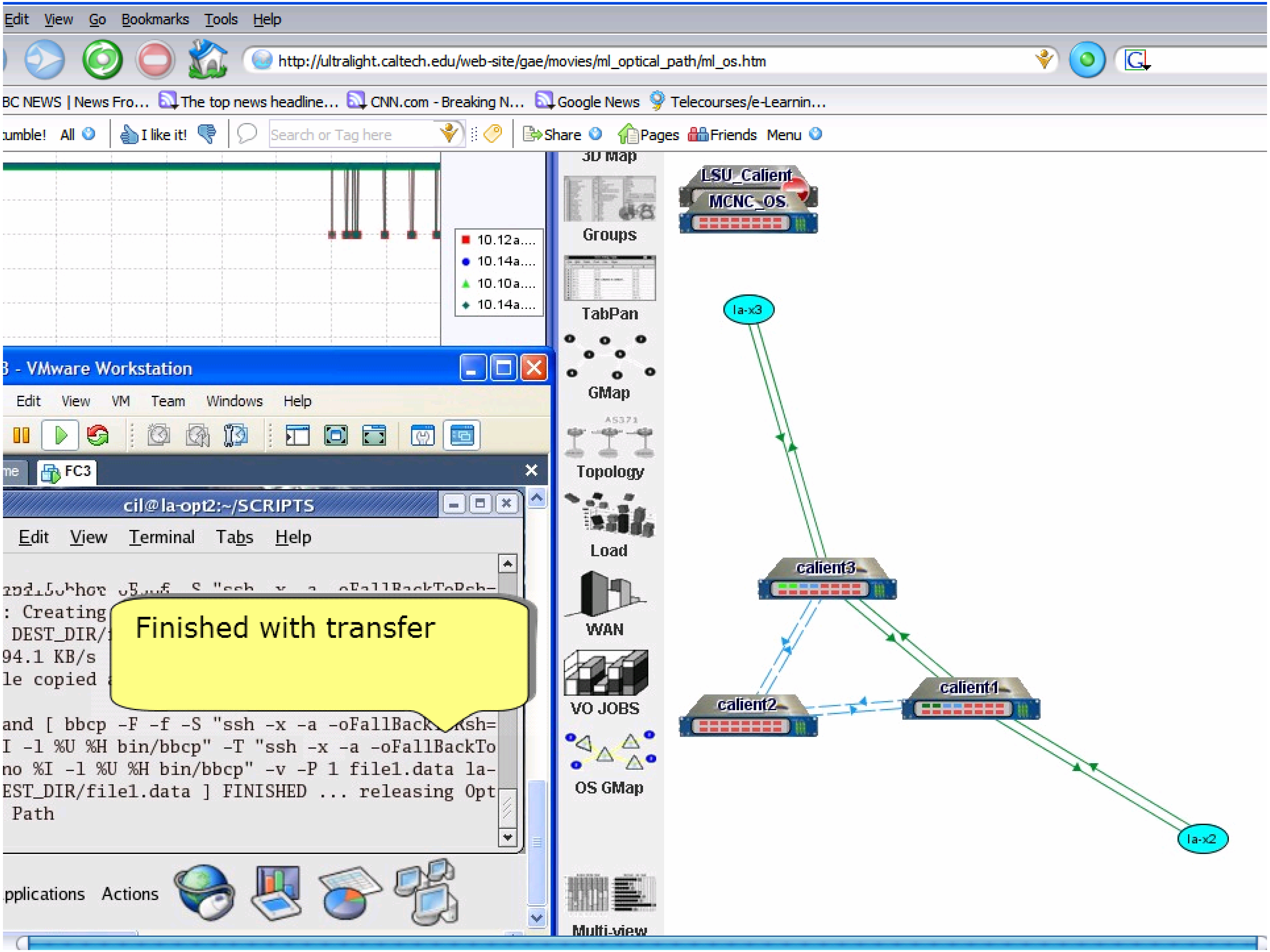

Edit View Go Bookmarks Tools Help
http://ultralight.caltech.edu/web-site/gae/movies/ml_optical_path/ml_os.htm
BC NEWS | News Fro...
The top news headline...
CNN.com - Breaking N...
Google News
Telecourses/e-Learnin...
All
I like it!
Search or Tag here
Share
Pages
Friends
Menu
10.12a....
10.14a....
10.10a....
10.14a....
- VMware Workstation
Edit View VM Team Windows Help
FC3
cil@la-opt2:~/SCRIPTS
Edit View Terminal Tabs Help
and [bbcp -F -f -S "ssh -x -a -oFallBackToRsh=
I -l %U %H bin/bbcp" -T "ssh -x -a -oFallBackTo
no %I -l %U %H bin/bbcp" -v -P 1 /dev/zero la-x
ration took 56 ms
@la-opt2 SCRIPTS]$ ./MLBBCOPY la-x3 file1.data
Copy data from host la-x2 to la-x3
pplications Actions
3D Map
Groups
TabPan
GMap
Topology
Load
WAN
VO JOBS
OS GMap
Multi-view
LSU_Calient
MCNC_OS
la-x3
calient3
calient2
calient1
la-x2

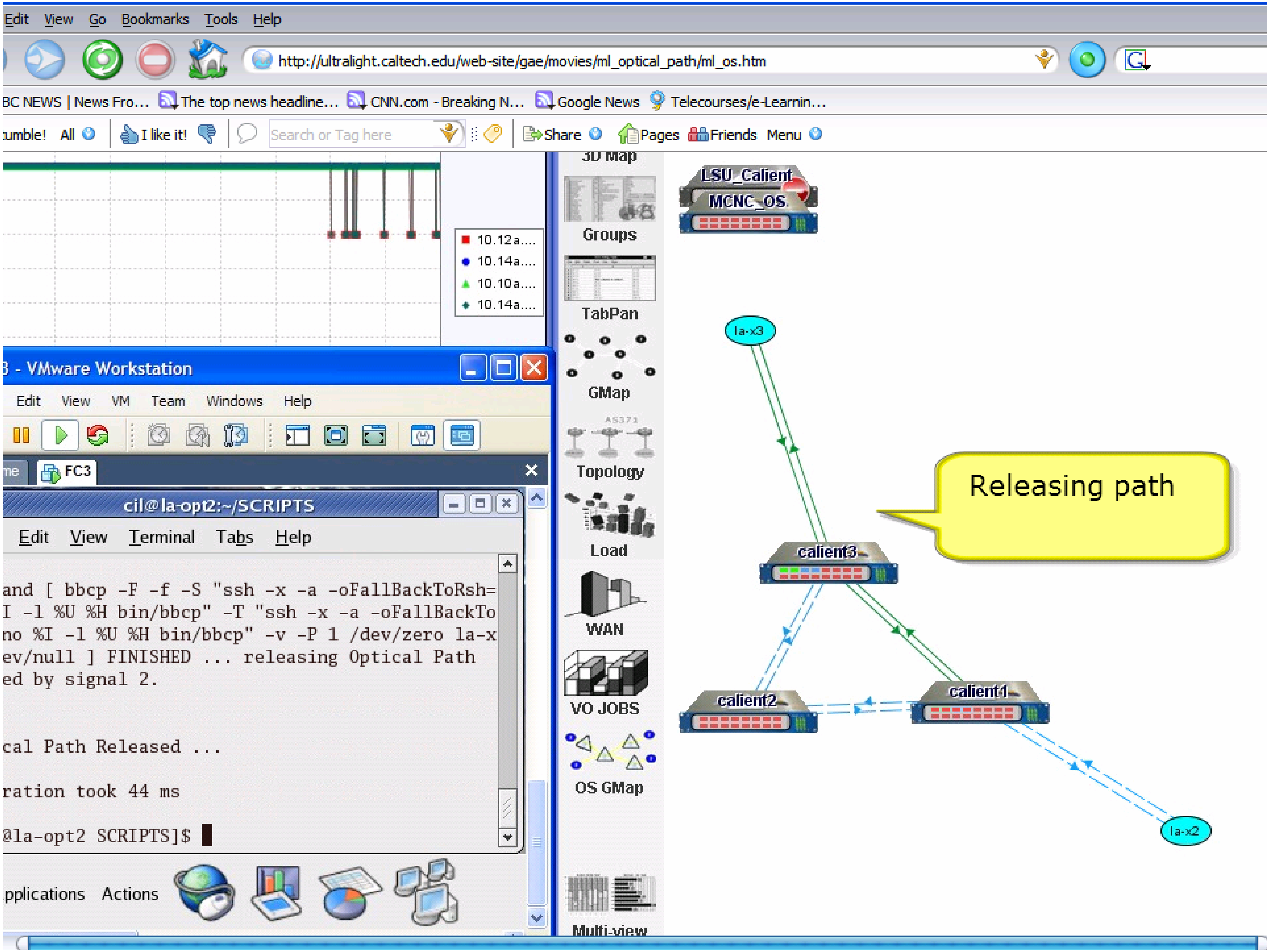

Edit View Go Bookmarks Tools Help
http://ultralight.caltech.edu/web-site/gae/movies/ml_optical_path/ml_os.htm
BC NEWS | News Fro...
The top news headline...
CNN.com - Breaking N...
Google News
Telecourses/e-Learnin...
umble! All
I like it!
Search or Tag here
Share
Pages
Friends
Menu
10.12a....
10.14a....
10.10a....
10.14a....
3 - VMware Workstation
Edit View VM Team Windows Help
FC3
cil@la-opt2:~/SCRIPTS
Edit View Terminal Tabs Help
@la-opt2 SCRIPTS]$ ./MLBBCOPY la-x3 file1.data
cal Path [la-x2 - calient1 - calient3 - la-x3
ration took 50 ms
uting command [bbcp -F -f -S "ssh -x -a -oFall
ToRsh=no %I -l %U %H bin/bbcp" -T "ssh -x -a -o
BackToRsh=no %I -l %U %H bin/bbcp" -v -P 1 file
ta la-x3:DEST_DIR/file1.data]
pplications Actions
3D Map
Groups
TabPan
GMap
Topology
Load
WAN
VO JOBS
OS GMap
Multi-view
LSU_Calient
MCNC_OS
la-x3
calient3
calient2
calient1
la-x2
Global path creation done in less 1 second.
Dynamically creating an optical path (segment 1)

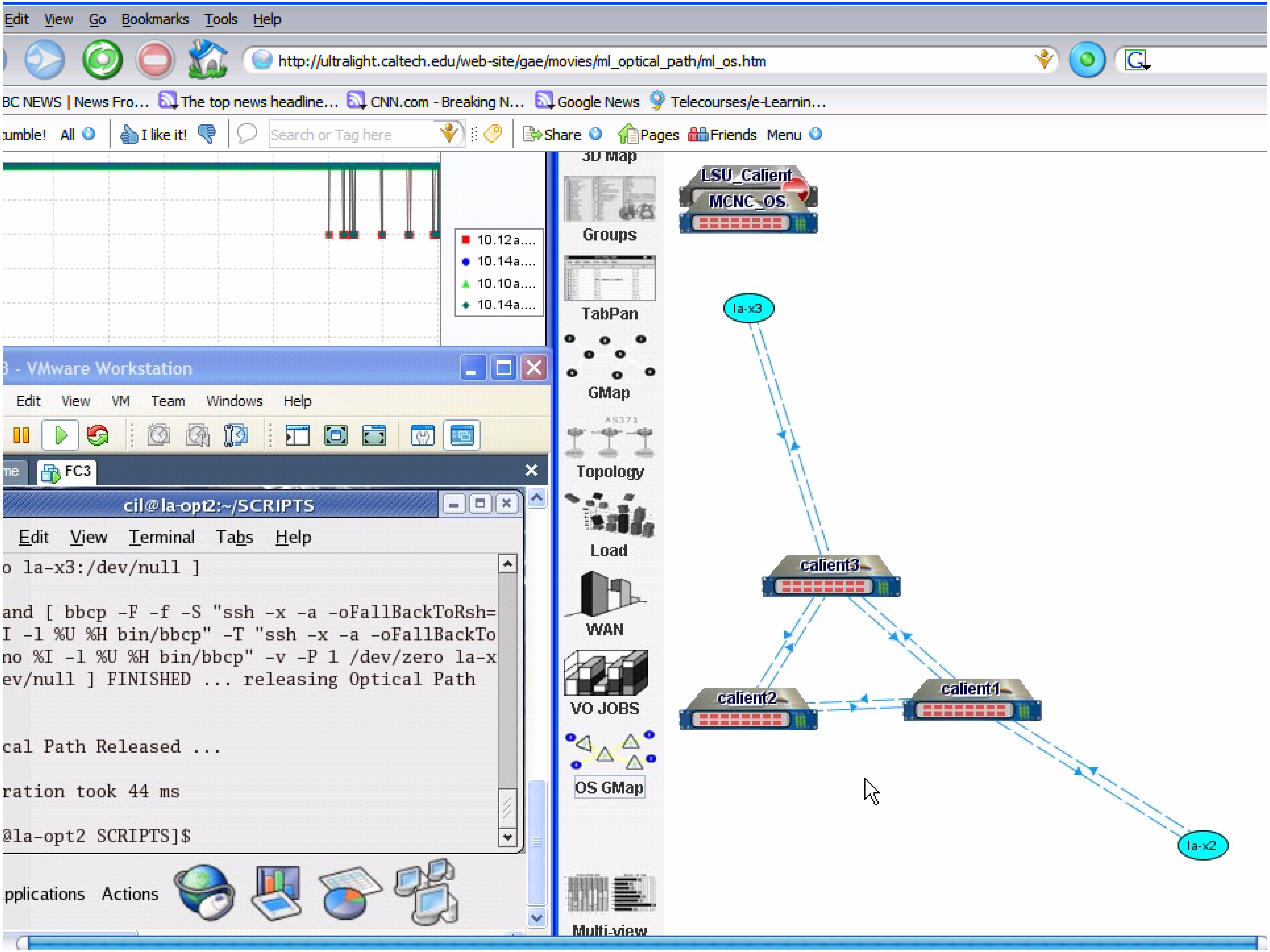

Edit View Go Bookmarks Tools Help
http://ultralight.caltech.edu/web-site/gae/movies/ml_optical_path/ml_os.htm
BC NEWS | News Fro... The top news headline... CNN.com - Breaking N... Google News Telecourses/e-Learnin...
umble! All I like it! Search or Tag here Share Pages Friends Menu
10.12a....
10.14a....
10.10a....
10.14a....
3 - VMware Workstation
Edit View VM Team Windows Help
FC3
cil@la-opt2:~/SCRIPTS
Edit View Terminal Tabs Help
@la-opt2 SCRIPTS]$ ./MLBBCOPY la-x3 file1.data
cal Path [la-x2 - calient1 - calient3 - la-x3
ration took 50 ms
uting command [bbcp -F -f -S "ssh -x -a -oFall
ToRsh=no %I -l %U %H bin/bbcp" -T "ssh -x -a -o
BackToRsh=no %I -l %U %H bin/bbcp" -v -P 1 file
ta la-x3:DEST_DIR/file1.data]
pplications Actions
3D Map
Groups
TabPan
GMap
Topology
Load
WAN
VO JOBS
OS GMap
Multi-view
LSU_Calient
MCNC_OS
la-x3
Creating path segment 2 and 3
calient3
calient2
calient1
la-x2



Edit View Go Bookmarks Tools Help
http://ultralight.caltech.edu/web-site/gae/movies/ml_optical_path/ml_os.htm
3 - VMware Workstation
cil@la-opt2:~/SCRIPTS
@la-opt2 SCRIPTS]$ ./MLBBCOPY la-x3 file1.data
cal Path [la-x2 - calient1 - calient3 - la-x3
ration took 50 ms
uting command [bbcp -F -f -S "ssh -x -a -oFall
ToRsh=no %I -l %U %H bin/bbcp" -T "ssh -x -a -o
BackToRsh=no %I -l %U %H bin/bbcp" -v -P 1 file
ta la-x3:DEST_DIR/file1.data]
Groups
TabPan
GMap
Topology
Load
WAN
VO JOBS
OS GMap
LSU_Calient
MCNC_OS
la-x3
calient3
calient2
calient1
la-x2
Once path is set, transfer data


Edit View Go Bookmarks Tools Help
http://ultralight.caltech.edu/web-site/gae/movies/ml_optical_path/ml_os.htm
The top news headline… CNN.com - Breaking N… Google News Telecourses/e-Learnin…
I like it! Share Pages Friends Menu
10.12a….
10.14a….
10.10a….
10.14a….
VMware Workstation
Edit View VM Team Windows Help
FC3
cil@la-opt2:~/SCRIPTS
Edit View Terminal Tabs Help
: Creating
DEST_DIR/
94.1 KB/s
le copied
and [bbcp -F -f -S "ssh -x -a -oFallBack
I -l %U %H bin/bbcp" -T "ssh -x -a -oFallBackTo
no %I -l %U %H bin/bbcp" -v -P 1 file1.data la-
EST_DIR/file1.data] FINISHED ... releasing Opt
Path
Finished with transfer
Applications Actions
3D Map
Groups
TabPan
GMap
Topology
Load
WAN
VO JOBS
OS GMap
Multi-view
LSU_Calient
MCNC_OS
la-x3
calient3
calient2
calient1
la-x2


Edit View Go Bookmarks Tools Help
http://ultralight.caltech.edu/web-site/gae/movies/ml_optical_path/ml_os.htm
BC NEWS | News Fro... The top news headline... CNN.com - Breaking N... Google News Telecourses/e-Learnin...
umble! All I like it! Search or Tag here Share Pages Friends Menu
10.12a....
10.14a....
10.10a....
10.14a....
3 - VMware Workstation
Edit View VM Team Windows Help
me FC3
cil@la-opt2:~/SCRIPTS
Edit View Terminal Tabs Help
and [bbcp -F -f -S "ssh -x -a -oFallBackToRsh=
I -l %U %H bin/bbcp" -T "ssh -x -a -oFallBackTo
no %I -l %U %H bin/bbcp" -v -P 1 /dev/zero la-x
ev/null] FINISHED ... releasing Optical Path
ed by signal 2.
cal Path Released ...
ration took 44 ms
@la-opt2 SCRIPTS]$
pplications Actions
3D Map
Groups
TabPan
GMap
Topology
Load
WAN
VO JOBS
OS GMap
Multi-view
LSU_Calient
MCNC_OS
la-x3
Releasing path
calient3
calient2
calient1
la-x2



Edit View Go Bookmarks Tools Help
http://ultralight.caltech.edu/web-site/gae/movies/ml_optical_path/ml_os.htm
BC NEWS | News Fro... The top news headline... CNN.com - Breaking N... Google News Telecourses/e-Learnin...
umble! All I like it! Search or Tag here Share Pages Friends Menu
10.12a....
10.14a....
10.10a....
10.14a....
3 - VMware Workstation
Edit View VM Team Windows Help
FC3
cil@la-opt2:~/SCRIPTS
Edit View Terminal Tabs Help
o la-x3:/dev/null]
and [bbcp -F -f -S "ssh -x -a -oFallBackToRsh=
I -l %U %H bin/bbcp" -T "ssh -x -a -oFallBackTo
no %I -l %U %H bin/bbcp" -v -P 1 /dev/zero la-x
ev/null] FINISHED ... releasing Optical Path
cal Path Released ...
ration took 44 ms
@la-opt2 SCRIPTS]$
pplications Actions
3D Map
Groups
TabPan
GMap
Topology
Load
WAN
VO JOBS
OS GMap
Multi-view
LSU_Calient
MCNC_OS
la-x3
calient3
calient2
calient1
la-x2

UltraLight Storage R&D Global Planning Services

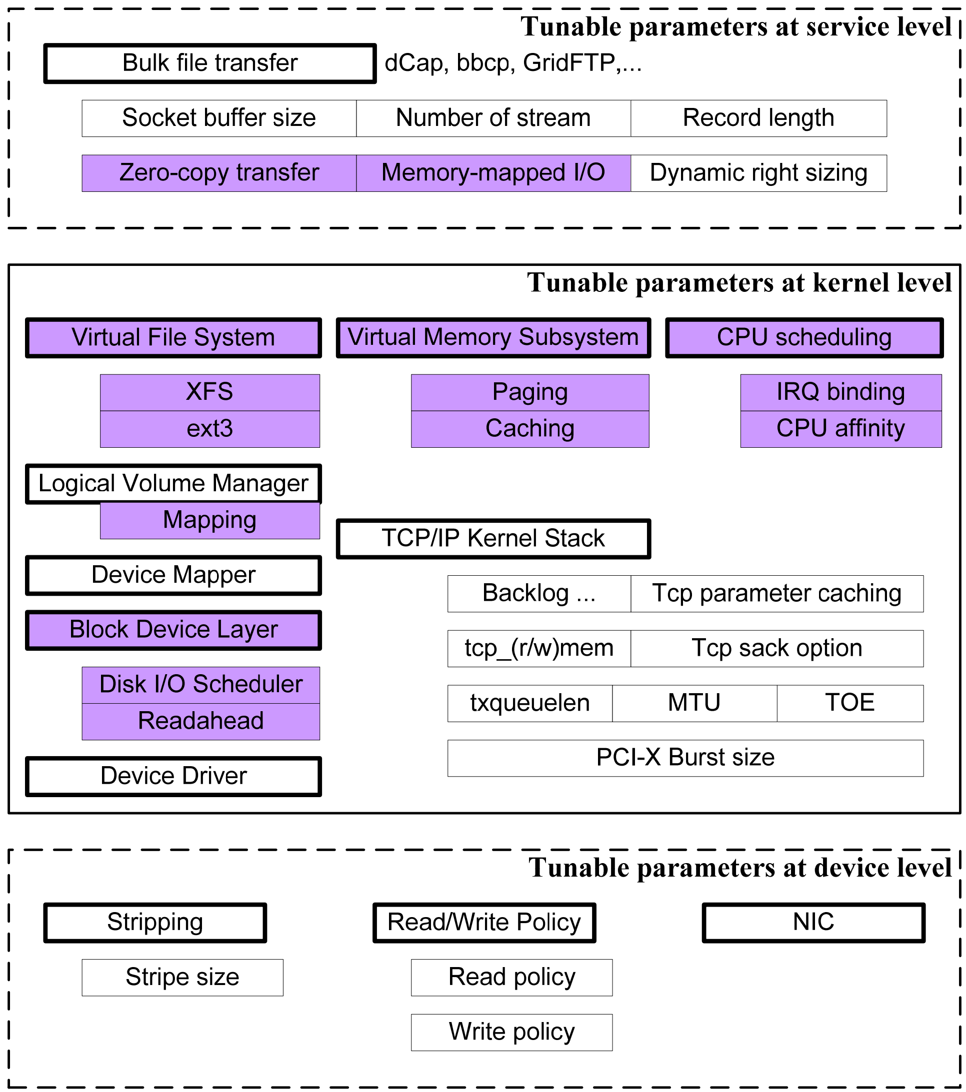
- UltraLight (optical networks in general) moving towards a managed “control plane”
  - Light-paths will be allocated/scheduled to data-flow requests via policy based priorities, queues, and advanced reservations
  - Clear need to match “Network Resource Management” with “Storage Resource Management”
    - Well known co-scheduling problem!
    - In order to develop an effective NRM, must understand and interface with SRM!
- End systems remain the current bottle-neck for large scale data transport over the WAN
  - Key to effective filling/draining of the pipe
  - Need highly capable hardware (servers, etc)
  - Need carefully tuned software (kernel, etc)

UltraLight using a Wholistic, Multi-level Approach

- End-host Device Technologies
  - Choosing right H/W platform for the price ($20K)
- End-host Software Stacks
  - Tuning storage server for stable and high throughput
- End-Systems Management
  - Specifying quality of service (QoS) model for Ultralight storage server
  - SRM/dCache
  - LSTORE (& SRM/LSTORE)
- Wide Area Testbeds (REDDnet)

Tunable Parameter Space

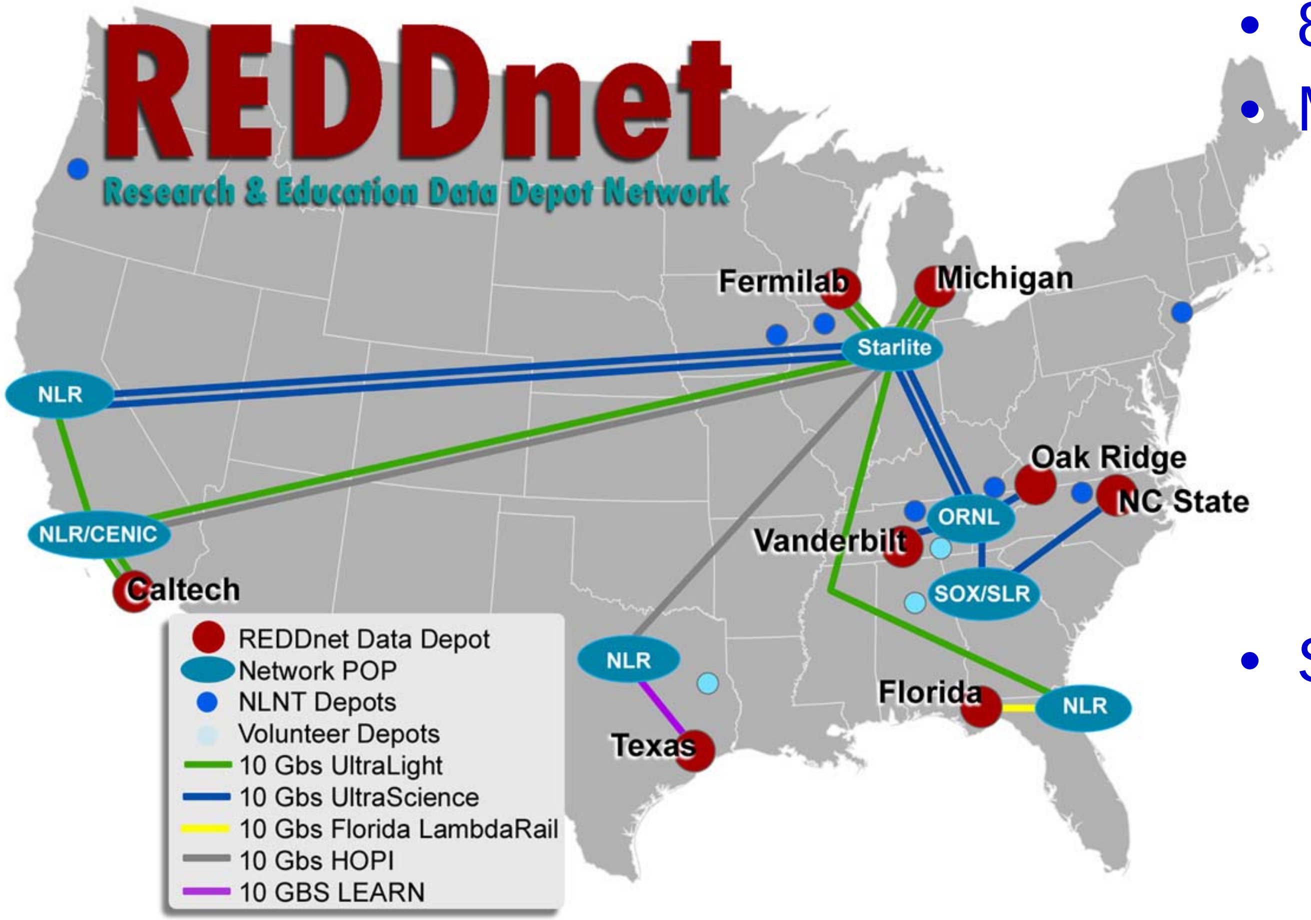
- Multiple layers
  - Service/AP level
  - Kernel level
  - Device level
- Complexity of tuning
  - Fine tuning is very complex task
  - Now investigating the possibility of *auto-tuning* daemon for storage server



REDDnet

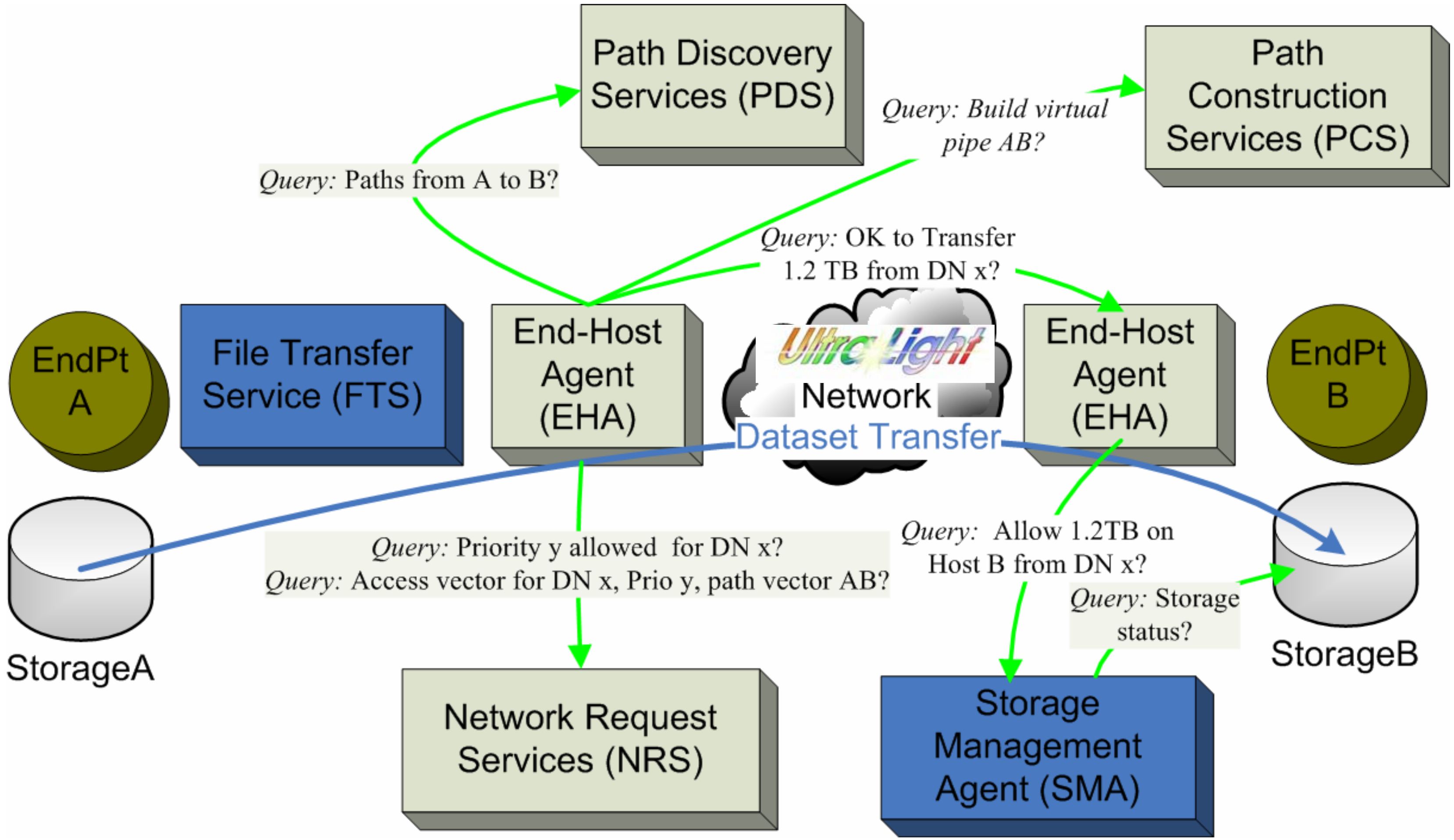
Research and Education Data Depot Network

Powered by UltraLight



- Led by Vanderbilt
- 8 initial sites
- Multiple disciplines
  - Sat imagery (AmericaView)
  - HEP
  - Terascale Supernova Initative
  - Structural Biology
  - Bioinformatics
- Storage
  - 500TB disk
  - 200TB tape

Sets the Stage for Testing Fully Integrated System



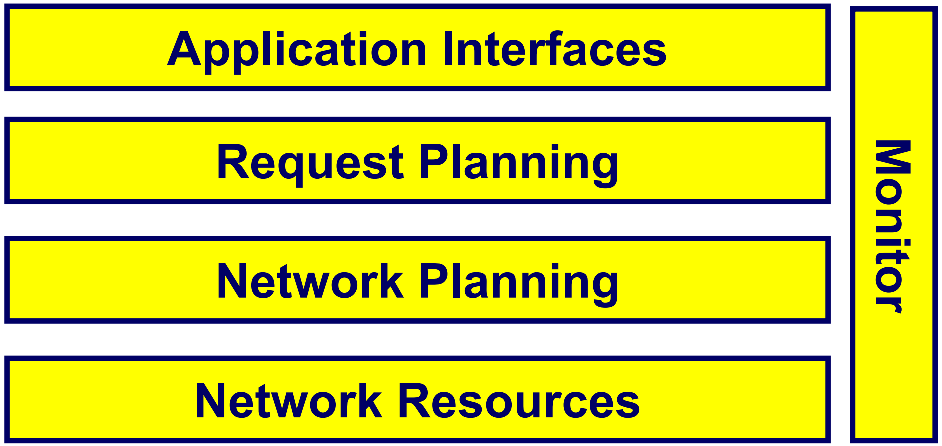
Summary

- UltraLight is a global Laboratory, uniquely positioned
  - Spans Tier-0, some Tier-1s, several Tier-2s, and some Tier-3s
  - Includes participation from ATLAS (not discussed in this talk)
- End-hosts remain serious bottleneck in delivering CMS data to the higher Tiers for physics data analysis
  - Human in the loop problem
  - Incapable hardware (sometimes, perhaps even often)
  - Fine tuning of services
- UltraLight working to address these (generic) problems by
  - Researching and developing Global Planning Services
  - Using a wholistic approach (devices, parameters, services, WAN)

- Final thought:
  - Not only critical for LHC, also important for preparing for SLHC! (HEP always asks: how much time to 4x our data sample?)

Make UltraLight available to Physics applications and their environments

- Unpredictable multi user analysis
- Overall demand typically fills the capacity of the resources
- Real time monitor systems for networks, storage, computing resources,… : E2E monitoring



Support data transfers ranging from predictable movement of large scale (simulated and real) data, to highly dynamic analysis tasks initiated by rapidly changing teams of scientists