Lessons learned from TeraGrid and I-WIRE

Engineering and Evaluation

Tony Rimovsky

NCSA

Lessons Learned

Three major lessons

- Be prepared to drop old ideas
- Keep the goals in mind. Keep asking what the goals are.
- Lean on experience

Extensible TeraGrid Overview

~23 Teraflop distributed cluster

- NCSA, Caltech, SDSC, ANL under TeraGrid
- Added PSC under ETF

Fastest research network in the US.

- 40 Gigabit/s (4xOC-192c) Chicago <-> LA
- 30 Gigabit/s to each site
- Using both CENIC and I-WIRE regional optical networks

NCSA

Be prepared to drop old ideas

Original TeraGrid network concept was full-mesh, point-to point 10Gigabit Ethernet connectivity using Ethernet switches between 4 sites.

Key Points:

- Qwest provided OC-192c lambdas
- I-WIRE and CENIC were both evaluating optical designs
- TeraGrid networkers involved in both I-WIRE and CENIC engineering

Technical Evaluation

Several problems were discovered:

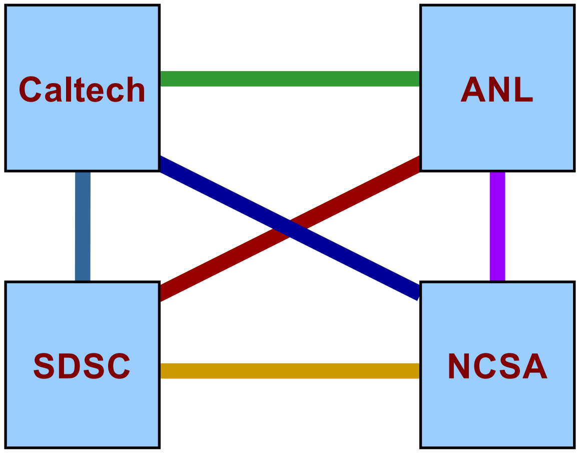
- SONET vs Ethernet industry preferences
  - Theoretically 10GigE could work over both LAN and OC-192
  - Ethernet vendors prioritized for LAN
  - SONET vendors prioritized for WAN
- 10GigE interfaces and switches didn’t exist yet.
- Buffering in Ethernet switches was going to be an issue

NCSA

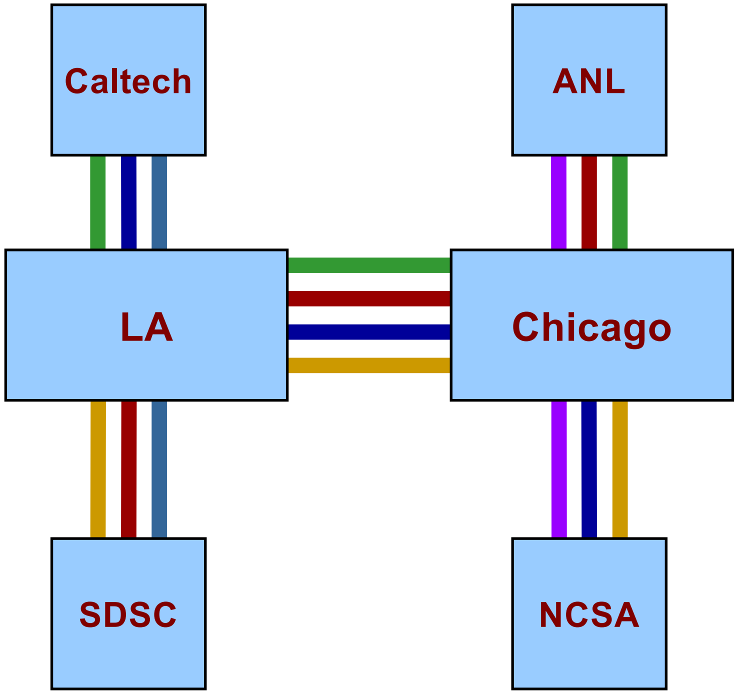
Keep the goals in mind

(and keep checking the goals)

Original concept was 4 sites and cluster traffic as in a machine room, with dedicated links in full-mesh.



Logical Lambda Topology
(Full Mesh)



Physical Lambda Topology

Goal Evolution

Priorities and goals developed over time

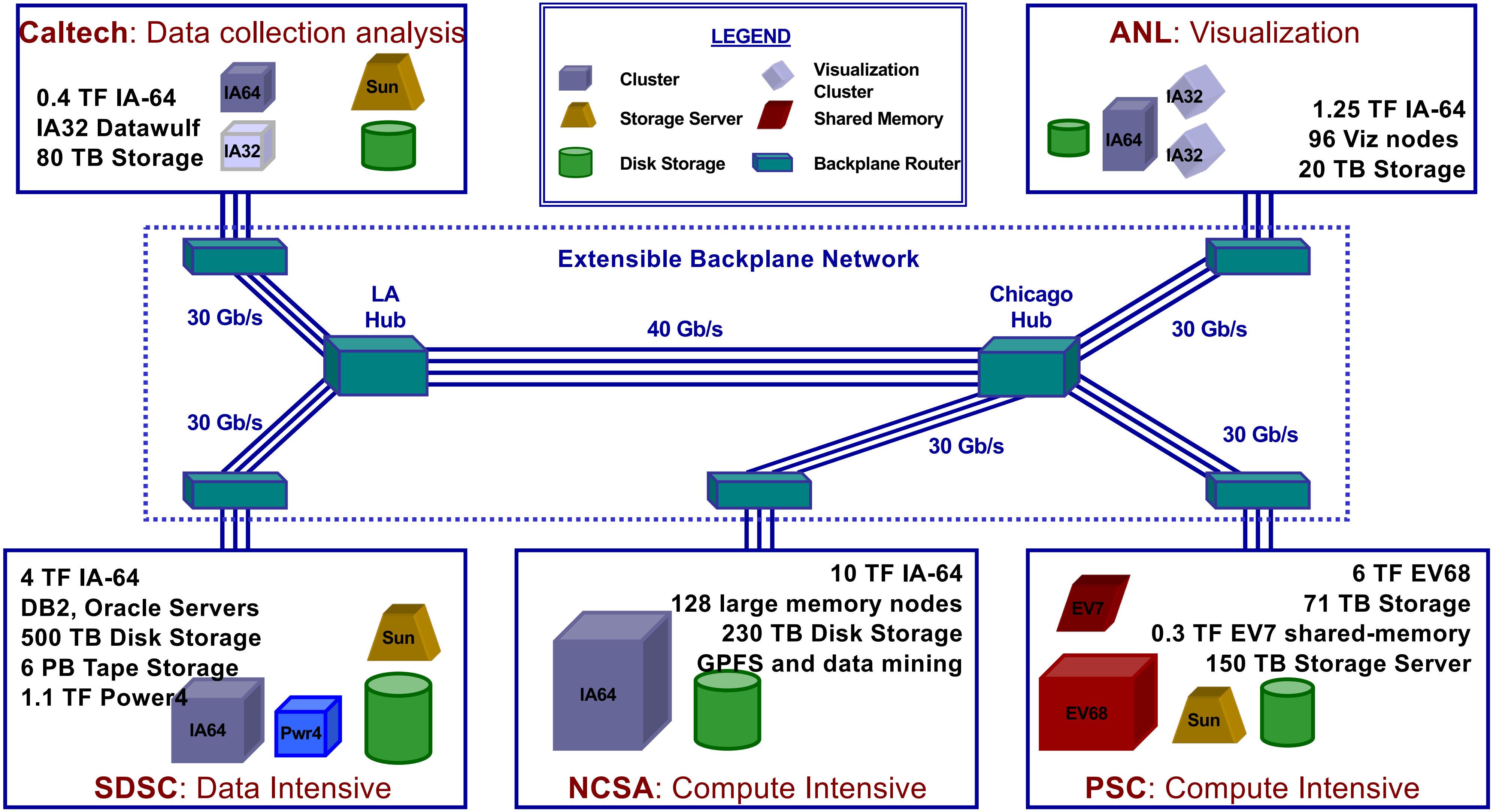
- Desire to maximize bandwidth usability among sites
- The network is not the research/risk area for ETF
- Suddenly needed to scale to more sites

As a result

- IP core put in place over the lambda topology
- Still able to experiment technology like SANs over lambdas and over IP
- PSC was added with minimal additional networking cost

Extensible TeraGrid Facility (ETF)

Proposed, 2002, Operational in 2003



Caltech: Data collection analysis

0.4 TF IA-64
IA32 Datawulf
80 TB Storage

LEGEND

Cluster		Visualization Cluster	
Storage Server		Shared Memory	
Disk Storage		Backplane Router	

ANL: Visualization

1.25 TF IA-64
96 Viz nodes
20 TB Storage

Extensible Backplane Network

4 TF IA-64
DB2, Oracle Servers
500 TB Disk Storage
6 PB Tape Storage
1.1 TF Power4

SDSC: Data Intensive

10 TF IA-64
128 large memory nodes
230 TB Disk Storage
GPFS and data mining

NCSA: Compute Intensive

6 TF EV68
71 TB Storage
0.3 TF EV7 shared-memory
150 TB Storage Server

PSC: Compute Intensive

I-WIRE Note

Evolution also affected I-WIRE development.

We kept finding new or better fiber deals

- Changed the topology, which changed the engineering/design.
- Fiber was driving the design and the priorities for a while.

This was both good…

We found opportunities we originally didn't know existed

… and bad

Design team had to keep starting over

NCSA

Lean on Experience

We brought in other people for I-WIRE and TeraGrid hardware evaluations

Valuable for several reasons

- In the I-WIRE case, we brought in people who had operational experience with optical nets.
- With TeraGrid, we had the routing people from four major sites involved, and asked for input on questions to ask from other teams.

Value of Experience

- **Experience with specific vendors**

  A lot of promises get made at the leading edge. Reputation and experience mean a lot.

- **Experience with technology**

  Having people with optical experience on the I-WIRE evaluation team helped tremendously.

- **Experience with evaluating proposals**

  Learned better ways to evaluate technical proposals.

NCSA

Conclusion

Questions?

- Teragrid: networking-wg@teragrid.org
- I-WIRE: iwire-eng@mcs.anl.gov
- Tony Rimovsky: tony@ncsa.edu