

TeraGrid Architecture

Charlie Catlett
Argonne National Laboratory
Executive Director, TeraGrid Project
Chair, Global Grid Forum

May 2002

TeraGrid Objectives

- Create unprecedented capability
 - integrated with extant PACI capabilities
 - supporting a new class of scientific research
- Deploy a balanced, distributed system
 - not a “distributed computer” but rather
 - a distributed “system” using Grid technologies
 - computing and data management
 - visualization and scientific application analysis
- Define an open and extensible infrastructure
 - an “enabling cyberinfrastructure” for scientific research
 - extensible beyond the original four sites
 - NCSA, SDSC, ANL, and Caltech



Strategies: An Extensible Grid

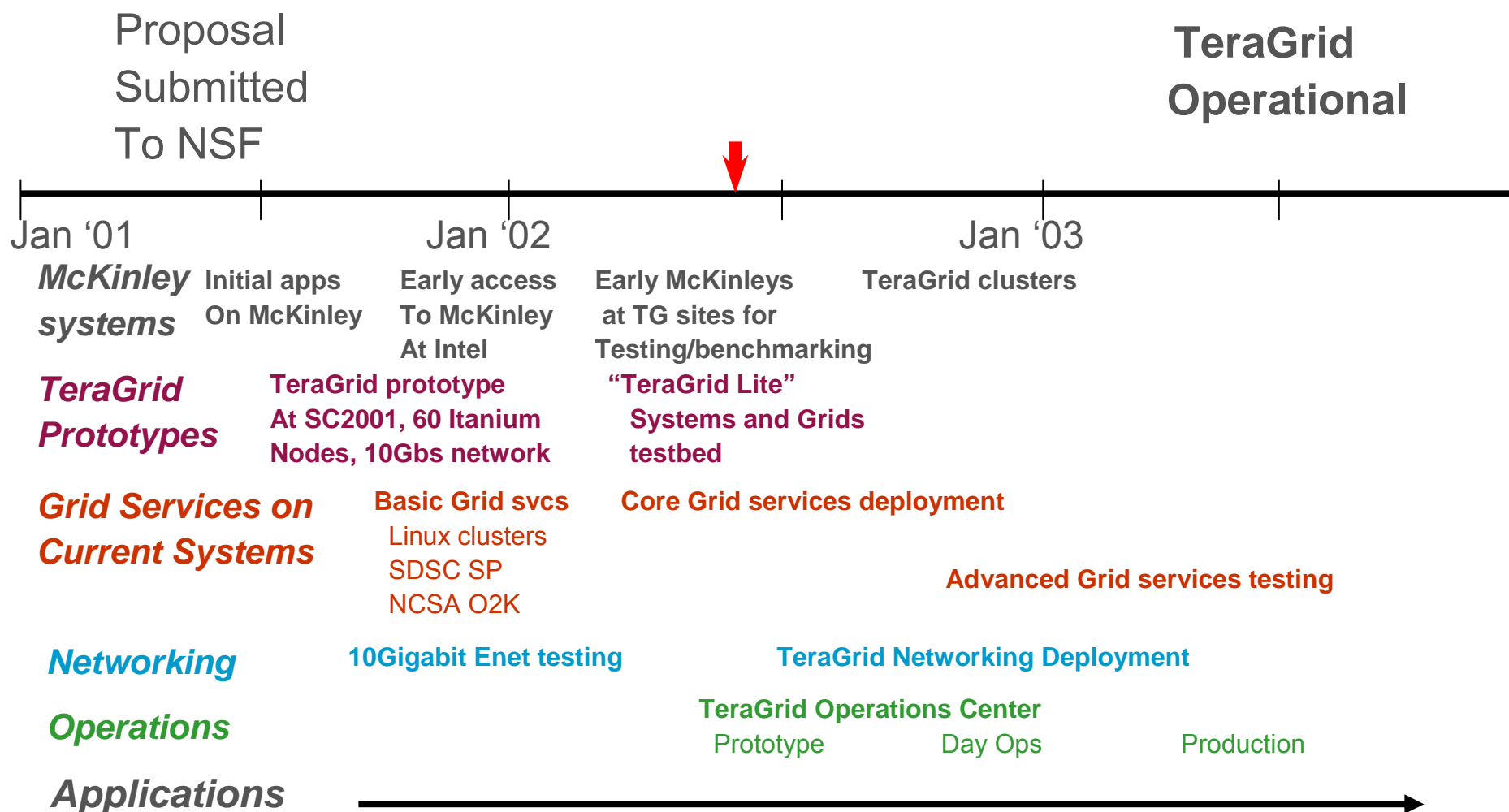


- Achievable goals
 - exploit DTF homogeneity where possible
 - accelerate deployment and lower user entry barriers
 - example: enabling program executable portability among TeraGrid sites
 - leverage GGF and software efforts such as NSF NMI, Condor
 - TeraGrid serves as a driver for NMI functionality and schedule
 - defining TeraGrid specifications via GGF
 - Partnerships with PPDG, GriPhyn, European DataGrid, others
- Design for expansion
 - focus on integrating “resources” rather than “sites”
 - adding resources will require significant, but not unreasonable, effort
 - supporting key protocols and specifications (e.g. authorization, accounting)
 - requires software implementation or porting to local platforms
 - supporting heterogeneity while exploiting homogeneity
 - balancing complexity and uniformity

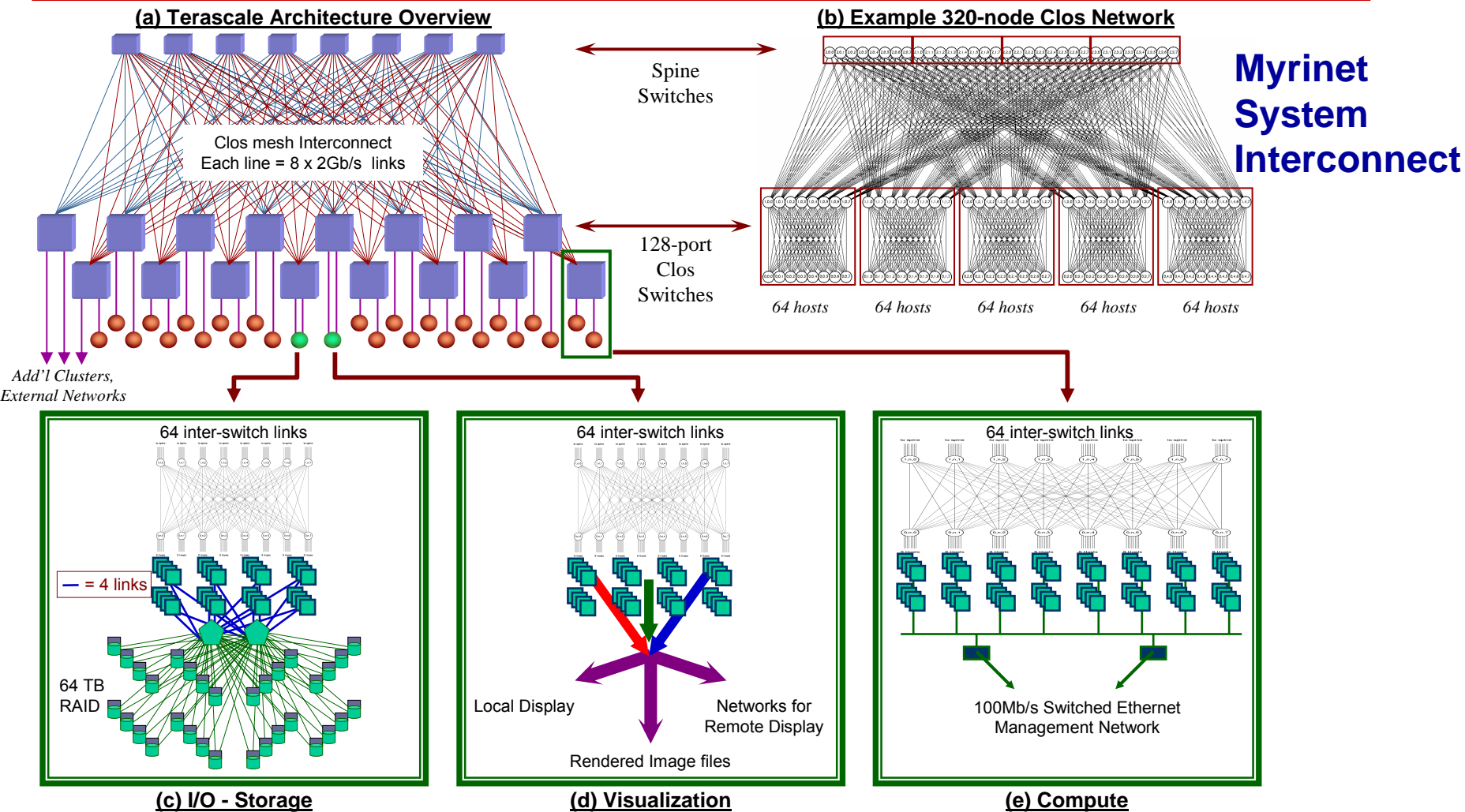
Condor Project is Doing Well in Europe



TeraGrid Timelines

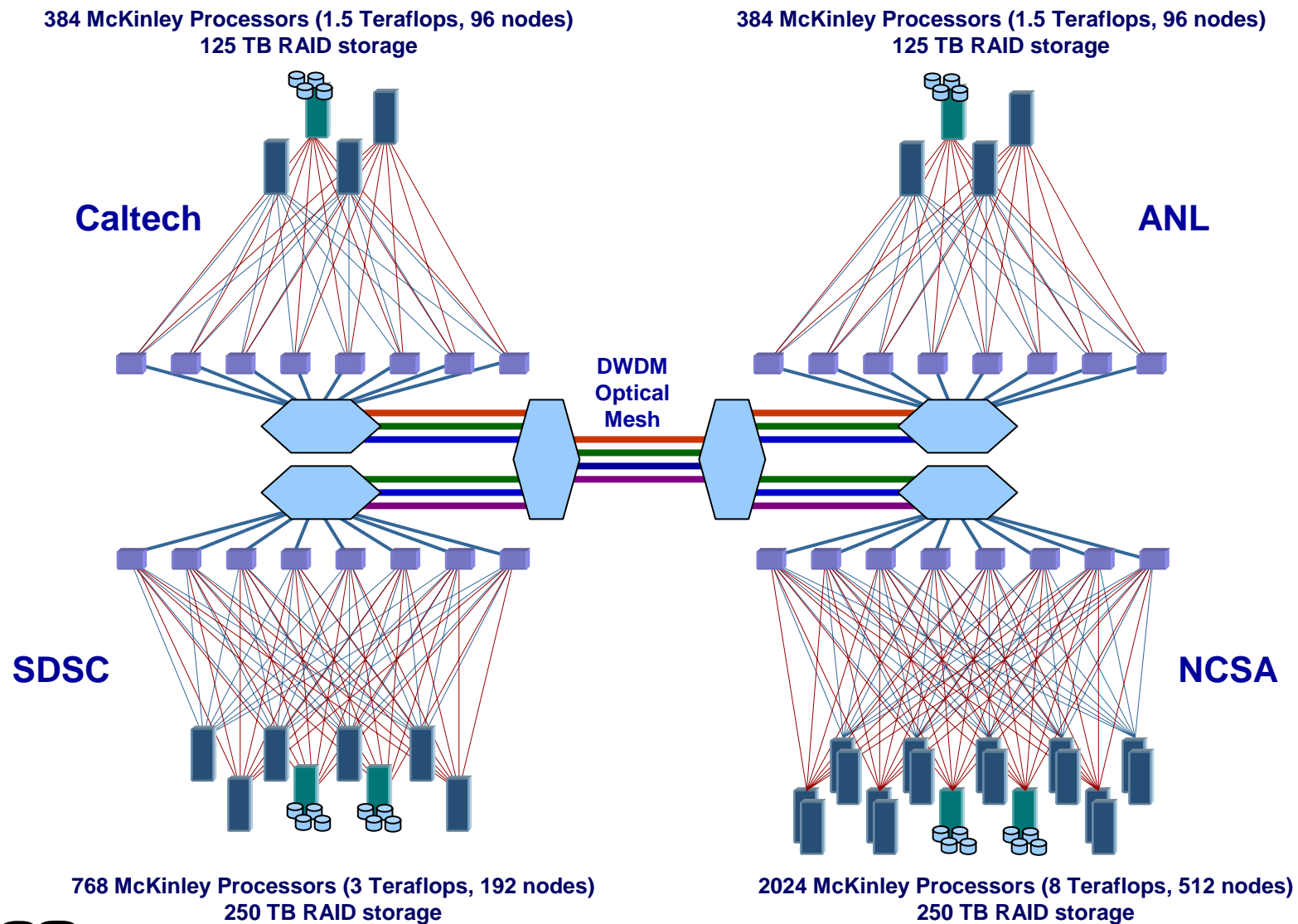


Terascale Cluster Architecture

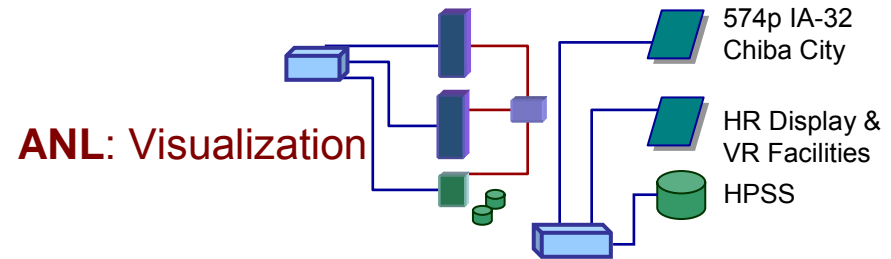
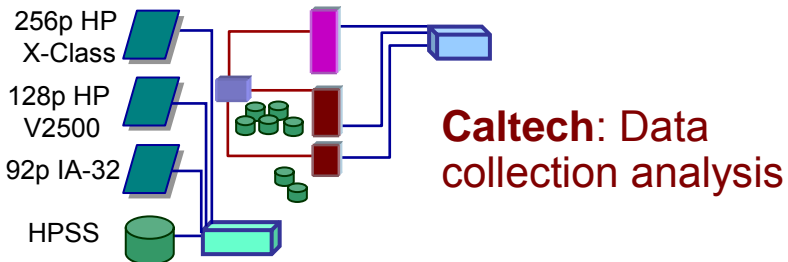


- FCS Storage Network
- GbE for external traffic

OK... What about the WAN?



NSF TeraGrid: 14 TFLOPS, 750 TB

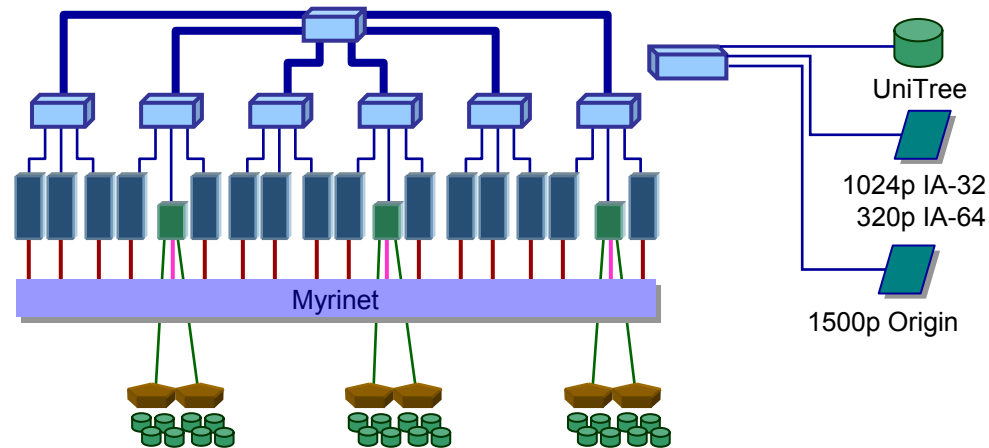
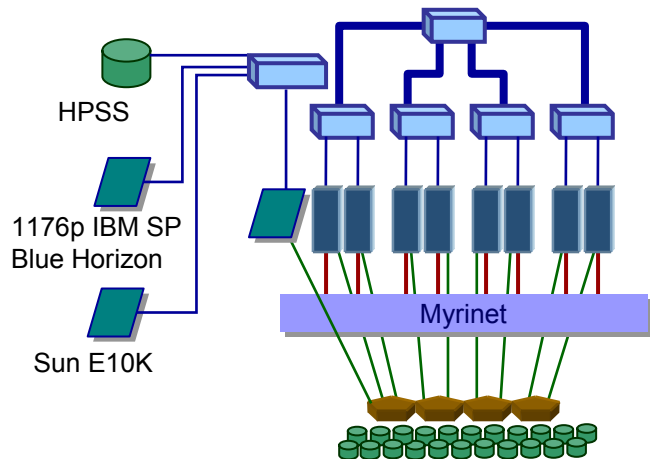


WAN Architecture Options:

- Myrinet-to-GbE; Myrinet as a WAN
- Layer2 design
- Wavelength Mesh
- Traditional IP Backbone

WAN Bandwidth Options:

- Abilene (2.5 Gb/s, →10Gb/s late 2002)
- State and regional fiber initiatives plus CANARIE CA*Net
- Leased OC48
- Dark Fiber, Dim Fiber, Wavelengths



NSFNET 56 Kb/s Site Architecture

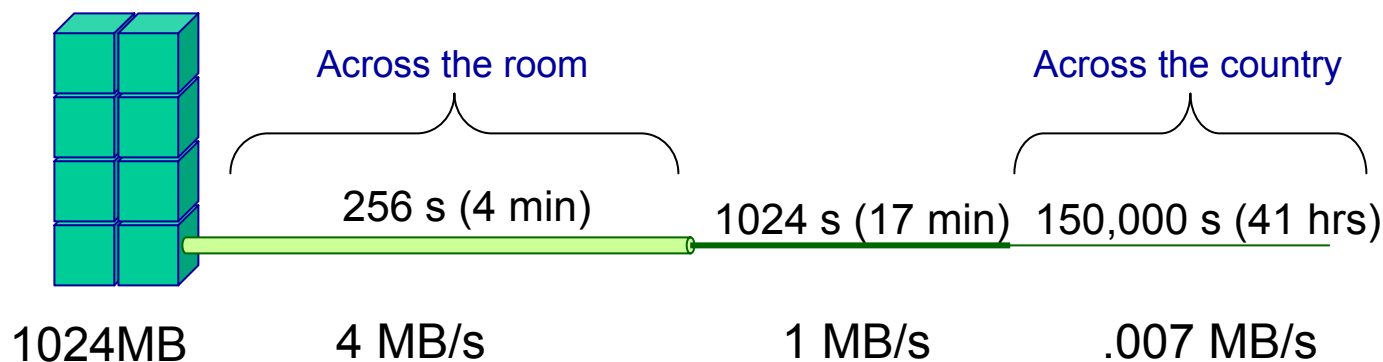


Bandwidth in terms of **burst** data transfer and user **wait time**.

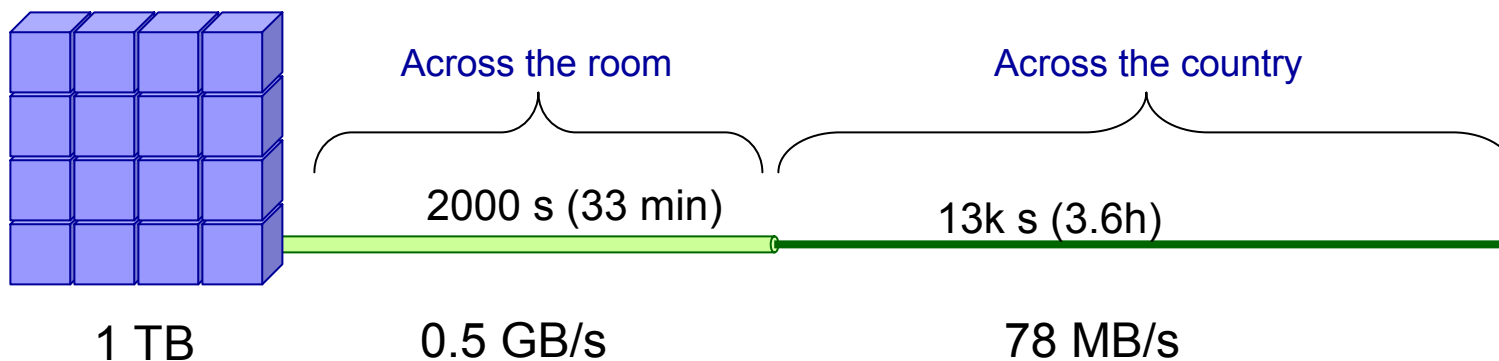
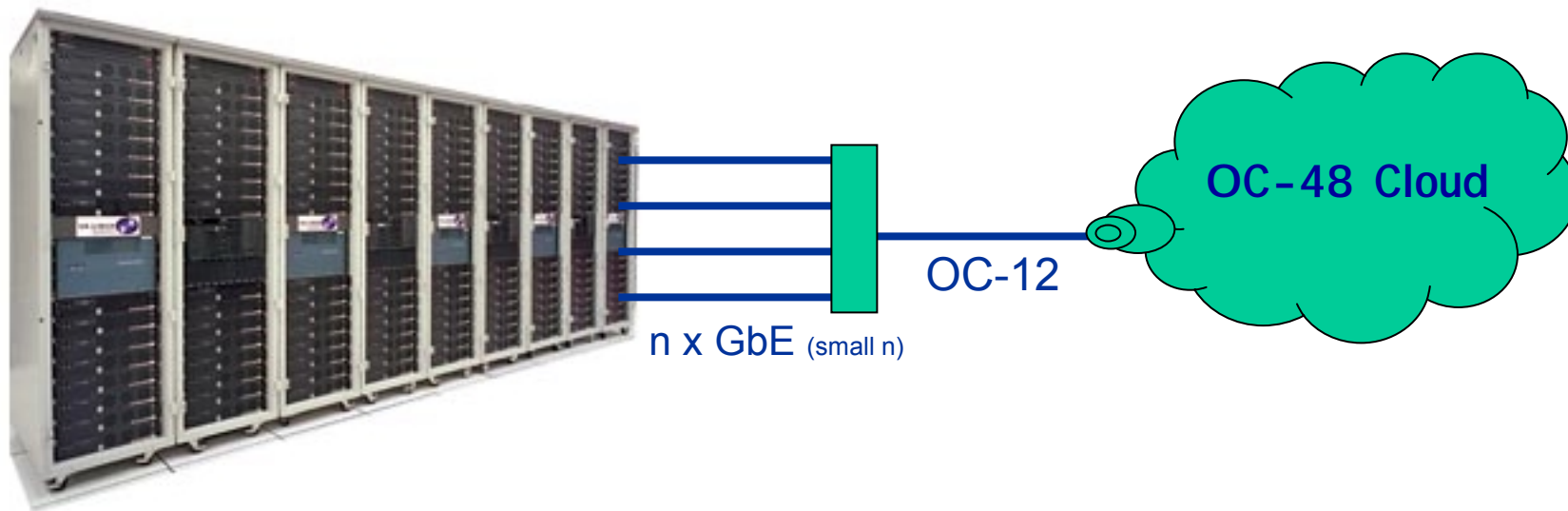


Fuzzball

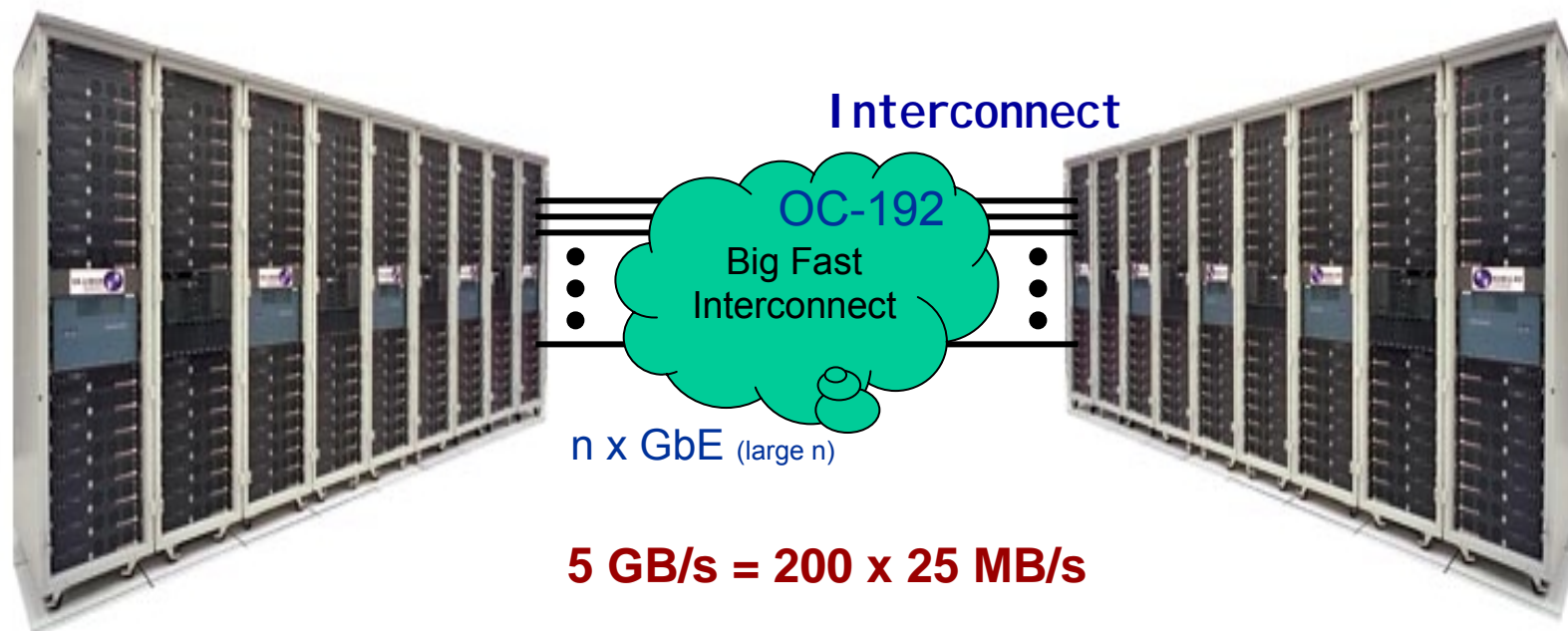
VAX



2002 Cluster-WAN Architecture

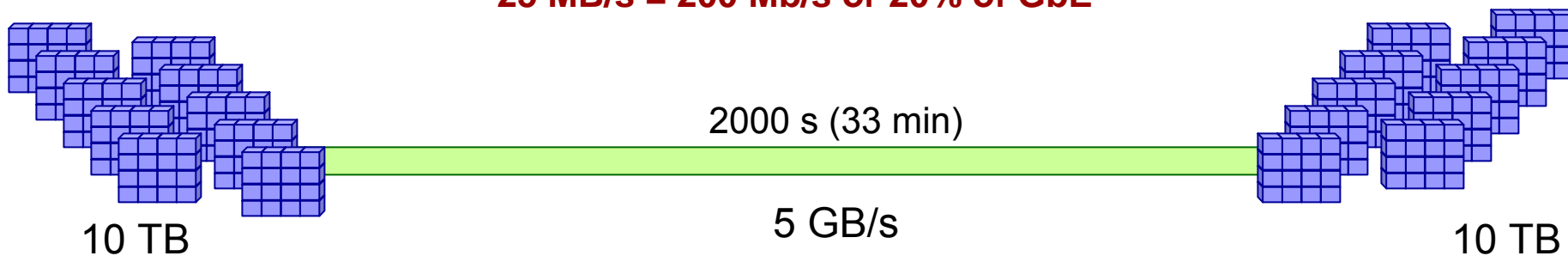


To Build a Distributed Terascale Cluster...



5 GB/s = 200 x 25 MB/s

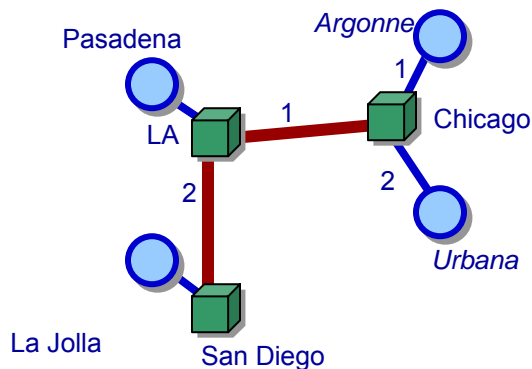
25 MB/s = 200 Mb/s or 20% of GbE



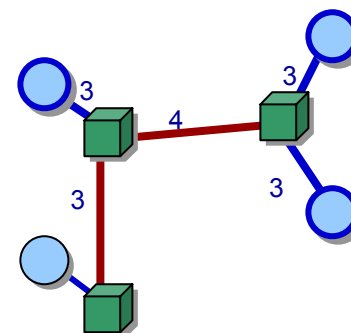
TeraGrid Interconnect: Qwest Partnership

Physical
denotes λ count

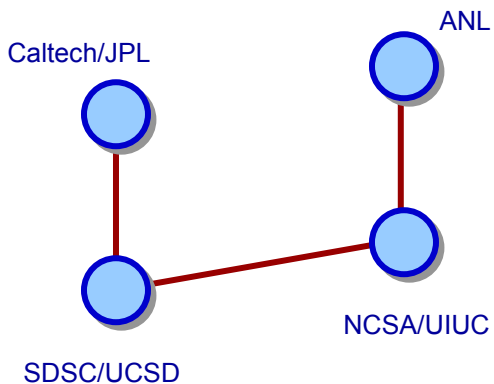
Phase 0 (June 2002)



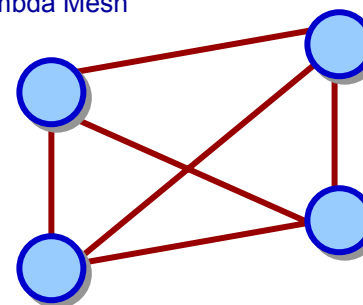
Phase 1 (November 2002)



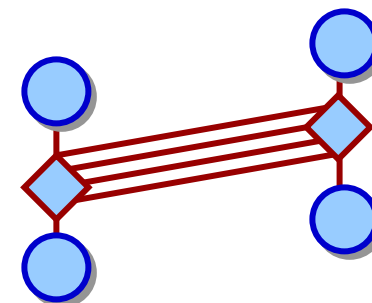
Light Paths (Logical)



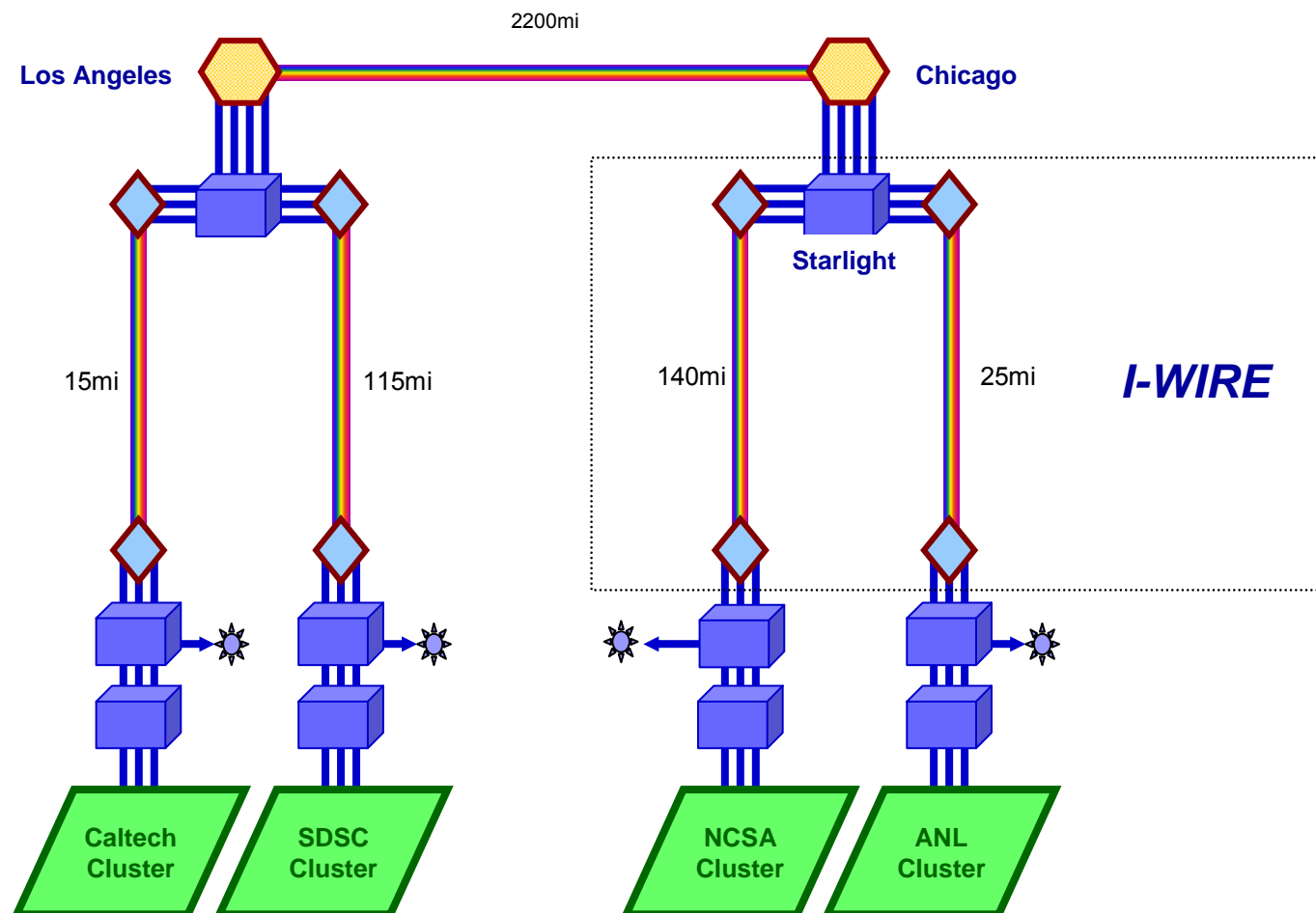
Original: Lambda Mesh



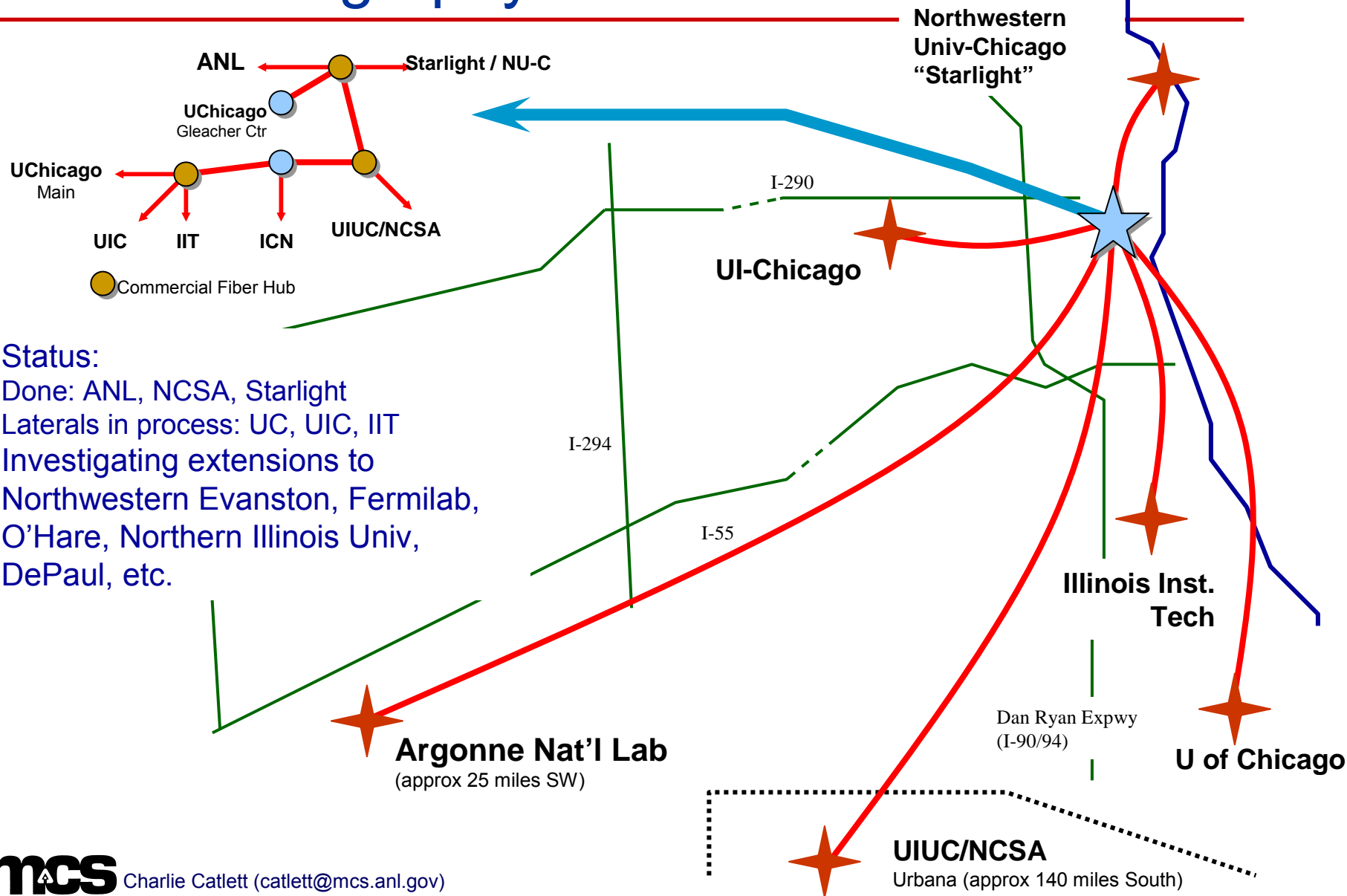
Extensible: Central Hubs



But You can't buy a lambda to Urbana...

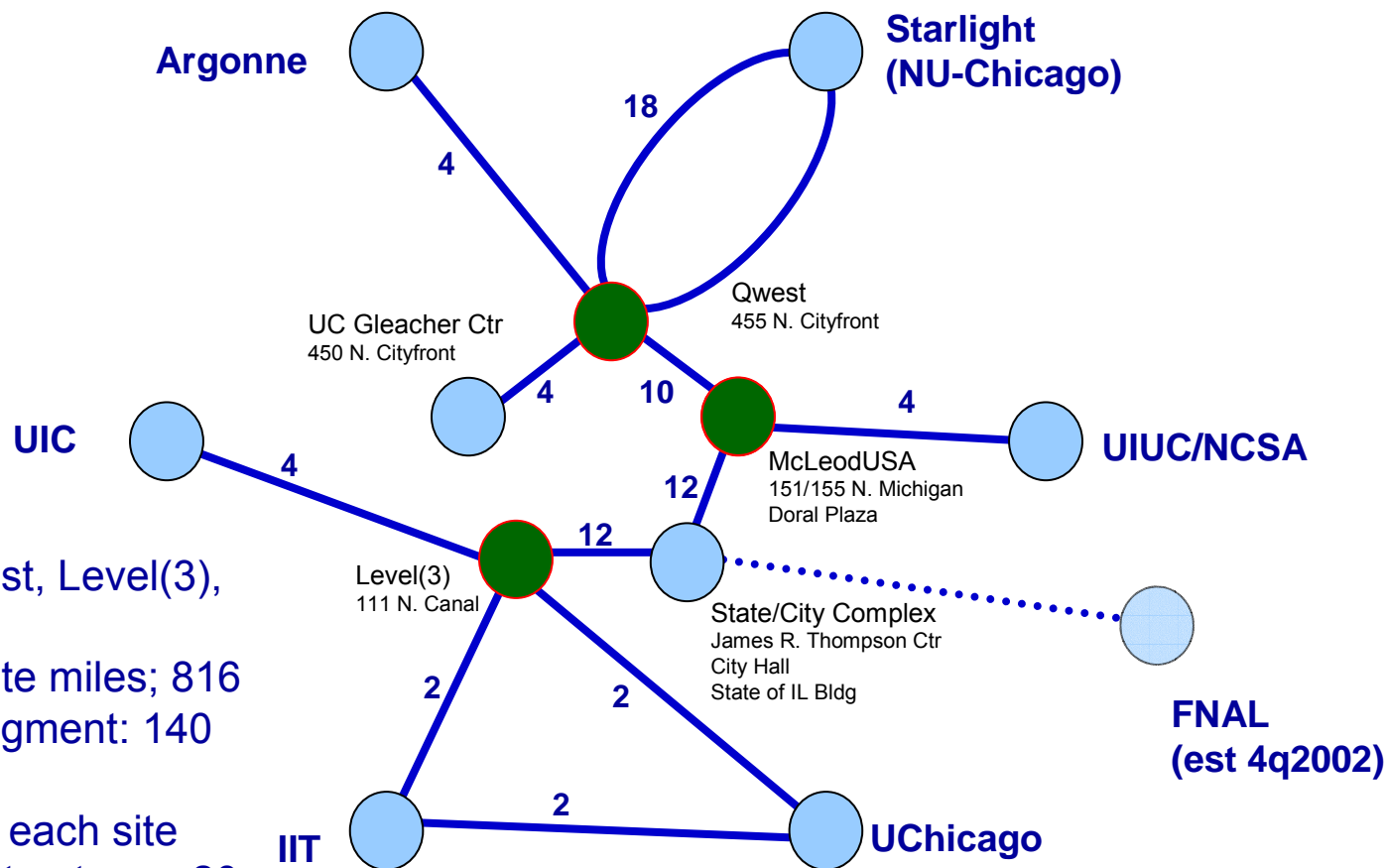


I-WIRE Geography



- Status:
- Done: ANL, NCSA, Starlight
- Laterals in process: UC, UIC, IIT
- Investigating extensions to Northwestern Evanston, Fermilab, O'Hare, Northern Illinois Univ, DePaul, etc.

I-Wire Fiber Topology



- Fiber Providers: Qwest, Level(3), McLeodUSA
- 10 segments, 190 route miles; 816 fiber miles, longest segment: 140 miles
- 4 strands minimum to each site
- ~\$4M for fiber- all contracts are 20 year "IRU"

Numbers indicate fiber count (strands)

I-Wire Transport

TeraGrid Linear

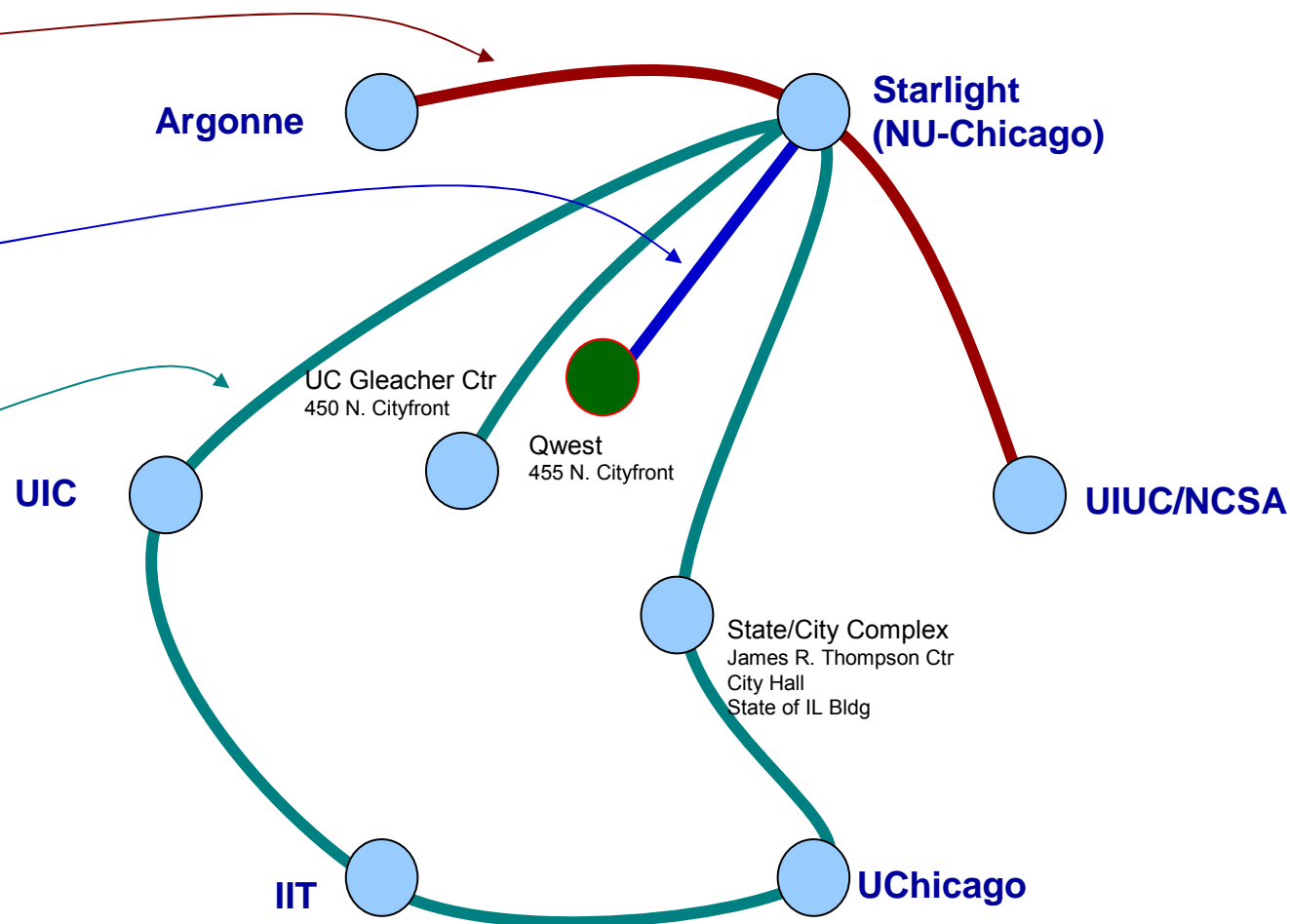
3x OC192
1x OC48
First light: 6/02

Starlight Linear

4x OC192
4x OC48 (→8x GbE)
Operational

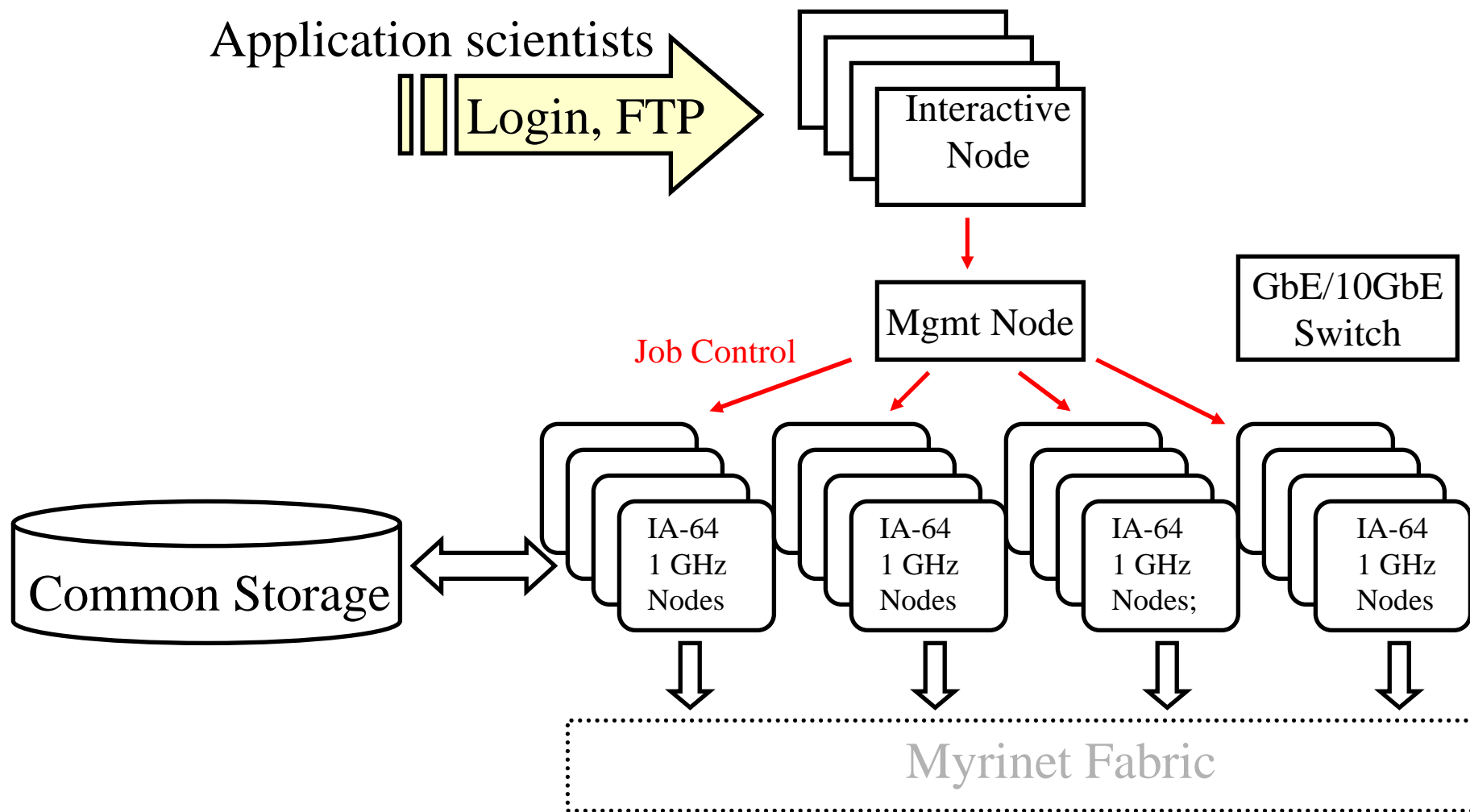
Metro Ring

1x OC48 per site
First light: 8/02



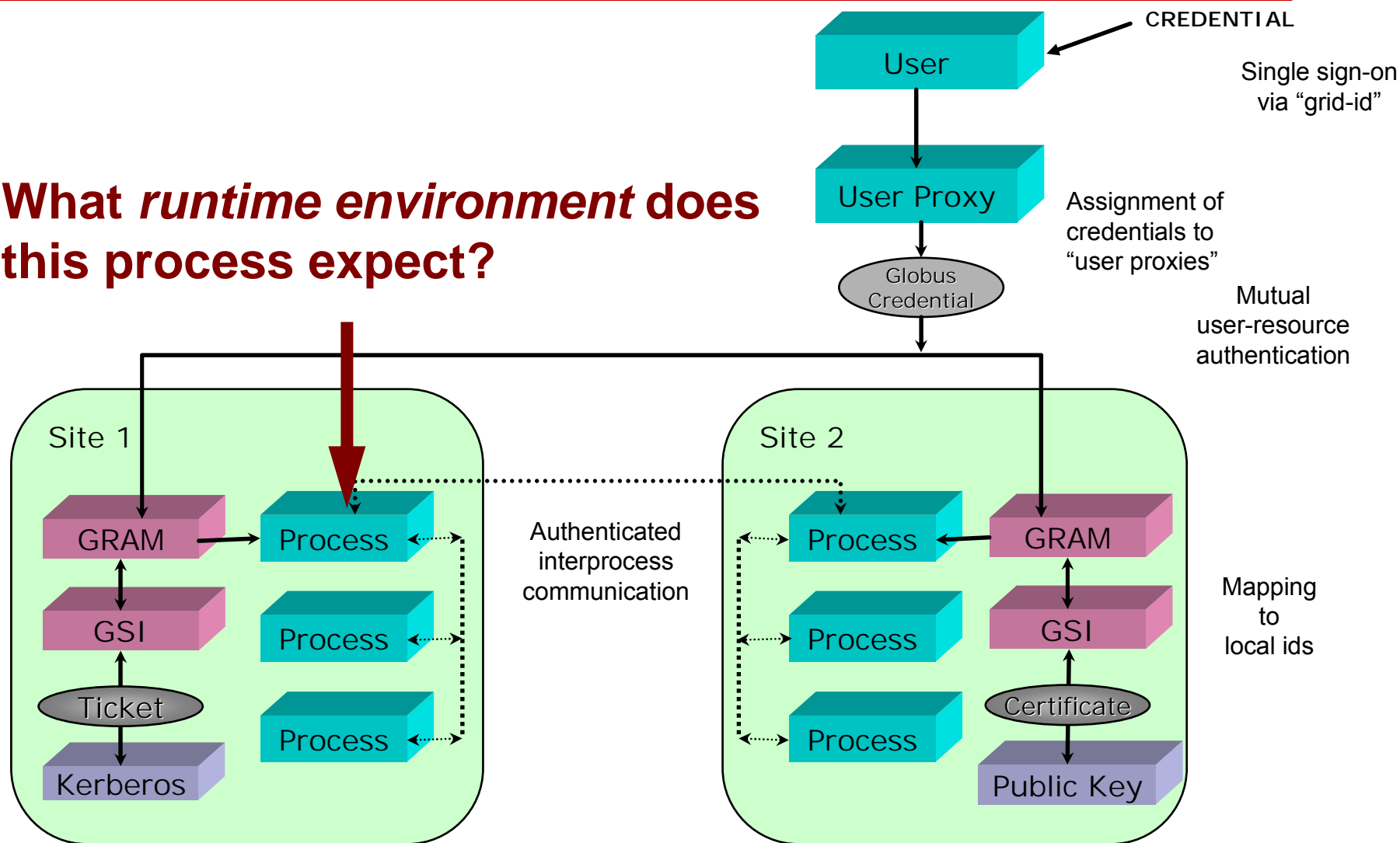
- Each of these three ONI DWDM systems have capacity of up to 66 channels, up to 10 Gb/s per channel
- Protection available in Metro Ring on a per-site basis

Back to those clusters: How to use?

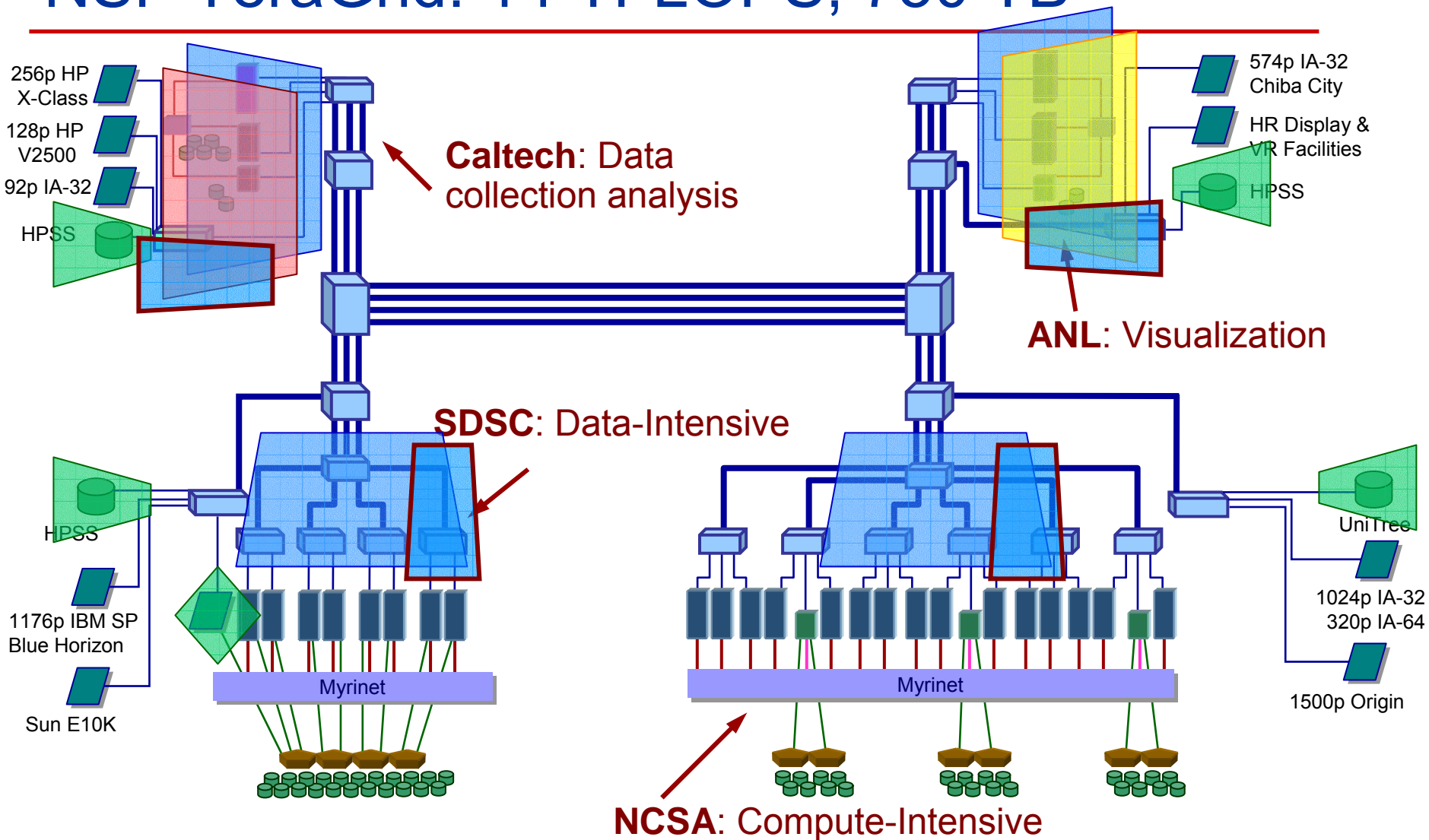


Example: Runtime Environment (*hosting*)

What *runtime environment* does this process expect?



NSF TeraGrid: 14 TFLOPS, 750 TB



Strategy: Define Standard Services

- Finite Number of TeraGrid Services
 - Defined as specifications, protocols, API's
 - Separate from implementation (magic software optional)
- Extending TeraGrid
 - Adoption of TeraGrid specifications, protocols, API's
 - What protocols does it speak, what data formats are expected, what features can I expect (how does it behave)
 - Service Level Agreements (SLA)
 - Extension and expansion via:
 - Additional services not initially defined in TeraGrid
 - e.g. Alpha Cluster Runtime service
 - Additional instantiations of TeraGrid services
 - e.g. IA-64 runtime service implemented on cluster at a new site
- Example: File-based Data Service
 - API/Protocol: Supports *FTP* and *GridFTP*, *GSI* authentication
 - SLA
 - All TeraGrid users have access to N TB storage
 - available 24/7 with $M\%$ availability
 - $\geq R$ Gb/s read, $\geq W$ Gb/s write performance

Defining and Adopting Standard Services

Finite set of TeraGrid services-
applications see *standard services* rather
than *particular implementations...*

Grid Applications

...but sites also provide additional services
that can be discovered and exploited.



IA-64 Linux TeraGrid Cluster Runtime



File-based Data Service



IA-64 Linux Cluster Interactive Development



Interactive Collection-Analysis Service

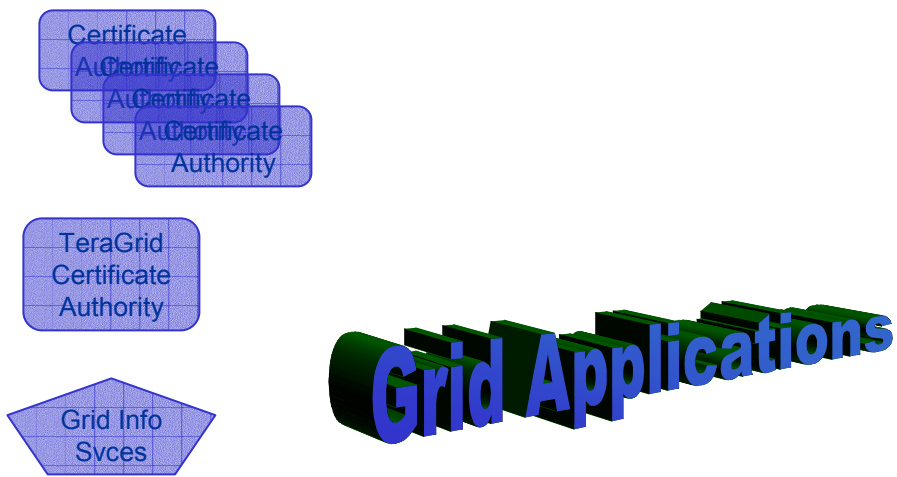


Volume-Render Service



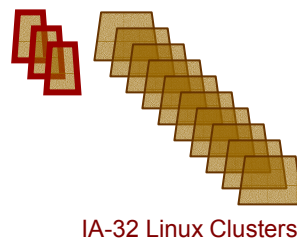
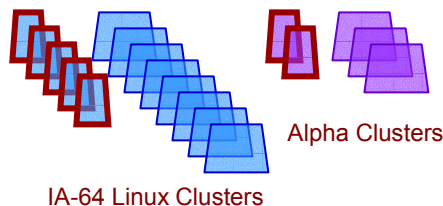
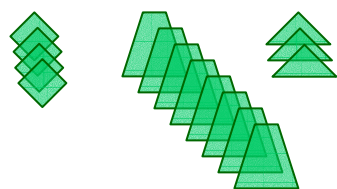
Collection-based Data Service

Standards → Cyberinfrastructure



If done openly and well...

- other IA-64 cluster sites would adopt TeraGrid service specifications, increasing users' leverage in writing to the specification
- others would adopt the framework for developing similar services on different architectures



Data/Information

- File-based Data Service
- Collection-based Data Service
- Relational dBase Data Service

Compute

- Interactive Development
- Runtime

Analysis

- Interactive Collection-Analysis Service
- Visualization Services

General TeraGrid Services

- Authentication
 - If required, must support GSI
 - Requires TeraGrid CA policy and services
- Resource Discovery and Monitoring
 - TeraGrid services/attributes published in Globus directory services
 - Define attributes to be published (some may be TG specific)
 - NMI release (Globus Toolkit 2.0)
 - Standard account information exchange to map use to allocation
- Status Query
 - For many services, publish query interface
 - Scheduler: queue status
 - Compute, Visualization, etc. services: attribute details
 - Network Weather Service
 - Allocations/Accounting Database: for allocation status

General TeraGrid Services

- **Advanced Reservation**
 - On-Demand services
 - Staging data - coordination of storage and compute resources
- **Communication and Data Movement**
 - All services assume any TeraGrid cluster node can talk to any TeraGrid cluster node
 - MPI communication
 - Technically this supports streaming, but requires availability of space at destination (e.g. staging service)
 - Note requires ability to open ports (including listening, access from outside of TG)
 - All resources support GridFTP

Summary and Status

- Finalizing node architecture (e.g. 2p vs 4p nodes) and site configurations (due by end of June)
 - Expecting initial delivery of cluster hardware late 2002
 - IA-32 “TeraGrid Lite” clusters up today, will be expanded in June/July
 - Testbed for service definitions, configuration, coordination, etc.
- NSF has proposed “Extensible TeraScale Facility”
 - Joint between ANL, Caltech, NCSA, PSC, and SDSC
 - PSC Alpha-based TCS1 as test case for heterogeneity, extensibility
 - Transformation of 4-site “internal” mesh network to extensible hub-based hierarchical network
- First Chicago-to-LA lambda on order
 - Expect to order 2nd lambda in July, 3rd and 4th in September
- Initial definitions being developed
 - Batch_runtime
 - Data_intensive_on-demand_runtime
 - Interactive_runtime
 - “Makefile Compatibility” service
- Merging “Grid,” “Cluster,” and “Supercomputer Center” cultures is challenging!