Tier2 Centers

Rob Gardner

University of Chicago

LHC Software and Computing Review

UC San Diego

Feb 7-9, 2006

Outline

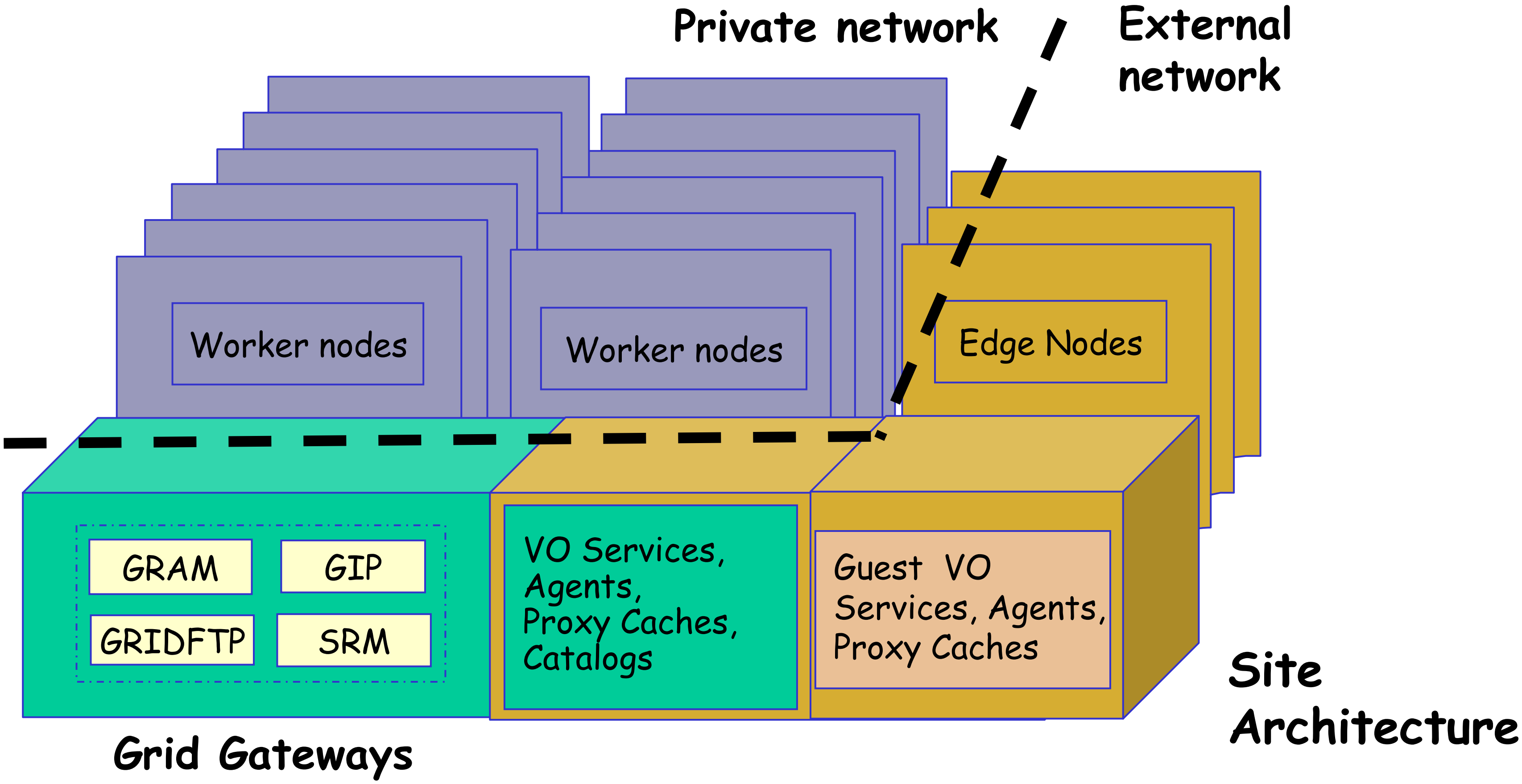
- ATLAS Tier2 centers in the computing model
- Current scale and deployment plans
  - Southwest (Arlington, Oklahoma, New Mexico)
  - Northeast (Boston, Harvard)
  - Midwest (Chicago, Indiana)
- Software, Services, Operations
- Conclusions

Tier2 Centers in the ATLAS Computing Model

- Tier2 role
  - To provide a computing resource for Monte Carlo production and facilitate physics analysis of AOD samples
  - Users: PanDA production team, general ATLAS and OSG users
- Worldwide: approximately 30 T2 Centers
  - Approximate Overall CAPACITY in 2008:
    - 20 MSi2k CPU
    - 9 PB Disk
  - US to satisfy commitments to ATLAS:
    - 3.3 MSi2k CPU
    - 1.5 PB Disk
  - Additional U.S. ATLAS physicists needs
- Current estimate of our average T2 in 2008
  - 1M SI2k CPU, 500 TB disk

Tier2 Site Architecture

- Commodity hardware for CPU and storage
- Provision for hosting persistent “Guest” VO services & agents



Capacity Profile

Tier 2	2005	2006	2007	2008	2009
Northeast					
CPU (kSi2k)	210	350	730	1,090	1,600
Disk (TB)	40	170	370	480	630
Southwest					
CPU (kSi2k)	600	1,000	1,600	1,700	2,100
Disk (TB)	60	200	380	540	700
Midwest					
CPU (kSi2k)	100	240	465	700	1,050
Disk (TB)	50	130	260	465	790

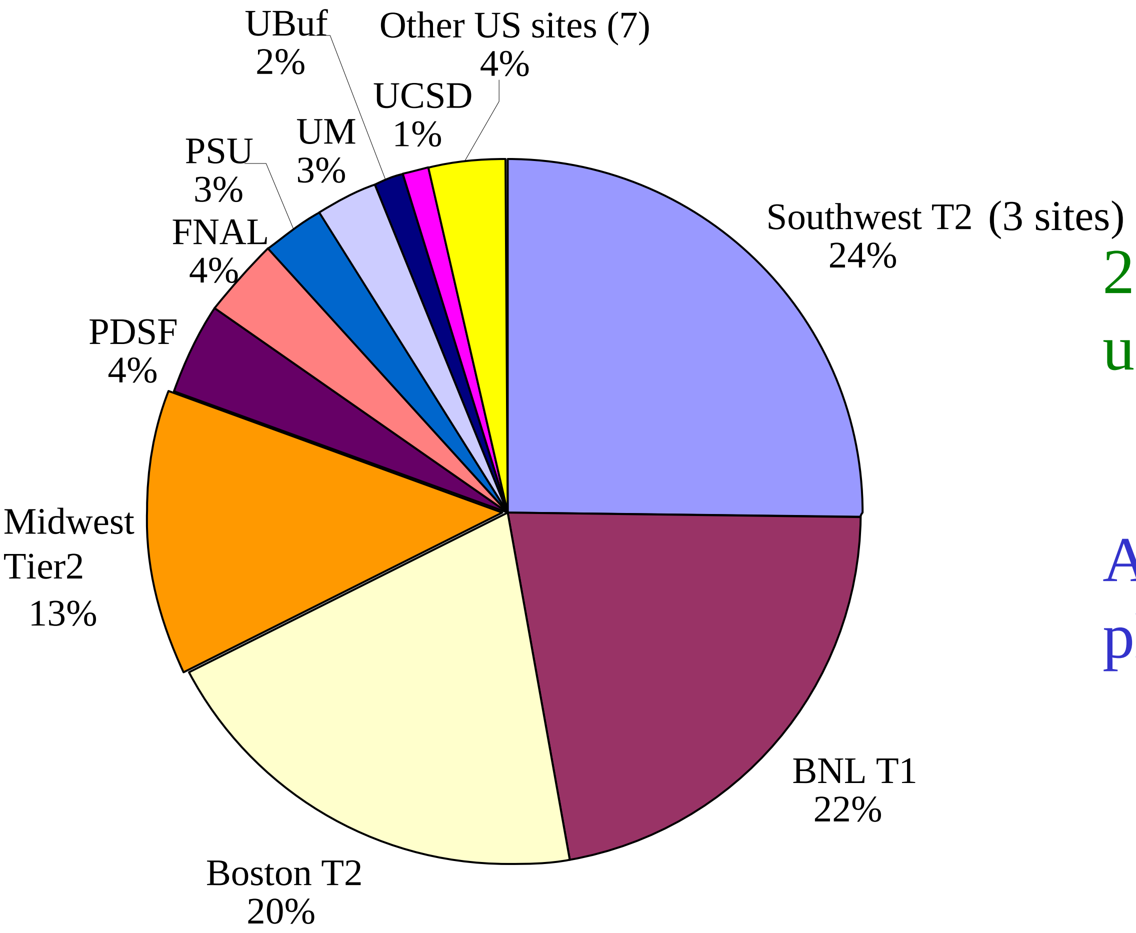
- Assumes Moore’s law doubling of CPU and disk capacity every 3 years at constant cost
- Assumes replacement of hardware every 3 years
- Program-funded resources shown only (total capacity to include leveraged resources)

Tier 2 Funding Profile

	FY03	FY04	FY05	FY06	FY07	FY08	FY09	FY10
Facilities (Tier 2)								
Tier 2 UC/IU								
Personnel (FTEs)	-	-	2.0	2.0	2.0	2.0	2.0	2.0
Personnel (cost)	-	-	300	125	300	300	300	300
Hardware (cost)	-	-	300	125	300	300	300	300
Tier 2C BU/HU								
Personnel (FTEs)	-	-	1.5	2	2	2	2	2
Personnel (cost)	-	-	169	256	300	300	300	300
Hardware (cost)	-	-	169	256	300	300	300	300
Tier 2C SW								
Personnel (FTEs)	-	-	2.0	3.0	3.0	3.0	3.0	3.0
Personnel (cost)	-	-	100	325	300	300	300	300
Hardware (cost)	-	-	238	187	300	300	300	300
Tier 2 Site D								
Personnel (FTEs)	-	-	-	1.0	2.0	2.0	2.0	2.0
Personnel (cost)	-	-	-	150	300	300	300	300
Hardware (cost)	-	-	-	150	300	300	300	300
Tier 2 Site E								
Personnel (FTEs)	-	-	-	-	2.0	2.0	2.0	2.0
Personnel (cost)	-	-	-	-	300	300	300	300
Hardware (cost)	-	-	-	-	300	300	300	300
Tier 2 Central								
Personnel (FTEs)	-	-	-	-	-	-		
Personnel (cost)	-	-	-	-	-	-		
Hardware (cost)	-	-	-	-	-	-		

	FY03	FY04	FY05	FY06	FY07	FY08	FY09	FY10
Total Tier 2								
Personnel (FTEs)	-	-	5.5	8.0	11.0	11.0	11.0	11.0
Personnel (cost)	-	-	569	856	1,500	1,500	1,500	1,500
Hardware (cost)	-	-	707	718	1,500	1,500	1,500	1,500
Total Costs	-	-	1,276	1,574	3,000	3,000	3,000	3,000

U.S. Grid Production (Rome/DC2 combined)



20 different sites used in the U.S.

ATLAS Tier 2's played dominant role

Southwest:UTA

- Purchased 175 node cluster from Dell
  - 160 Worker nodes
    - Dual Xeon 3.2GHz, 4GB RAM, 160GB Storage
  - 8 Front-end Nodes
    - Dual Xeon 3.2GHz, 8GB RAM, 73GB RAID1 Storage
  - 6 I/O servers
  - 1 management node
  - 20TB (raw) disk system
  - Gigabit Ethernet interconnection
- Computing center expected ready now (Feb 3).
- Summary
  - 496 kSI2K CPU
  - 55 TB disk

Southwest:OU

- Purchased 44 Node cluster from Dell
  - 40 Worker nodes
    - Dual Xeon 3.2GHz, 4GB RAM, 160GB Storage
  - 2 Front-end nodes
    - Dual Xeon 3.2GHz, 8GB RAM, 73GB RAID1 Storage
  - 2 I/O Nodes
  - 5TB Disk subsystem
  - Gigabit Ethernet interconnection
- Awaiting delivery of Disk subsystem
- Temporary facility with adequate power/cooling being used
- Permanent Facility available 5/06
- Summary (deployed)
  - 135 kSI2K
  - 11 TB disk

Southwest:Non-dedicated

- UTA-DPCC
  - ❑ 80 Node (dual 2.4GHz Xeon), 45 TB Storage
  - ❑ Expansion expected in 2Q 2006
    - ⌘ ~64 processors
    - ⌘ ~9TB storage
- OU-Boomer
  - ❑ 135 Node (dual 2.0GHz Xeon), 2TB
  - ❑ Will be dedicated to Physics in 2Q (DØ)
- OU-Topdawg
  - ❑ 512 Node (dual 2.2GHz Xeon), 10TB
  - ❑ Expected to come online in 2Q

Southwest:UTA Facility

Southwest Tier2 Staff

- Staffing
  - One person hired
  - Two positions currently advertised
  - May add a fourth person later this year

Northeast Tier 2

Boston University + Harvard University

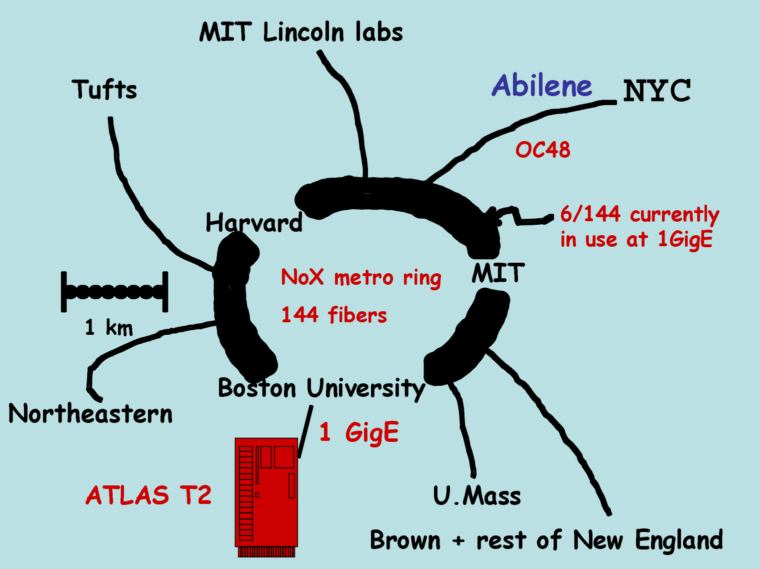
General Contacts: Jim Shank, **Saul Youssef**

shank@bu.edu , youssef@bu.edu

WAN: Charles Von Lichtenberg

chuckles@bu.edu

LAN & Security: Augustine Abaravicyus
systems@scv.bu.edu



Northeast Staff

- ~1.5 FTE sys admin, ~1 of which is contributed
- Management of the facility with fractional effort from Shank & Youssef
- Part of 1 physics grad student (ITR paid) helps with non-cluster admin and Pacman-related work.
- Two hires expected over the next six months, the first of which will be an additional systems administrator.

Northeast Tier2

- Project is building from iVDGL prototype Tier2 at Boston University
- 56 processor Xeon in IBM blade center
- Dual Xeon Gatekeeper node
- Dual Xeon File server node
- 32 processor Xeon development cluster
- 9TB NFS mounted disk
- 4TB Local disk on worker nodes
- WAN 1 GigE to campus core, OC48 to Abilene
- 56 AMD dual core just added
- Summary
  - 181 kSI2K CPU
  - 13 TB disk

Northeast Next Purchases

- 28 “standard” nodes, 10K local disks, 1G Ram per core, 2.2 GHz AMD CPUs.
- 28 “high end” nodes, 15K local disks, 4G RAM/Core, >2.2 GHz AMD CPUs.
- 20 TB of NFS mounted storage
- Summary expected additional capacity

Midwest Tier2:Chicago

- Project is building from iVDGL prototype Tier2 center
- 64 dual 3.0 GHz Xeon nodes, 2 GB memory
- Interconnected by Cisco Catalyst 3750G
- A storage system consists of four servers with dual 3Ware disk controllers, providing 16 TB of attached RAID storage.
- Eight front-end nodes provide grid GRAM and GridFTP servers from VDT (OSG and OSG Integration testbed gateways)
- The facility also has interactive login nodes to support local batch analysis that are provided on a best effort basis
- 4 machine development cluster (Rocks and Condor configuration trials, OSG integration testbed)
- Summary (deployed)
  - 128 kSI2K
  - 16 TB disk

Midwest Tier2:Indiana

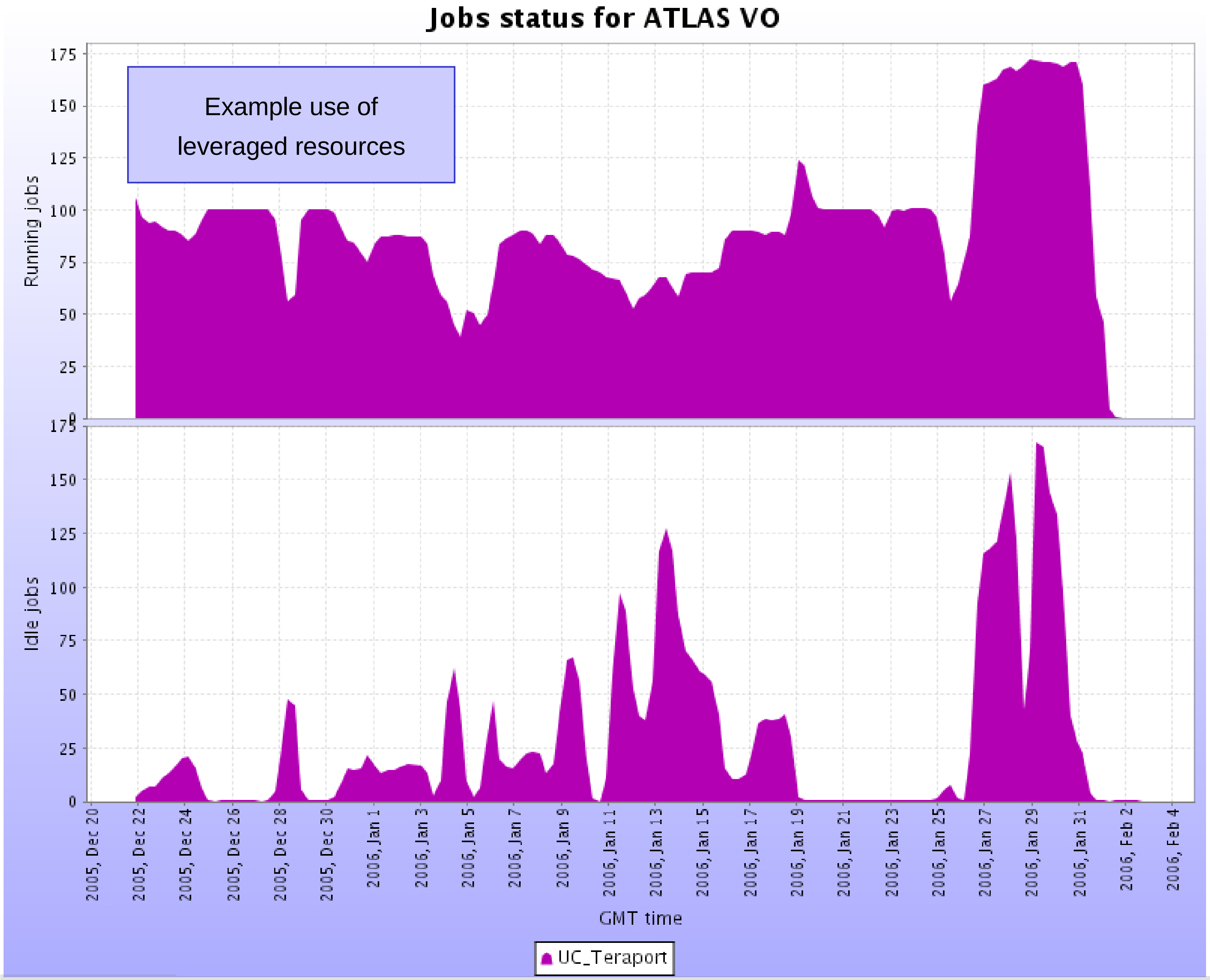
- Project is building from iVDGL prototype Tier2 center
- Dedicated use of 64 2.4 GHz Xeon processors possible through an in-kind contribution by IU (AVIDD, NSF-MRI) to ATLAS
- 1.5 TB of dedicated fiber channel disk
- Additional 8.0 TB of Network Attached Storage (NAS)
- HPSS Archival tape system
- New MWT2 equipment will be located on the Indianapolis campus of IU (IUPUI)
- Summary (deployed)
  - 51 kSI2K CPU
  - 9.5 TB disk

Non-dedicated:MWT2:Chicago

- UC Teraport cluster (NSF-MRI)
  - 128 node IBM e325 Opteron Model 248 (2.2 GHz)
  - 14 TB fiber channel storage
  - GPFS filesystem, SLES9 OS

- Support for opportunistic ATLAS last 3 months

Account	#jobs	#CPU-days	fraction
usatlas3	20307	5622	35.75%
usatlas1	8495	1296	8.24%



Example use of
leveraged resources

Midwest Tier2 Staff and Next Purchases

- Three dedicated project FTEs hired
  - ❑ Marty Dippel @ UC; Kristy Kallback-Rose @ IU
  - ❑ Dan Shroeder @ IU
- Next procurement in progress
  - ❑ Adding ~100 kSI2K and 50 TB of storage
  - ❑ Smaller capacity to achieve 10 Gb/s switching infrastructure at cluster
  - ❑ RFP being finalized now

Midwest Tier2 Facilities

Indiana
Machine Room

Chicago Tier2 Cluster

Conclusions

- 3 Tier2 centers deployed and functional
  - Each site is significantly expanding capacity now and for the remainder of 2006 (~1M SI2K, 100 TB now deployed)
  - ATLAS software and services deployed - functional, heavily used
  - All integrated into OSG and providing services and resources to the common US and WLCG grid infrastructure
  - Network upgrades in progress - each site with aggressive plans to provide 10 Gb/s services in 2006, 2007
  - Operations and policies - good start made, more work to do
  - Major upcoming focus - production dCache-SRM storage services
- Call for 2 additional Tier 2 centers sent out

Additional Slides

Software and Services

- ATLAS releases provided at all Tier2 sites
  - Distribution kit releases for production
- Services deployed
  - DDM managed agents and catalogs
  - Small scale dCache services deployed; gaining experience
- OSG production and OSG Integration testbed (ITB) instances
- Interactive support (some)
  - Login accounts, local job submission
  - OSG Grid client tools

Policy Issues

- Resource allocation determined by US ATLAS policy (RAC)
- Condor and PBS will have different implementations (near term), but will effectively set usage for:
  - ❑ for production managers
  - ❑ for software maintainers
  - ❑ for US ATLAS users
  - ❑ for general ATLAS
  - ❑ for general OSG access
- Set by role based authorization system (VOMS, GUMS, and local Unix accounts)
  - ❑ Have configured UC GUMS to support US ATLAS roles

Operations

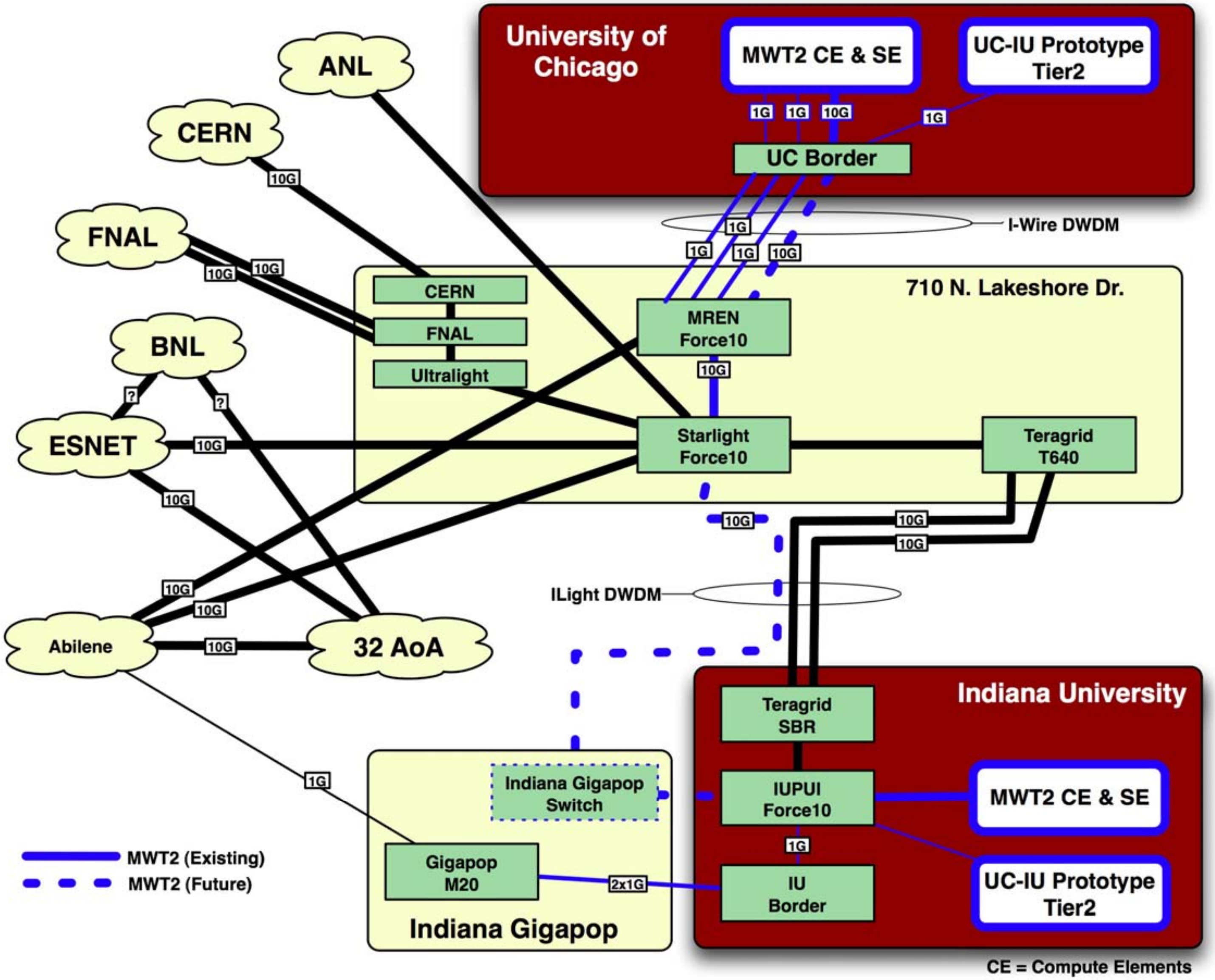
- Systems administrator response: 9-5 M-F and best effort
- General grid problems go to the US ATLAS VO:
  - US ATLAS support center @ BNL trouble ticket system
  - http://www.usatlas.bnl.gov/twiki/bin/view/
  - http://www.usatlas.bnl.gov/twiki/bin/view/Support/WebHome

Network Status and Plans

- Overall MWT2 network architecture
- UC Connectivity status and plans
- IU Connectivity status and plans
- Starlight Configuration Issues
- IU and UC network support organizations, represented here today

Dec 2005

MWT2 Network Architecture



CE = Compute Elements
SE = Storage Elements

UC Connectivity to Starlight - Status

- Prototype Tier2 Cisco 3750 connected via 1 Gbps path to campus border router (Cisco 6509)
  - Once at the border, we share with the rest of UC
- Campus to 710 N. Lakeshore Drive via 2 x 1 Gbps fiber provided by I-WIRE
  - State of Illinois project: http://www.iwire.org/
- At 710 NLSD
  - DWDM output connected to MREN Force10, providing L2 Ethernet connectivity
  - MREN Force10 provides L3 routing
  - 10 Gbps link between MREN and Starlight Force10, shared

UC-Starlight Connectivity Upgrade

- 10 Gbps upgrade funded by UC Teraport project (NSF-MRI)
- Agreements with Starlight 9/8/05
  - One time cost for 10 Gbps interface at Starlight: 10K
  - First year of operation: 10K
- Components ordered from Qwest 9/15/05
  - DWDM 10 Gbps Lambda transceivers for UC edge routers at Starlight and at campus border: 56.1K
- Delivery on these components delayed by Cisco-Qwest
  - Current ETA: ~ month
- Service brought to Research Institutes building, RI-050, where new Tier2 equipment will be housed
- Tier2 to fund RI-050 connectivity w/ Cisco 6509

IU Connectivity to Starlight

- Currently there are 2x10 Gbps lambdas between IU GigaPoP and 710 NLSD
  - On campus, 10 GigE connections from GigaPoP to ICTC, home of IU.MWT2
  - Connected to Starlight Force10
- Current UC-IU connectivity thus goes via Abilene
- IU dedicated (shared with other IU projects) in place between IUPUI and 710 NLSD
  - Expect by Feb 06

Viewed from the machine room

- Both UC and IU are still finalizing cluster & storage design
- At UC, will likely build cluster around Cisco 6509
  - Cisco chassis plus supervisor engine
  - 4 port 10 Gbps interface card
  - 1 x 48 port 1 Gbps interface card
  - Total switching costs: ~60K, negotiating with other projects in the UC Computation Institute for sharing the costs for 10 Gbps card
  - Separate network for storage-compute node (backplane interference?)
- At IU
  - MWT2 infrastructure costs cover basic network connectivity
  - Currently, this goes via Force10, E600
  - Considering upgrade to E1200
  - Depends on final cluster and storage architecture

Starlight Peering

- Goal is to setup VLAN peering to enable 10 Gbps virtual circuits between each of the major nodes on the network:
  - ❑ Between UC and IU hosts
  - ❑ For either UC or IU to BNL hosts
  - ❑ For either UC or IU to CERN hosts
- For UC, setup rules on MREN router
- For IU, setup with dedicated IU router at Starlight
- All this should be straightforward to establish for UC-IU link
- Not as clear the CERN or BNL links (why we’re here!)