MWT2 Project Update

Rob Gardner

US ATLAS Tier2 Meeting @ UTA

December 8, 2006

Hardware Profile

- Phase I (operational)
  - Processors
    - 28 Dual CPU, dual core AMD Opteron 285 (2.6 GHz): 154k SI2K
    - 112 batch slots
  - Storage
    - 80 GB local scratch
    - 5 x 500GB Hardware RAID5 / node (2.5TB/node)
    - 65 TB dCache-based
  - Edge servers for dCache, DQ2, NFS (OSG, /home), mgt services
  - Gigabit switching Cisco 6509/UC, Force10/IU; 10G blades (for four hosts, 2 at each site)
  - Cluster management
    - Cyclades terminal servers for console logging
    - Ethernet accessible power distribution units for power management
- Phase II (ordered)
  - Orders placed for an additional 44 nodes (308k SI2K), compute only
  - Additional scratch disk for all worker nodes (500 GB)
  - Expect delivery mid-January

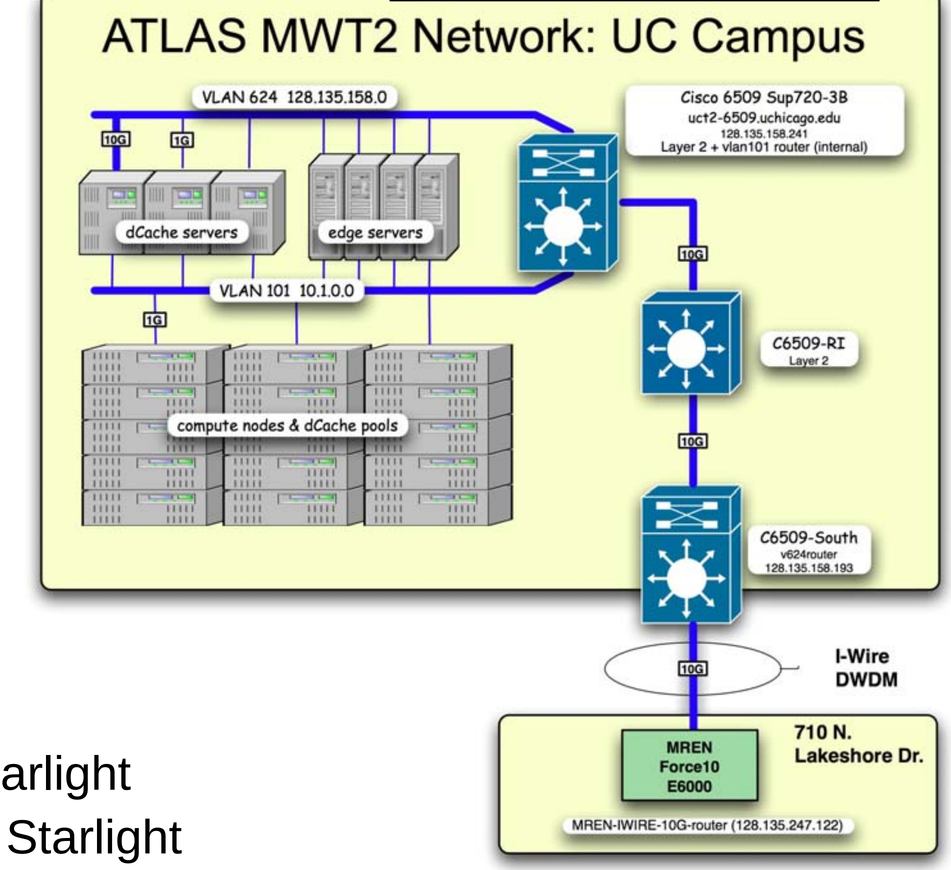
Software Profile

- Platform: SLC4
  - Linux uct2-grid6 2.6.9-42.0.3.EL.cernsmp #1 SMP Fri Oct 6 12:07:54 CEST 2006 i686 athlon i386 GNU/Linux
  - xfs filesystem: benchmarked at 133 MB/s R/W
- OpenPBS
  - Simple: one queue with a 72 hour wall-time limit
- Cluster management tools from ACT
  - Image “cloner” and “beo_exec” command script
- dCache 1.6.6 full bundle (server, client, postgres, dcap)
- OSG 0.4.1
- GUMS
  - Configured to authorize only usatlas1, usatlas2 proxies
- ATLAS
  - Releases: 11.0.3 11.0.42 11.0.5 12.0.3 12.0.31 12.3.0 kitval
  - DQ2 site services installed via dq2.sh

Phase I

- Dual role for worker nodes
  - Four processing cores
  - dCache R/W pool (2.5 TB)
  - 500 GB scratch
- Edge servers
  - 3 dCache services nodes
    - dc1: gridFTP, dcap, SRM
    - dc2: pnfs server, Postgres
    - dc3: admin, gridFTP, dcap
  - DQ2
  - OSG gatekeeper
  - Login
- Network
  - UC: Cisco, w/10G iWIRE to Starlight
  - IU: Force10, w/10G iLIGHT to Starlight
- Other services deployed:
  - OpenPBS, Ganglia, Nagios

IU site same except
all nodes on public network
and Force10 switch



Creation Date: 10/20/06
Contact Information: R. Gardner

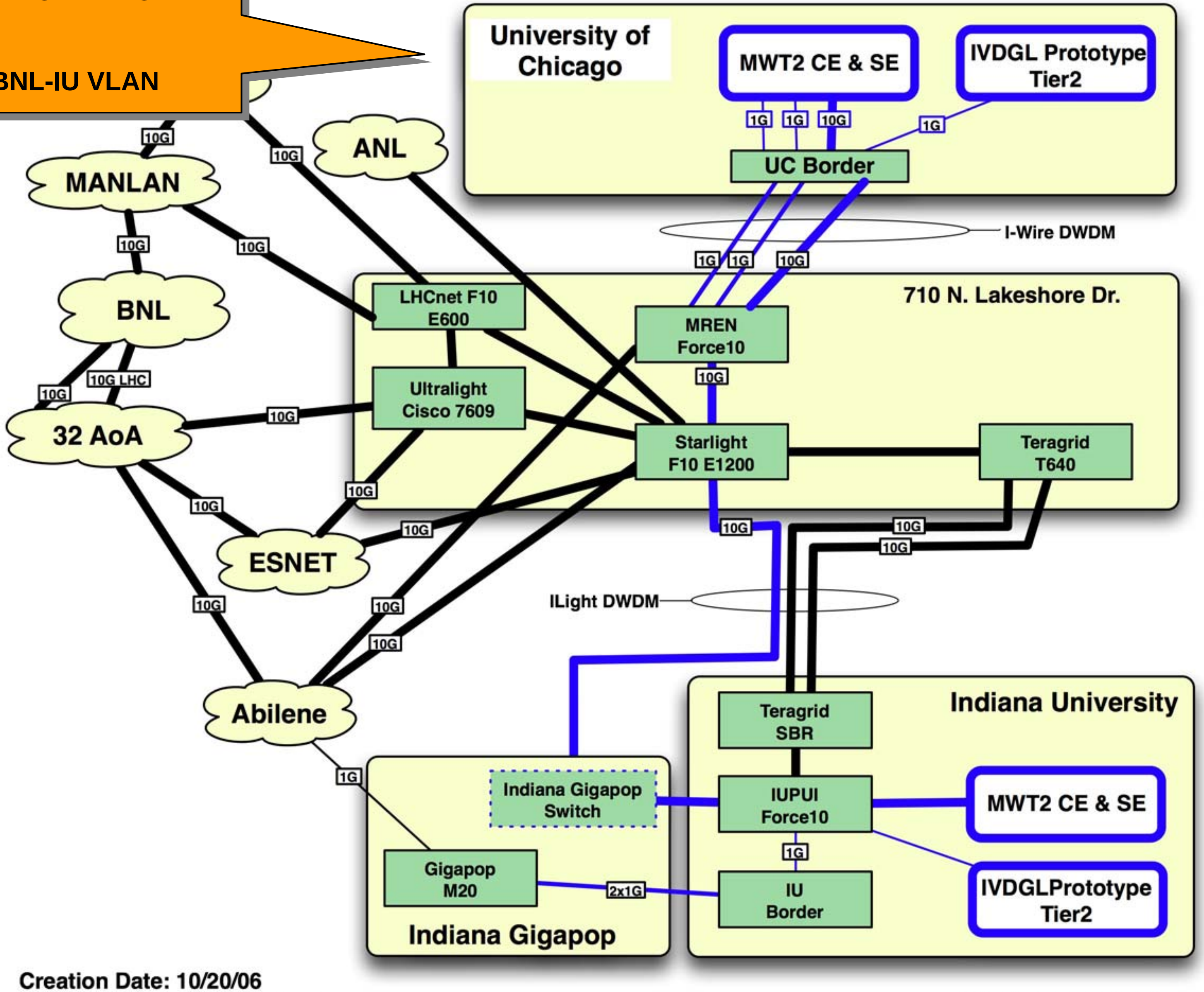
http://plone.mwt2.org/monitors

10G network now operational

IU-UC VLAN via Starlight configured

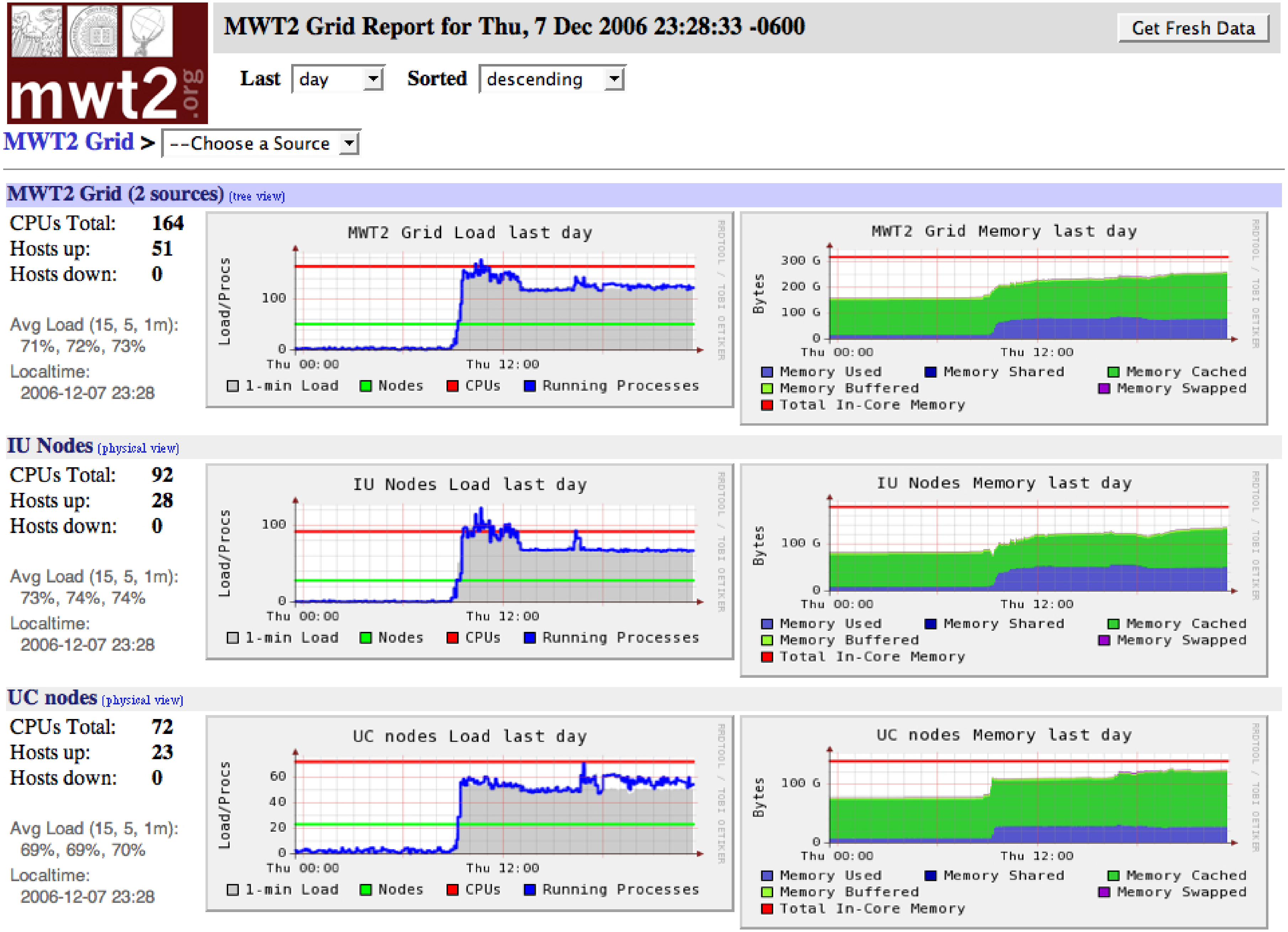
Next:
BNL-UC VLAN & BNL-IU VLAN

MWT2 Network Architecture



Creation Date: 10/20/06
Contact Information: R. Gardner

CE = Compute Element
SE = Storage Element



MWT2 Grid Report for Thu, 7 Dec 2006 23:28:33 -0600

Last day Sorted descending

MWT2 Grid > --Choose a Source

MWT2 Grid (2 sources) (tree view)

CPUs Total:	164
Hosts up:	51
Hosts down:	0

Avg Load (15, 5, 1m):
71%, 72%, 73%

Localtime:
2006-12-07 23:28

IU Nodes (physical view)

CPUs Total:	92
Hosts up:	28
Hosts down:	0

Avg Load (15, 5, 1m):
73%, 74%, 74%

Localtime:
2006-12-07 23:28

UC nodes (physical view)

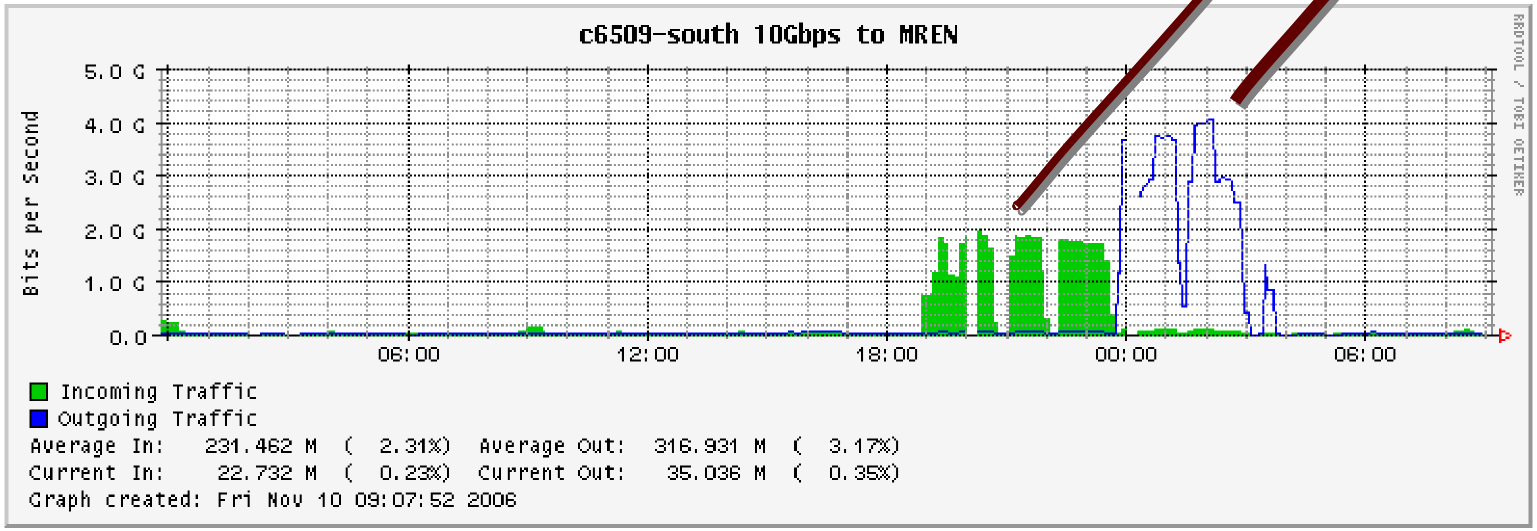
CPUs Total:	72
Hosts up:	23
Hosts down:	0

Avg Load (15, 5, 1m):
69%, 69%, 70%

Localtime:
2006-12-07 23:28

10G Network Tests

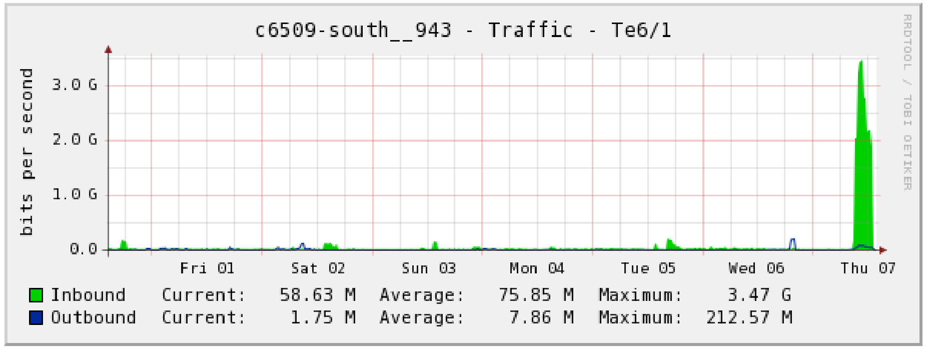
- Tests using griftpPRO using several hosts at each end
- Plots show copy rates ~200 MB/s IU to UC
- Another test UC to IU ~400 MB/s
- One 30 minute interval achieved 539 MB/s

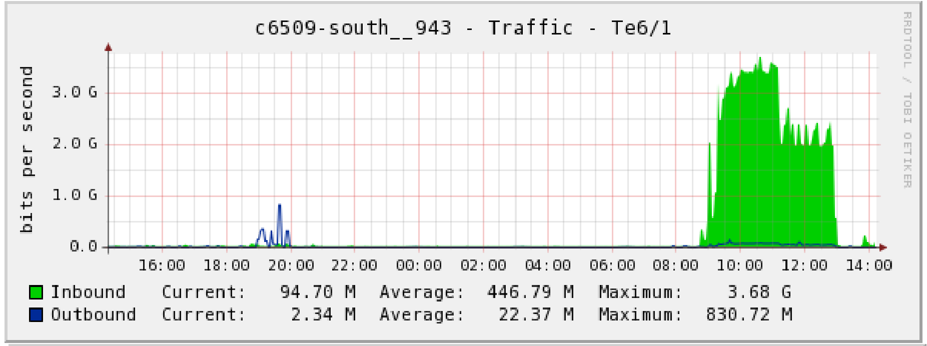


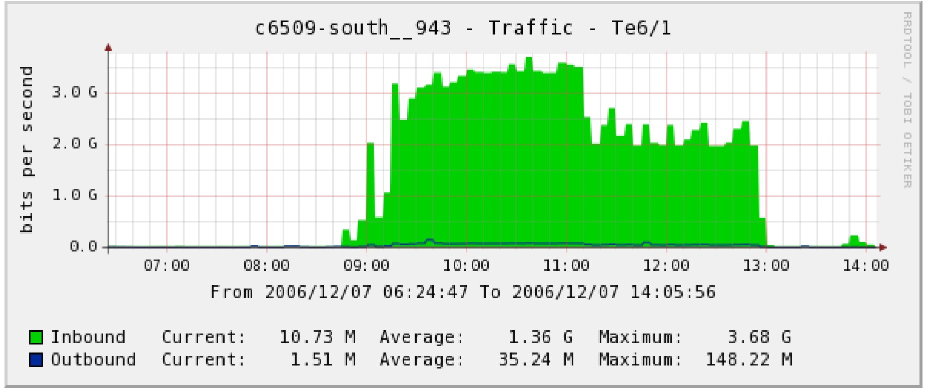
```
🟩 Incoming Traffic
🟦 Outgoing Traffic
Average In:    231.462 M  (  2.31%)  Average Out:  316.931 M  (  3.17%)
Current In:     22.732 M  (  0.23%)  Current Out:   35.036 M  (  0.35%)
Graph created: Fri Nov 10 09:07:52 2006
```

Network Testing II

- 10 simultaneous transfers executed during each iteration (there were 256 iterations in total) was based on a bbftPRO file transfer command.
- Each transfer has used 10 parallel streams to transfer a 1.7 GB file.
- Each file transfer was performed between two different hosts, one at UC, the other at IU.
- Each host had a 1Gbs capable NIC
- Have not adjusted TCP window size, MTU limits, etc.
- Have not used 10G NIC







Problems

- Memory faults
  - 8 x 1 GB DIMMs failing in three nodes “MCE errors”
  - Replaced in two servers at UC; third server returned to ACT
- Kernel panics
  - NFS servers at both UC and IU failed experienced failures
  - Experimenting with NFS parameters: # nfsd’s eg.
- Terminal server memory errors
  - Cyclades buffering doesn’t seem to work
  - Logging host console messages to NFS mounted directory
  - Reboot every three days
- Development pilot submit host
  - (non-BNL) dCache not supported by pilot production hosts
  - Current production done with modified pilot submitter
- DQ2
  - Managing this service continues to be perplexingly complicated

Plan and Capacity Profile

- Phase III (planned Febuary)
  - Fill Phase II nodes with dCache disk pools
  - Based on previous purchases, ~110 TB
- Phase IV (late spring)
  - Based on operational experience with a 175 TB scale dCache system we will evaluate technology options
  - If we continue with the same architecture
    - Increase CPU and storage capacity with a ~$135K purchase
    - Roughly 140k Si2K, 50TB
- Summary comparison (program-funded only)

Tier2 Facility	2005	2006	2007	2008	2009
CPU (Proposal 04) (SI2K)	97670	244439	465102	699327	1050185
CPU (Deployed 06-07)	0	473000	613000		
Disk (Proposal 04) (TB)	51	132	261	465	790
Disk (Deployed 06-07)	0	65	225		

MWT2 team

Kristy Kallback-Rose
Dan Schrager
Greg Cross
Joe Urbanski (1/07)

+ fred, rob