



Site Throughput Review and Issues

Shawn McKee/University of Michigan

US ATLAS Tier2/Tier3 Workshop

May 27th, 2008

US ATLAS Throughput Working Group



- ❄ In Fall 2007 a “Throughput” working group was put together to study our current Tier-1 and Tier-2 configurations and tunings and see where we could improve.
- ❄ First steps were identifying current levels of performance.
- ❄ Typically sites performed significantly below expectations.
- ❄ During weekly calls we would schedule Tier-1 to Tier-2 testing for the following week. Gather a set of experts on a chat line or phone line and debug the setup.
- ❄ [See http://www.usatlas.bnl.gov/twiki/bin/view/Admins/LoadTests.html](http://www.usatlas.bnl.gov/twiki/bin/view/Admins/LoadTests.html)

Roadmap for Data Transfer Tuning

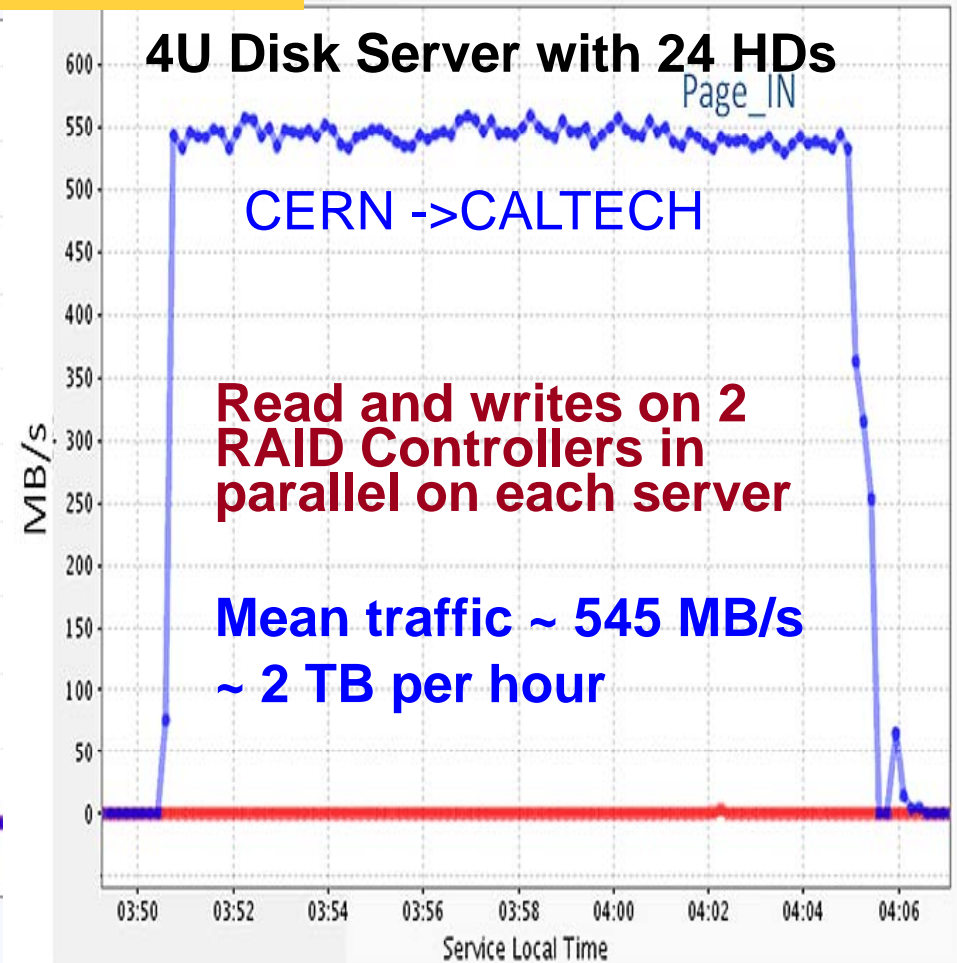
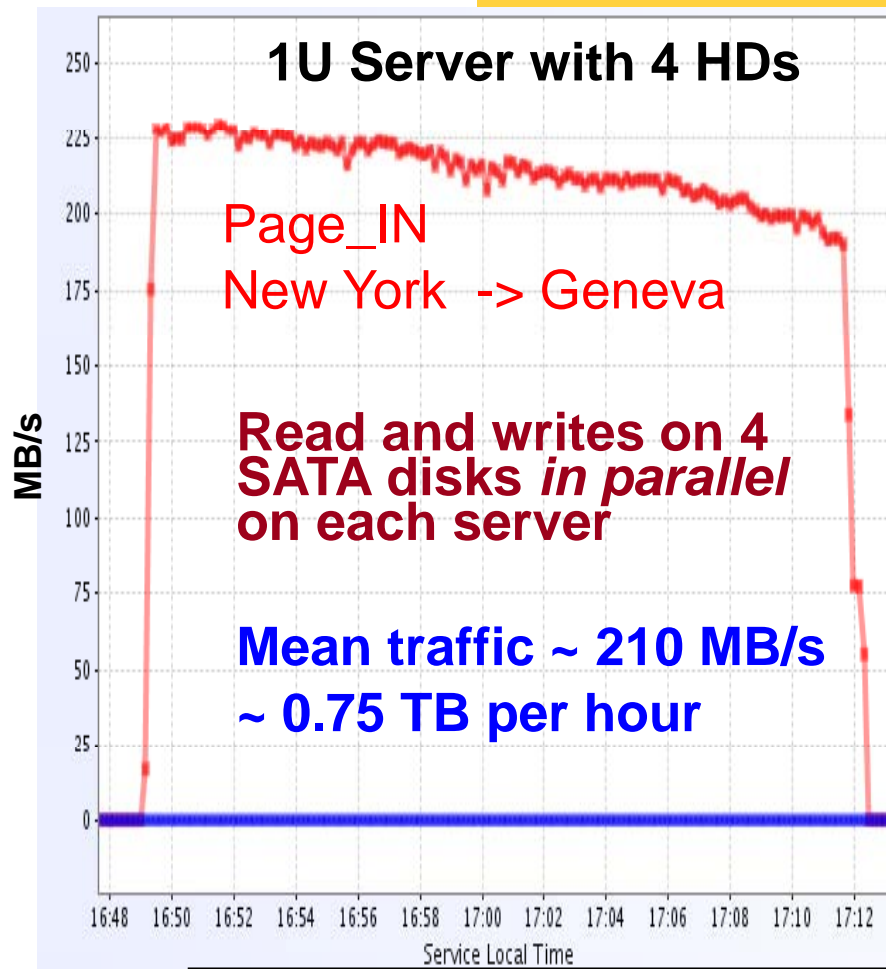


- ❄ Our first goal is to achieve **~200 MB (bytes) per second** from the Tier-1 to each Tier-2. This is achieved by either/both:
 - **Small number of high-performance systems (1-5)** at each site transferring data via GridFTP, FDT or similar applications.
 - ⌘ Single systems benchmarked with I/O > 200 MB/sec
 - ⌘ 10 GE NICs (“back-to-back” tests achieved 9981 Mbps)
 - **Large number of “average” systems (10-50)**, each transferring to a corresponding systems at the other site. This is a good match to an SRM/dCache transfer of many files between sites
 - ⌘ Typical current disks are 40-70 MB/sec
 - ⌘ Gigabit network is 125 MB/sec
 - ⌘ Doesn't require great individual performance / host: 5-25 MB/sec

FDT Examples: High Performance Possible



<http://monalisa.cern.ch/FDT/>



Working on integrating FDT with dCache

Data Transfers for US ATLAS (Oct07)



FTS Report



Disclaimer

This page contains a report generated from information stored in the FTS Database and is intended for reporting purposes only. Since the format will probably change in the future, it's therefore recommended not to use parsing robots on it.

Typical rate is
0.5 – 3 MB/s!!!

Statistics concerning all the transfers performed yesterday managed by "**lcg03.usatlas.bnl.gov**"
Between **2007-10-26 08:00:00 -04:00** and **2007-10-27 08:00:00 -04:00**

Channel Name	VO Name	Total	% Failures	# Succ.	# Fail.	1st Failure Reason	% 1st Failure Reason	2nd Failure Reason	% 2nd Failure Reason	Avg. Size (GiB)	Avg. Duration (sec)	Avg. Tx Rate (MB/sec)	Eff. Tx Bytes (GiB)	Tx Bytes (GiB)
BNLDCACHE-WIS	[All]	5479	97	137	5342	Transfer	91	Source SRM: Prep	8	0.1	143.64	0.95	13.36	13.5
STAR-BNLDCACHE	[All]	1246	96	47	1199	Source SRM: Prep	100	Transfer	0	0.04	62.87	0.68	1.9	1.9
BNLDCACHE-UMICHGFTP	[All]	157	89	17	140	Transfer	99	Source SRM	1	1.89	560.29	3.34	32.09	248.95
BNLDCACHE-HEAD01AGLT2	[All]	3216	83	540	2676	Dest SRM: Prep	92	Transfer	8	0.07	143.09	0.74	38.49	38.72
UTASW2-BNLDCACHE	[All]	1163	33	785	378	Transfer	85	Source SRM: Prep	14	0.09	274.17	0.27	73.5	176.81
UTADPCC2-BNLDCACHE	[All]	661	24	503	158	Transfer	82	Source SRM: Prep	18	0.1	316.03	0.22	50.26	104.54
BNLDCACHE-OUGFTP2	[All]	5	20	4	1	Transfer	100			0.11	126.25	1.31	0.44	0.44
RALLCG2-BNLDCACHE	[All]	10	20	8	2	Source SRM: Prep	100			0.03	117.13	0.29	0.27	0.27
UCT2DC1-BNLDCACHE	[All]	4467	10	4000	467	Dest SRM: Prep	43	Transfer	29	0.11	30.45	3.83	422.33	424.49
IUT2DC-BNLDCACHE	[All]	3287	9	2995	292	Source SRM: Prep	50	Transfer	33	0.12	55.09	2.02	346.57	353.24
NDGFT1DISK2-BNLDCACHE	[All]	256	5	244	12	Source SRM: Prep	50	Source SRM	25	0.03	48.68	0.67	6.45	6.45
PICSRM-BNLDCACHE	[All]	42	5	40	2	Source SRM: Prep	100			0.08	198.85	0.34	3.06	3.06
SERV04SLAC-BNLDCACHE	[All]	5154	5	4915	239	Transfer	62	Source SRM: Prep	38	0.08	35.04	2.04	374.76	375.76
BUATLASTIER2-BNLDCACHE	[All]	3400	4	3256	144	Transfer	59	Source SRM: Prep	39	0.08	45.33	2.18	272.29	272.91
UMICHGFTP-BNLDCACHE	[All]	7053	4	6737	316	Transfer	62	Source SRM: Prep	37	0.07	37.45	1.87	438.44	440.33
GRIDKASRM-BNLDCACHE	[All]	51	2	50	1	Source SRM	100			0.07	203.42	0.37	3.62	3.62
LYONDISK-BNLDCACHE	[All]	243	1	241	2	Source SRM: Prep	100			0.06	49.75	1.27	14.39	14.39
TRIUMFSRM-BNLDCACHE	[All]	50	0	50	0					0.07	54.28	1.24	3.66	3.66

Click on the Channel Name to show the VO details

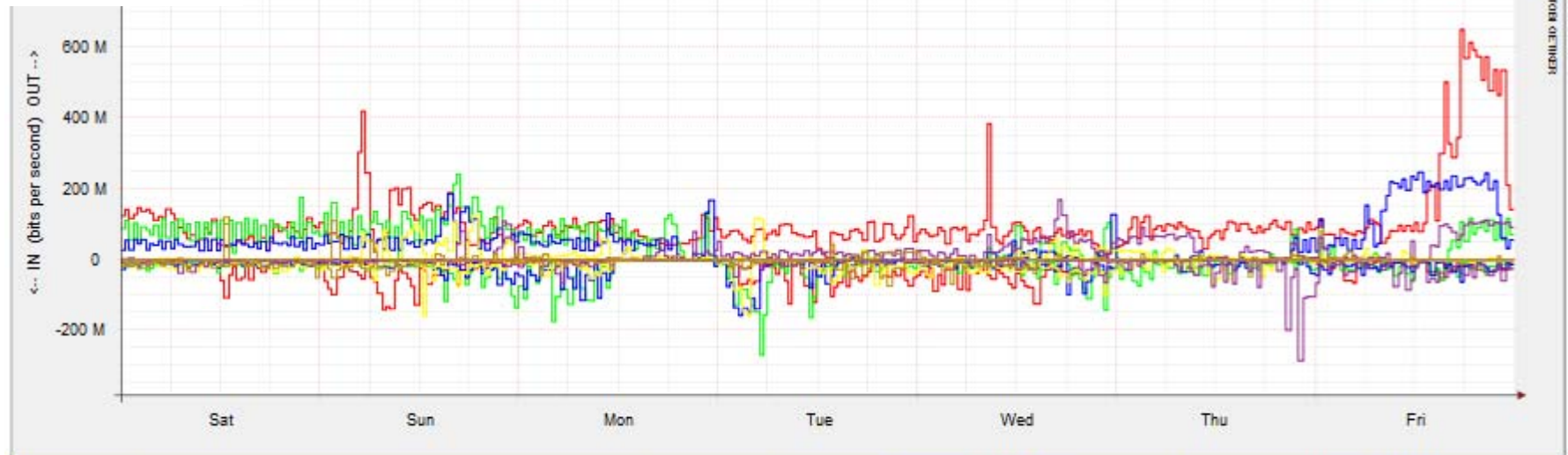
Why aren't we achieving ~30 MB/s ?

Beginning “Load Tests” Untuned



Time Series Bits Time Range: from 2007/12/01 00:00:00 to 2007/12/08 00:00:00 Null Values: Stack: width: 887 height: 354

800 Mbps ->



Color	Site(s)	Label	Average Rate (bps)		Volume (Bytes)	
			In	Out	In	Out
Blue	Indiana University	src	19.518 M	46.388 M	1.346 T	3.199 T
Brown	Oklahoma University	src	8.006 M	3.132 M	565.354 G	221.161 G
Purple	SLAC	src	14.569 M	21.909 M	1.005 T	1.511 T
Green	University Chicago	src	26.967 M	44.179 M	1.860 T	3.047 T
Red	University Michigan	src	22.643 M	106.770 M	1.562 T	7.363 T
Yellow	University Texas at Arlington	src	17.664 M	11.114 M	1.218 T	784.830 G

What are the Primary Issues?



❄ Network

- ❑ First documented what existing network capabilities were for each site
 - ⌘ Typically POOR by default.
- ❑ After tuning could achieve “wire-speed” in memory to memory tests

❄ Storage

- ❑ Highly variable. Individual disks can vary from **5 – 90 MB/sec** depending upon type of disk interface, RPMs, cache, etc.
- ❑ RAID controllers can utilize disks in parallel, in principle leading to linear increases in I/O performance as a function of # of spindles(disks)...
- ❑ Drivers/OS/Kernel can impact performance, as can various tunable parameters

❄ End-to-end Paths

- ❑ What bottlenecks exist between sites? Is competing traffic a problem?

❄ Overall operations

- ❑ Many “small” files limit achievable throughput
- ❑ How well does the system function when data has to move from disk-driver-memory-driver-network-wan-network-driver-memory-driver-disk?!

Tuning/Debugging Methodology



- ❄ First document site topology: Create network diagrams representative of the path to “doors” and “storage”
(See <http://www.usatlas.bnl.gov/twiki/bin/view/Admins/NetworkDiagrams>)
- ❄ Gather information on the servers involved in data transport at the site.
 - ❑ TCP stack settings
 - ❑ NIC model and firmware
 - ❑ OS, processors, memory
 - ❑ Storage details (including benchmarking)
- ❄ Schedule a working call involving the Tier-1 and Tier-2 under test. Perform a series of tests and tunings.

Network Tools and Tunings



- ❄ The network stack is the 1st candidate for optimization
 - ❑ Amount of memory allocated for data “in-flight” determines maximum achievable bandwidth for a given src-destination
 - ❑ Parameters (some example settings):
 - ⌘ `net.core.rmem_max = 20000000`
 - ⌘ `net.core.wmem_max = 20000000`
 - ⌘ `net.ipv4.tcp_rmem = 4096 87380 20000000`
 - ⌘ `net.ipv4.tcp_wmem = 4096 87380 20000000`
- ❄ Other useful tools: *lperf, NDT, wireshark, tracepath, ethtool, ifconfig, sysctl, netperf, FDT.*
- ❄ Lots more info/results in this area available online...
 - ❑ <http://www.usatlas.bnl.gov/twiki/bin/view/Admins/NetworkPerformanceP2.html>


Achieving Good Networking Results

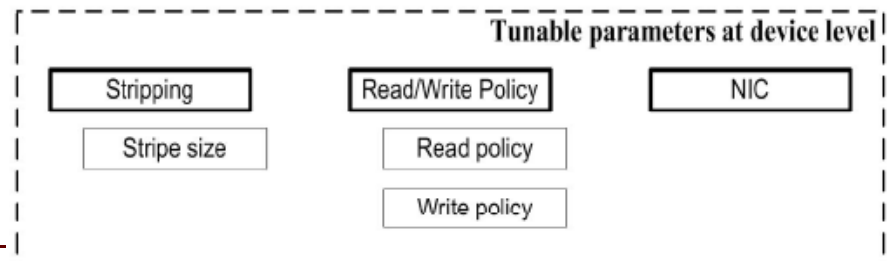
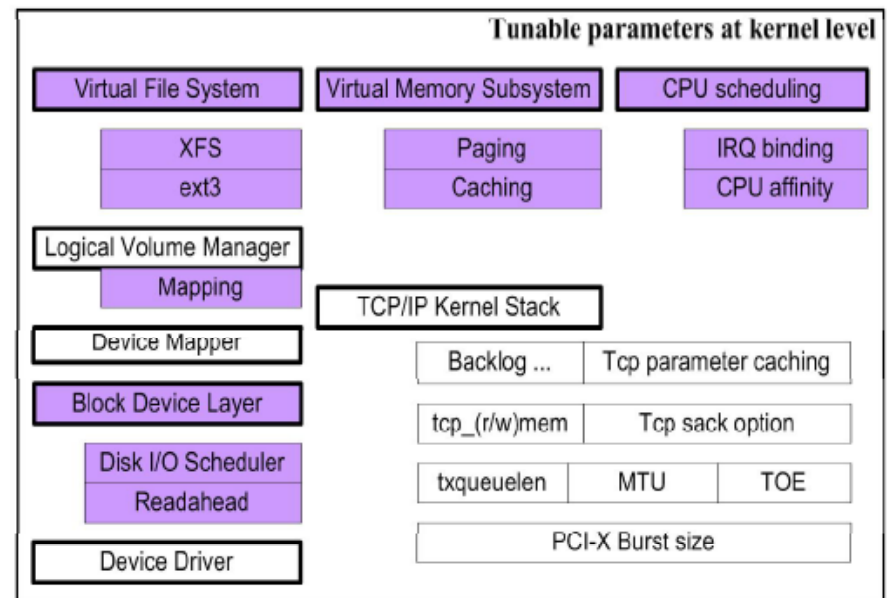
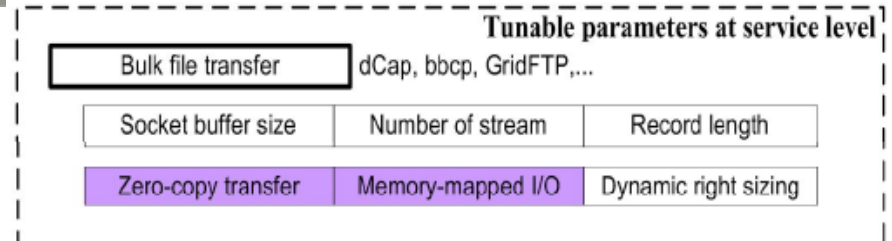


- * Test system-pairs with `lperf (tcp)` to determine achievable bandwidth
- * Check `'ifconfig <ethx>'` to see if errors or packet loss exists
- * Examine driver info with `'ethtool -i <ethx>'`
- * Set TCP stack parameters to allow full use of bottleneck bandwidth, typically 1 gigabit. The maximum expected round-trip-time (RTT) should be used to estimate the amount of memory for data “in flight” and this should be setup in the TCP stack parameters. NOTE: set the maximums large enough...leave the default and “pressure” values low.
- * Retest with `lperf` to determine effect of change.
- * Debug with `lperf (udp)`, `ethtool`, `NDT`, `wireshark` if there are problems
- * Remember to check **both** directions...

Tunable Parameters Impacting I/O



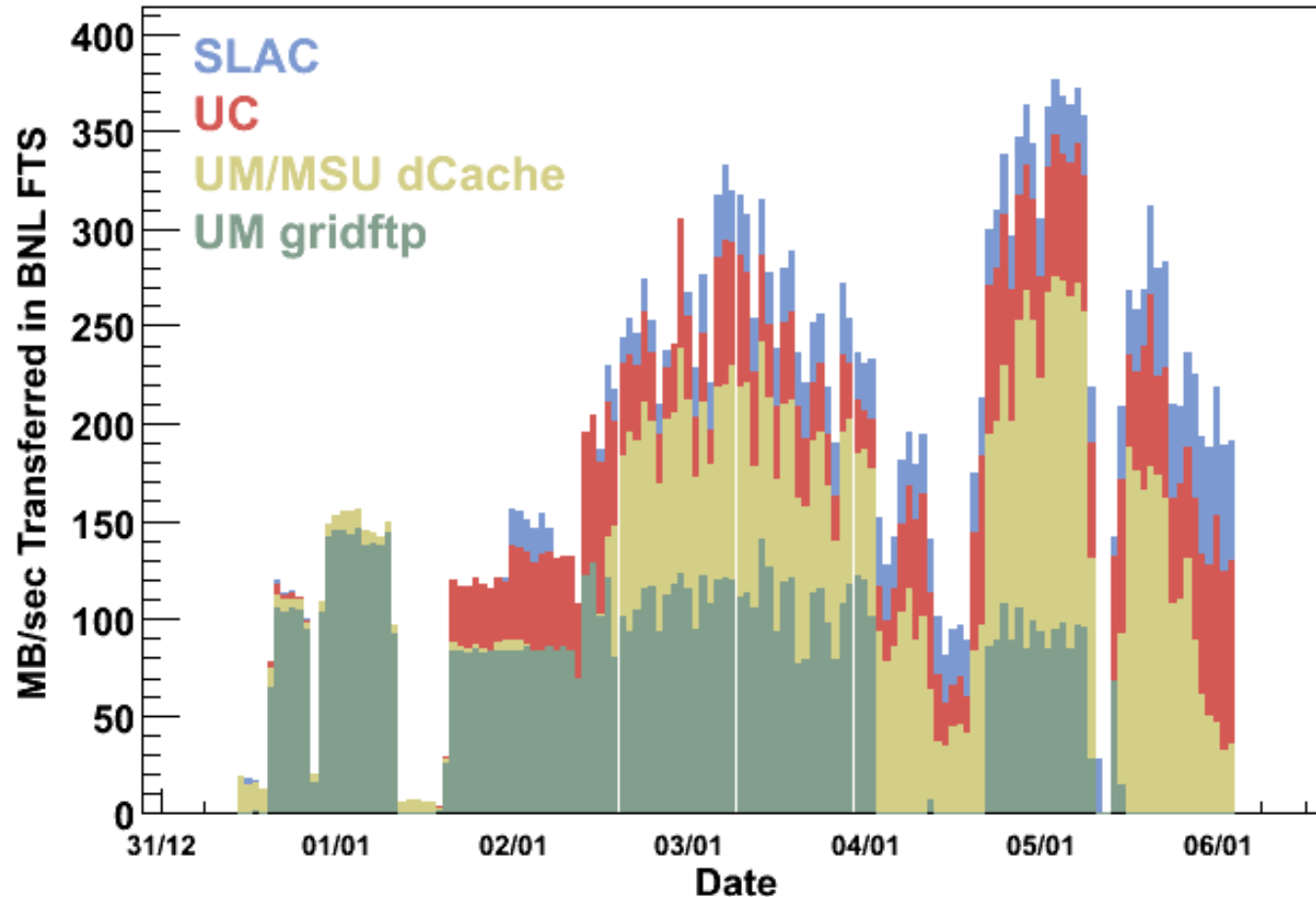
- ❄ There are potentially MANY places in a linux OS that can have an impact on I/O for WAN transfers...
- ❄ We need to explore the impact of various tunings and options . 
- ❄ The purple areas in the figure have been at least initially explored by **Kyu Park (UFL)**.
- ❄ **Wenjing Wu (UM)** is continuing work in this area.



Initial Tuned Tests (Dec 31 – Jan 6)



BNL ---> T2



We have setup manual (contact Hiro) “on-demand” load tests for this...

Initial Findings(1)



- ❄ Most sites had at least a gigabit capable path in principle.
 - ❑ Some sites had < 1Gbit/sec bottlenecks that weren't discovered until we starting trying to document and test sites
- ❄ Many sites had 10GE “in principle” but getting end-to-end connections clear at 10GE required changes
- ❄ Most sites were un-tuned or not properly tuned to achieve high throughput
- ❄ Sometimes flaky hardware was the issue:
 - ❑ Bad NICs, bad cables, underpowered CPU, insufficient memory, etc
- ❄ BNL was limited by the GridFTP doors to around 700 MB/s

Initial Findings(2)



- ❄ Network for **properly tuned hosts** is **not** the bottleneck
- ❄ Memory-to-disk tests interesting in that they can expose problematic I/O systems (or give confidence in them)
- ❄ Disk-to-disk tests do poorly. Still a **lot** of work required in this area. Possible issues:
 - ❑ Wrongly tuned parameters for this task (driver, kernel, OS)
 - ❑ Competing I/O interfering
 - ❑ Conflicts/inefficiency in the Linux “data path” (bus-driver-memory)
 - ❑ Badly organized hardware, e.g., network and storage cards sharing the same bus
 - ❑ Underpowered hardware or bad applications for driving gigabit links

T1/T2 Host Throughput Monitoring



❄ How it works

- ❑ Control plugin for Monalisa runs iperf and gridftp tests twice a day from select Tier 1 (BNL) to Tier 2 hosts and from each Tier 2 host to Tier 1, (production SE hosts).
- ❑ Results are logged to file.
- ❑ Monitoring plugin for Monalisa reads log and graphs results.

❄ What it currently provides

- ❑ Network throughput to determine if network tcp parameters need to be tuned.
- ❑ Gridftp throughput to determine degradation due to gridftp software and disk.
- ❑ Easy configuration to add or remove tests.

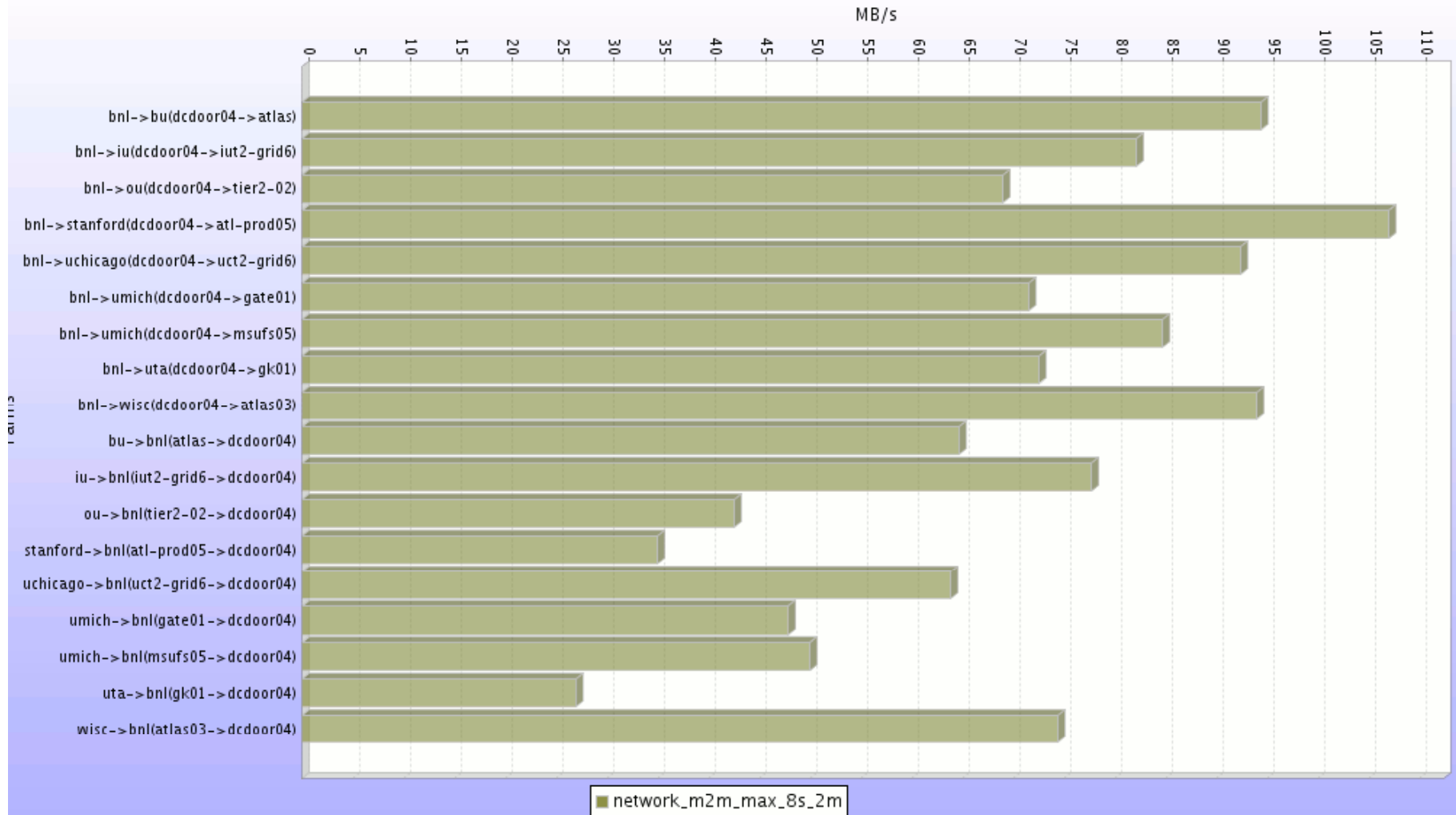
❄ What else can be added within this framework

- ❑ Gridftp **memory to memory** and **memory to disk (and vice-versa)** tests to isolate disk and software degradation separately.
- ❑ SRM tests to isolate SRM degradation
- ❑ FTS to isolate FTS degradation

Iperf (Mem-to-Mem) tests T1<->T2



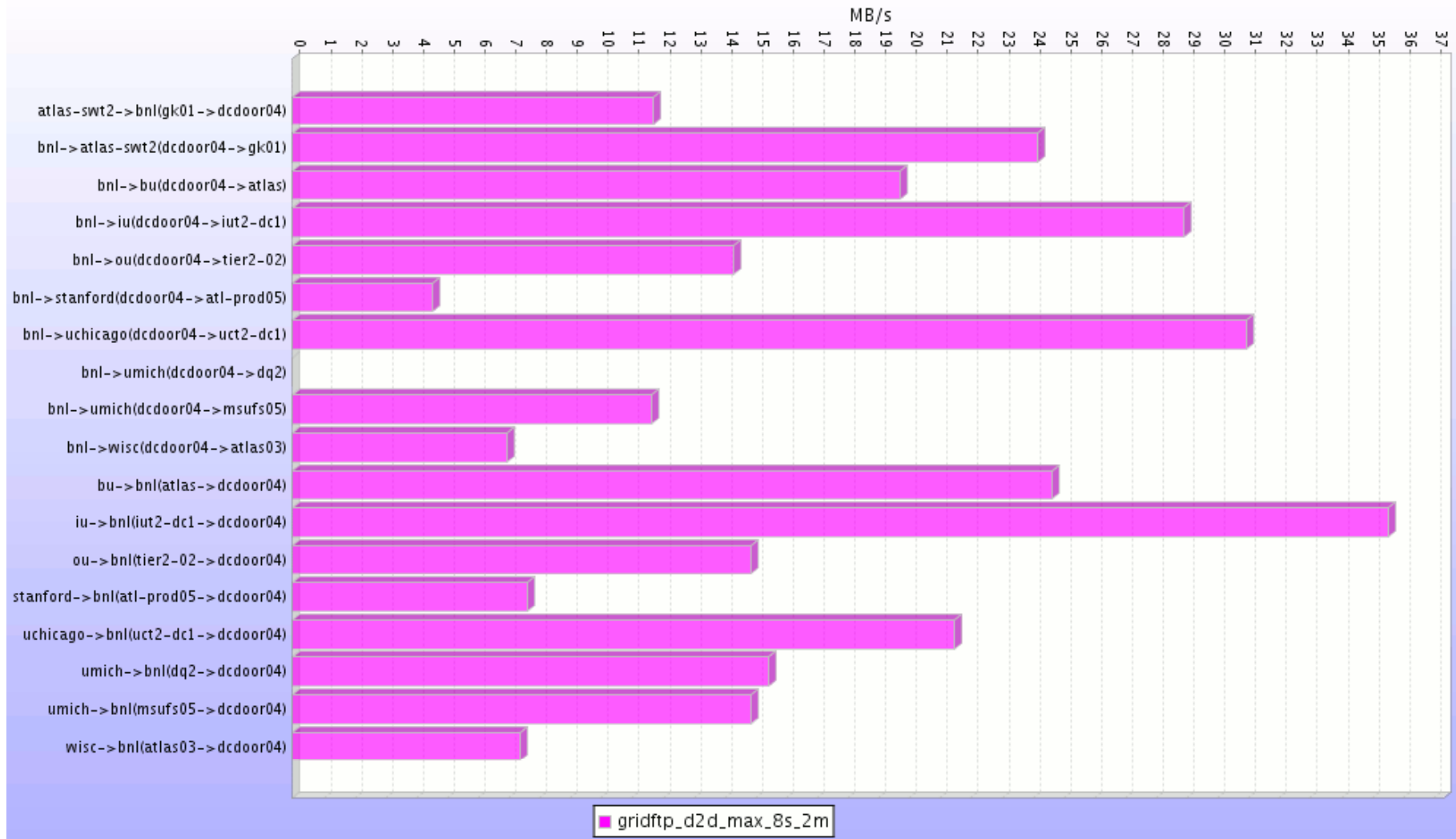
Network (8 streams, 2MB window)



GridFTP (Disk-to-Disk) T1<->T2



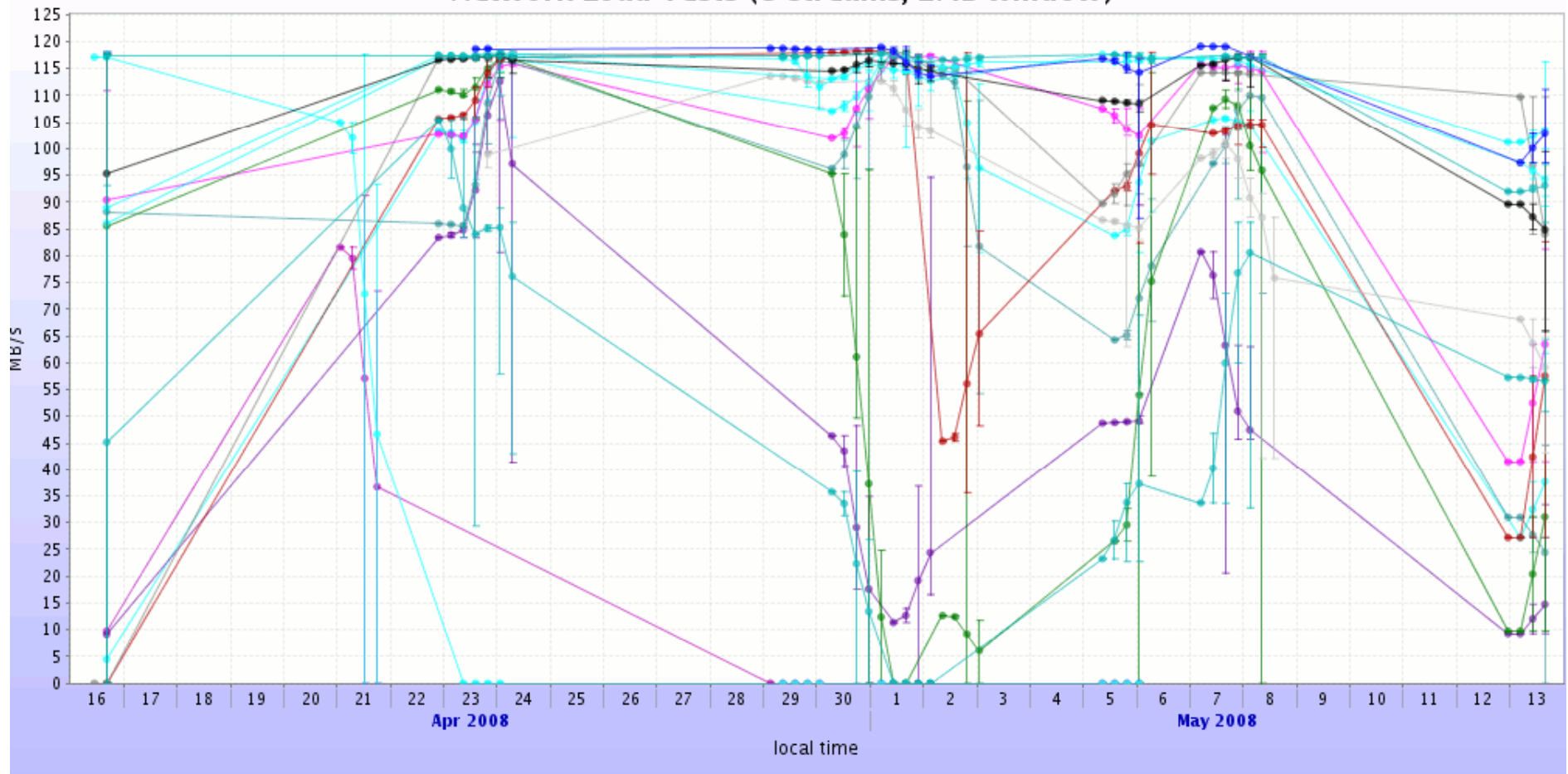
Gridftp_disk2disk (8 streams, 2MB window)



Iperf History Graphing



Network Load Tests (8 streams, 2MB window)

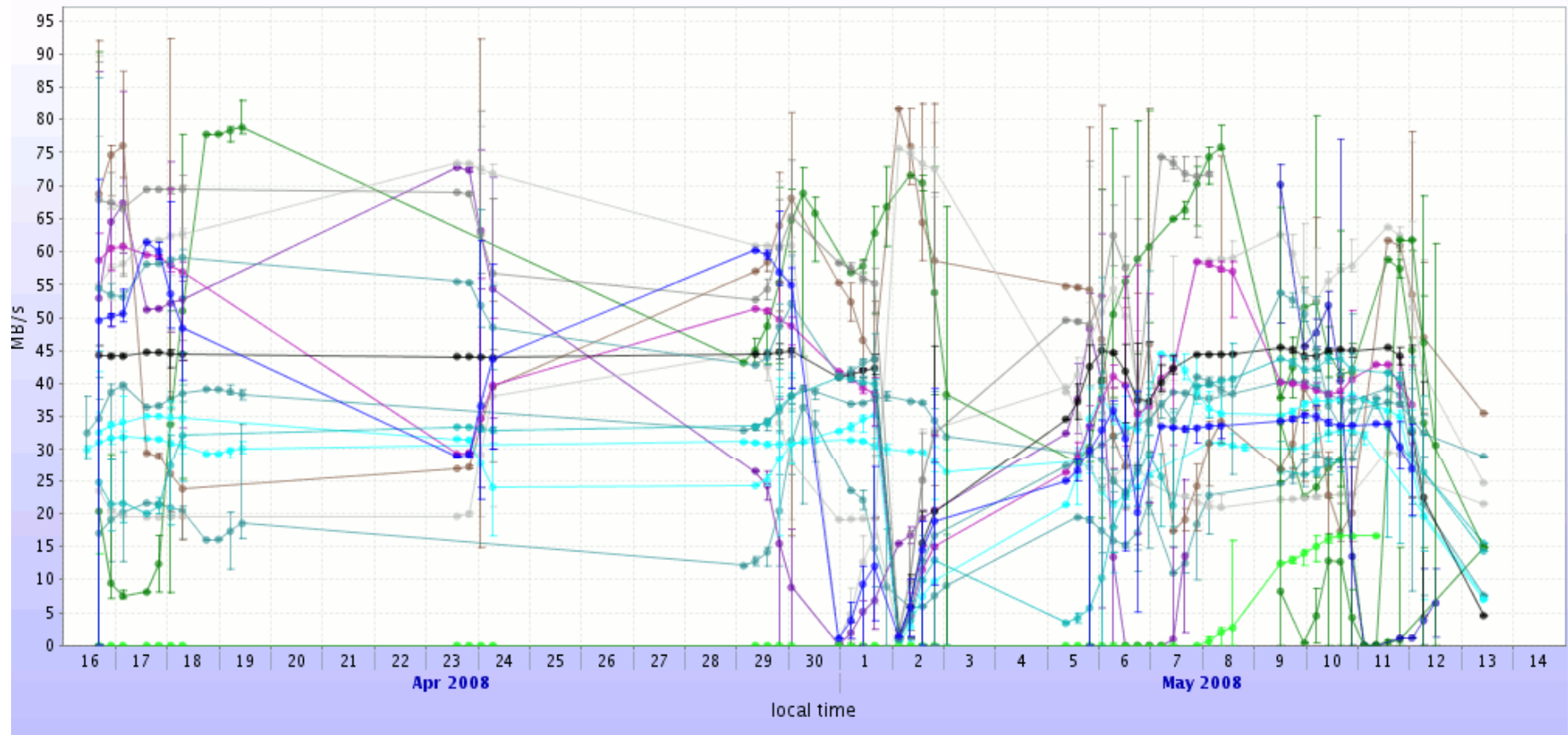


- bnl->bu(dcdoor04->atlas)
- bnl->iu(dcdoor04->iut2-grid6)
- bnl->ou(dcdoor04->tier2-02)
- bnl->stanford(dcdoor04->atl-prod05)
- bnl->uchicago(dcdoor04->uct2-grid6)
- bnl->umich(dcdoor04->gate01)
- bnl->umich(dcdoor04->umfs05)
- bnl->wisc(dcdoor04->atlas03)
- bu->bnl(atlas->dcdoor04)
- iu->bnl(iut2-grid6->dcdoor04)
- ou->bnl(tier2-02->dcdoor04)
- stanford->bnl(atl-prod05->dcdoor04)
- uchicago->bnl(uct2-grid6->dcdoor04)
- umich->bnl(gate01->dcdoor04)
- umich->bnl(umfs05->dcdoor04)
- wisc->bnl(atlas03->dcdoor04)

GridFTP History Plots



Gridftp Load Tests (8 streams, 2MB window)



- bnl->bu(dcdoor04->atlas) - bnl->iu(dcdoor04->iut2-dc1) - bnl->ou(dcdoor04->tier2-02) - bnl->stanford(dcdoor04->atl-prod05)
- bnl->uchicago(dcdoor04->uct2-dc1) - bnl->umich(dcdoor04->dq2) - bnl->umich(dcdoor04->gate01) - bnl->umich(dcdoor04->msufs05)
- bnl->uta(dcdoor04->gk01) - bnl->wisc(dcdoor04->atlas03) - bu->bnl(atlas->dcdoor04) - iu->bnl(iut2-dc1->dcdoor04) - ou->bnl(tier2-02->dcdoor04)
- stanford->bnl(atl-prod05->dcdoor04) - uchicago->bnl(uct2-dc1->dcdoor04) - umich->bnl(dq2->dcdoor04) - umich->bnl(gate01->dcdoor04)
- umich->bnl(msufs05->dcdoor04) - uta->bnl(gk01->dcdoor04) - wisc->bnl(atlas03->dcdoor04)

perfSONAR in USATLAS



- ❄ There is a significant, coordinated effort underway to instrument the network in a standardized way. This effort, call **perfSONAR** is jointly supported by DANTE, Esnet, GEANT2, Internet2 and numerous University groups.
- ❄ Within USATLAS we will be targeting implementation of a **perfSONAR** instance for our facilities. Its primary purpose is to aid in network diagnosis by quickly allowing users to isolate the location of problems. In addition it provides a standard measurement of various network performance related metrics
- ❄ **USATLAS Plan:** Each Tier-2, and the Tier-1, will provide a dedicated system to run perfSONAR services. (more soon)
- ❄ See <http://www.perfsonar.net/> for information on **perfSONAR**

Near-term Steps for Working Group



- ❄ Lots of work is needed for our Tier-2's in end-to-end throughput.
- ❄ Each site will want to explore options for system/architecture optimization **specific to their hardware**.
- ❄ Most reasonably powerful storage systems should be able to exceed **200MB/s**. *Getting this consistently across the WAN is the challenge!*
- ❄ For each Tier-2 we need to tune GSIFTP (or FDT) transfers between “powerful” storage systems to achieve “bottleneck” limited performance
 - ❑ Implies we document the bottleneck: I/O subsystem, network, processor?
 - ❑ Include 10GE connected pairs where possible...target **400MB/s/host-pair**
- ❄ For those sites supporting SRM, trying many host pair transfers to “go wide” in achieving high-bandwidth transfers.

Throughput Goals



We have a set of goals for our “Throughput” work:

<u>Goal</u>	<u>AGLT2</u>	<u>IU</u>	<u>UC</u>	<u>NET2</u>	<u>OU</u>	<u>UTA</u>	<u>WT2</u>	<u>WISC</u>
200MB/sec > 2 hrs	YES	100	YES	80	Not done	Not Done	YES	YES
400MB/sec (10GE) > 2 hrs	YES	NO	YES	NO	Not done	Not Done	NO	NO
500MB/sec (BNL->Mulit-Tier2) > 24 hrs					Yes			
1GB/sec BNL->All Tier-2s					No			
Max-rate to Tier-2 (30 minute avg)								

For the last two goals we are awaiting GridFTP upgrades at BNL to support this level of throughput. Right now this looks like mid-June

This is supposed to be a “living table” on our Twiki site...As we modify and upgrade our sites we retest to verify what the changes have done for our performance.

Plans and Future Goals



- ❄ We still have a significant amount of work to do to reach consistently high throughput between our sites.
- ❄ Performance analysis is central to identifying bottlenecks and achieving high-performance.
- ❄ Automated testing via BNL's MonALISA plugin will be maintained to provide both baseline and near-term monitoring.
- ❄ **perfSONAR will be deployed in the next month or so.**
- ❄ Continue “on-demand” load-tests to verify burst capacity and change impact
- ❄ **Finish all sites in our goal table & document our methodology**



Questions?

Some Monitoring Links



<http://gridtest01.racf.bnl.gov:8081/display?page=network>

http://gridtest01.racf.bnl.gov:8081/display?page=network_hist

https://svn.usatlas.bnl.gov/svn/usatlas-int/load-test/Control/conf/network_m2m

<http://gridtest01.racf.bnl.gov:8081/display?page=gridftp>

http://gridtest01.racf.bnl.gov:8081/display?page=gridftp_hist

https://svn.usatlas.bnl.gov/svn/usatlas-int/load-test/Control/conf/gridftp_d2d

<http://www.perfsonar.net/>