Argonne
NATIONAL
LABORATORY

... for a brighter future

Improving network for Tier3

D.Benjamin, S.Chekanov, R.Yoshida

U.S. ATLAS Tier 2/Tier 3 workshop
Chicago, August 19-20

U.S. Department
of Energy

UChicago ▸
Argonne LLC

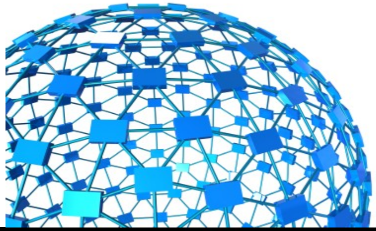
A U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC

Challenge II: How to bring data from Tier1/2 sites to ANL ASC & T3G sites. Can we use 1 Gbps full network bandwidth?

“Last-mile paradox”:

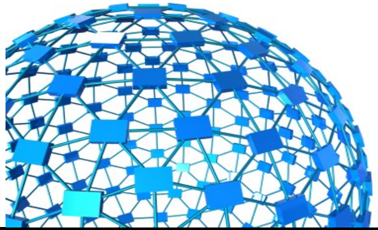
How 1 Gbps bandwidth translates into 20 Mbps for end users

- **10 Gbps up-link comes to ANL and connects HEP via 2 Gbps fibers**
- **1 Gbps Netgear switches and network cards**
- **Single-thread download rate for default (SL5.3) Linux installation:**
  - 600 Mbps for sites inside ANL
  - 100 Mbps with U.Chicago
  - 20-30 Mbps with any other remote site (BNL, SLAC, CERN etc)
- **Unacceptable taking into account our goal (~4 TB/day) for 1 Gbps**
- **Common problem for many Tier3 sites?**



Esnet recommendations: http://fasterdata.es.net/

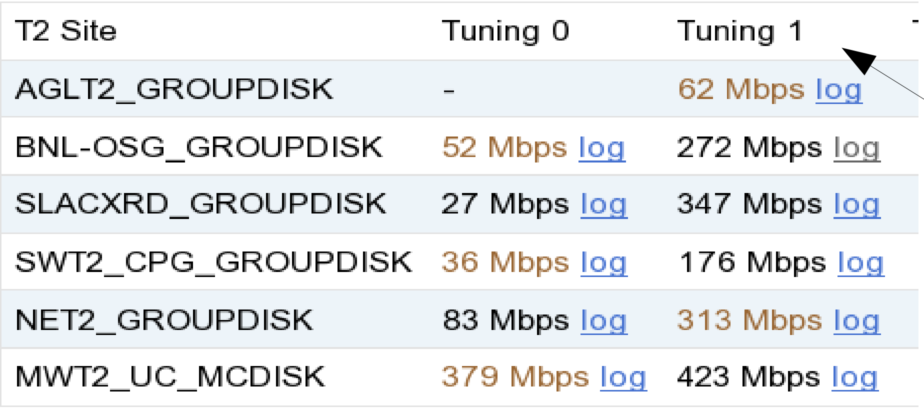
- **Esnet web site:**
  - "Moving a TeraByte between most large research institutions in the US should only take around 8 hours"
- **But network should be tuned (http://fasterdata.es.net/tuning.html):**
  - 1) Increase TCP buffer size for Linux
    - *For ANL, download rate increased by **factor 4** with SLAC/BNL!*
  - 2) Use newest Linux kernels
    - *Not tried. But some small difference between 2.6.9 and 2.6.18 kernels*
  - 3) Increase buffer size in 10 Gbps →1 Gbps switches (if applicable)
    - *Tried by network people. Unsuccessful for current switches*



**ASC network was tunned with the help of
Eli Dart and many other ESnet and ANL network people**

Getting data from Tier1/2 to ASC ANL

Recent stress tests using “dq2-get” (default: 3 threads)
Data: *user.RichardHawking.0108173.topmix_Egamma.AOD.v2 (125 GB)*
Use a bash script with dq2-get for benchmarking



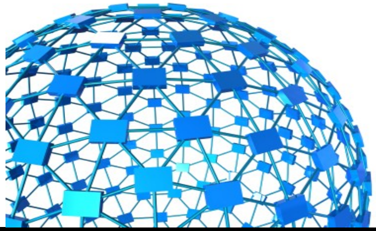
T2 Site	Tuning 0	Tuning 1
AGLT2_GROUPDISK	-	62 Mbps log
BNL-OSG_GROUPDISK	52 Mbps log	272 Mbps log
SLACXRD_GROUPDISK	27 Mbps log	347 Mbps log
SWT2_CPG_GROUPDISK	36 Mbps log	176 Mbps log
NET2_GROUPDISK	83 Mbps log	313 Mbps log
MWT2_UC_MCDISK	379 Mbps log	423 Mbps log

SL 5.3 TCP tune
Recommended
by ESnet

Brown color: at least one file has 0 size

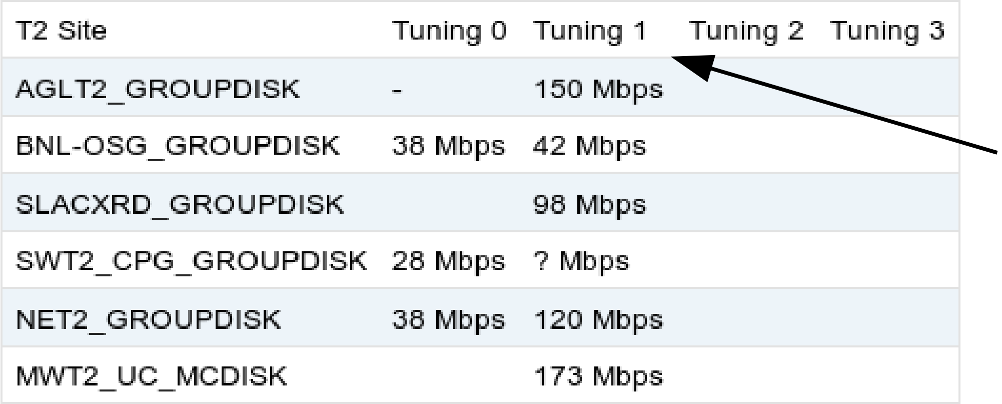
Satisfactory for MidWest Tier2 (UChicago) ~ 50 MB/s (4.5 TB/day, other sites ~3 TB/day)

For a single thread, the network speed is < 120 Mbps (using 1 Gbps uplink!)



Getting data from Tier1/2 to U.Duke

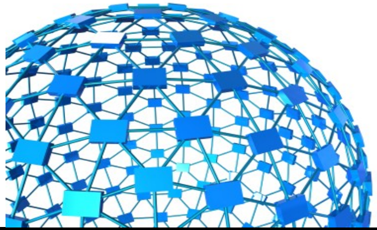
Recent stress tests using “dq2-get”: (3 threads)
Data: *user.RichardHawking.0108173.topmix_Egamma.AOD.v2 (125 GB)*
Use a bash script for benchmarking



T2 Site	Tuning 0	Tuning 1	Tuning 2	Tuning 3
AGLT2_GROUPDISK	-	150 Mbps		
BNL-OSG_GROUPDISK	38 Mbps	42 Mbps		
SLACXRD_GROUPDISK		98 Mbps		
SWT2_CPG_GROUPDISK	28 Mbps	? Mbps		
NET2_GROUPDISK	38 Mbps	120 Mbps		
MWT2_UC_MCDISK		173 Mbps		

SL 5.3 TCP tune
Recommended
by ESnet

Factor ~ 4 improvement with BNL, NET2 using Esnet recommendations
Below ANL numbers, but 2TB/day has achieved

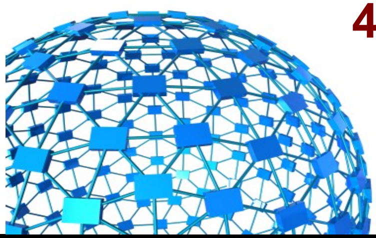


Getting data from Tier1/2 to ASC ANL

- **Even after TCP tunning, network bandwidth is ~100 Mbps for single thread download (~300 Mbps for dq2-get)**
  - Reason: packet loses in 10 Gbps →1 Gbps switches
- **Possible solution: use multiple dq2-get threads**
  - Split dataset on equal subsets. Create a file list
  - Run dq2-get on each PC farm node in parallel using the file list

- **ANL solution: Use a front-end of dq2-get included into the ArCond package:**
  - *arc_ssh -h hosts-file -l <user-name> -o /tmp/log "exec send_dq2.sh”*
    - *Gets a list of files. Splits in ranges depending on number of slaves.*
    - *Executes dq2-get on each slave using this list.*
  - Tested using 5 Linux boxes (five dq2-get threads)

4 TB/day from BNL/SLAC achieved after using 2-3 dq-get threads



Summary

- **Download rate is acceptable after TCP tunning of the PC farm**
  - A tool for downloads using multiple dq2-get was tested (included to ArCond)
- **ANL is moving towards 10 Gbps network setup:**
  - Network switch with 10 Gbps uplink & 1 Gbps ports
    - $9k for 48 Gbps ports, WS-C4948-10GE
  - ~25-30 TB/day using multiple dq2 threads?
- **dq2-get Stress Test documentation (including log files)**
  - https://atlaswww.hep.anl.gov/twiki/bin/view/ASC/Dq2_getStressTest
- **How to use dq2-get in multiple threads using ArCond and TCP recommendations:**
  - https://atlaswww.hep.anl.gov/twiki/bin/view/Tier3Setup/T3gGettingDataPCfarm