Argonne NATIONAL LABORATORY

… for a brighter future

U.S. Department of Energy

UChicago ▶ Argonne LLC

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

ANL T3g infrastructure

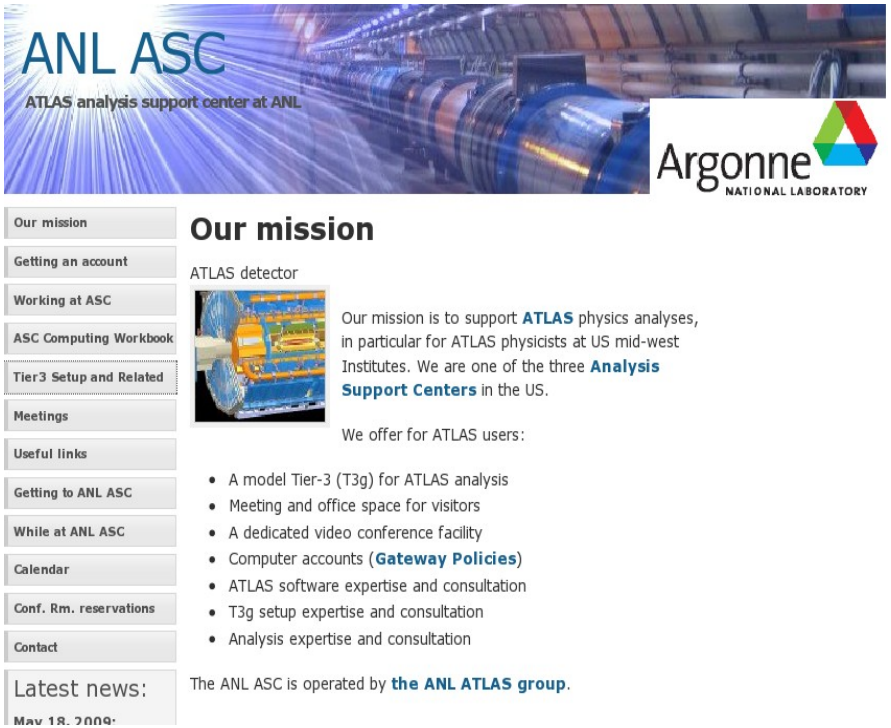
S.Chekanov

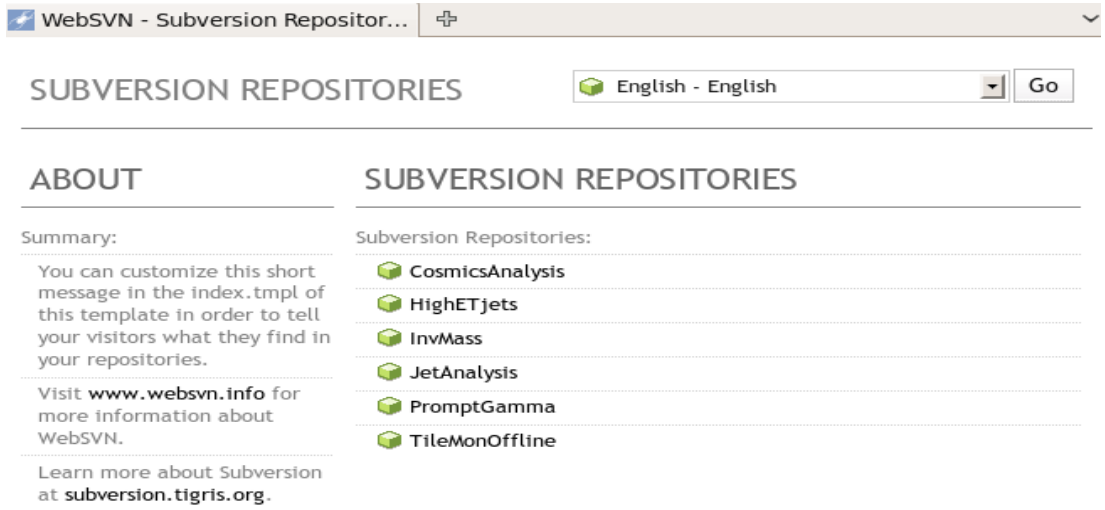
(HEP Division, ANL)

ANL ASC Jamboree
September 2009

ANL ASC T3g

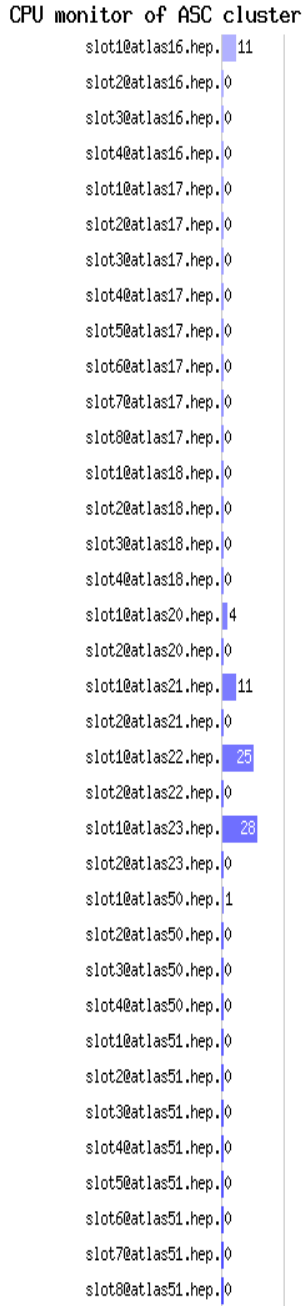
- ~90 registered users
- Twiki with:
  - ASC computer workbook
  - Tier3 setup guide
- Could provide Indico service
  - not activated
- CVS
- Together with CERN, migrated to SVN
- SVN and SVN browser: http://atlaswww.hep.anl.gov/asc/WebSVN/
  - 6 supported packages:
    - *HighETJets*
    - *PromptGamma,*
    - *InvMass*
    - *CosmicAnalysis*
    - *JetAnalysis*





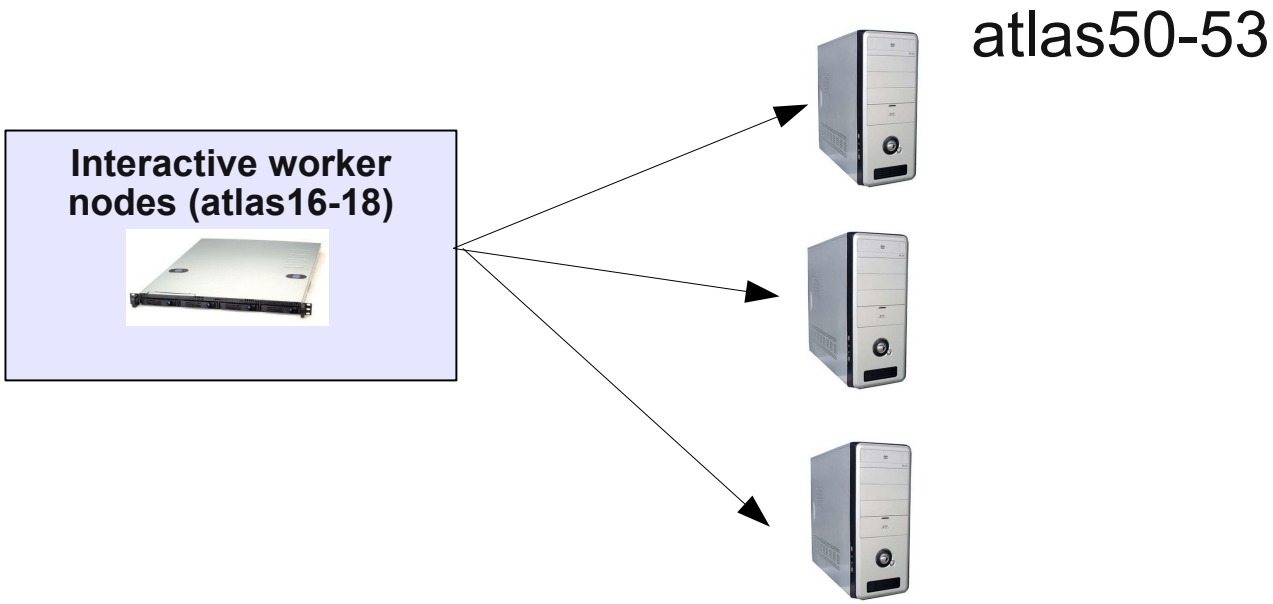
ANL T3g computing

- **About 50 cores chained by Condor**
- **Interactive nodes: atlas16,17,18 (16 cores), SL4.8**
- **User scratch disk space (~5 TB)**
  - Excluding NFS, data storage
- **PC farm prototype**
  - ArCond/Condor
  - 24 cores, 6TB data storage
- **Software:**
  - Grid, pathena, OSG-client, dq2-get
  - All major atlas releases including 15.4.0
    - atlas18 contains locally-installed releases
  - SL5.3: validation computers (atlas11,50) + all desktop computers
  - only 15.2.0 release. Setup is exactly the same
    - Change: set_atlas.sh to set_atlas_sl5.sh
    - gcc432!
- **Migration to SL5.3 in ~1 month**



ANL T3g cluster design

- **Prototype was designed (24 cores) and operational since Sep. 2008**
- **Fully satisfies to the T3G requirements:**
  - Grid access
  - 24-core cluster with Arcond/Condor
  - 2 TB/8 cores, data “pre-staged” (local to disks). No network load.
  - Other 25 cores (atlas16-27) can also be used by Condor, but no local data



Data sets on the PC farm

- **Since Sep. 2008, we store 17422 AOD MC files**
  - ~ 4M Monte Carlo AOD events (+ few ESD sets)
  - Corresponds to ~**25%** of the total capacity of the PC farm prototype

/data1/mc/gamma_jet/pt17/AOD	atlas52	gamma+jet samples, r14.2, pt>17 GeV. Also available: pt40, pt8 pt600
/data1/mc/pythia_gfilter/pt17/AOD	atlas51	Filtered background sample, r14.2, pt>17 GeV. Also available: pt pt400, pt600
/data1/mc/PythiaZeegam25/AOD	atlas51-52	Z+gamma+X samples, r14.2, pt>25 GeV
/data1/mc/BaurZeegam/AOD	atlas51	Z+gamma+X, Baur MC, r14.2, pt>25 GeV, X-section=463.622 p each file
/data1/mc/mc08.105802.JF17_pythia_jet_filter.recon.AOD.e347_s462_r541/AOD	atlas51-53	~1.5 M events, inc.Pythia after JetFilter, r14.2, pt>17
/data1/mc/mc08.106070.PythiaZeeJet_Ptcut.recon.AOD.e352_s462_r541/AOD	atlas51-53	Z->e+e- + jet events, r14.2.20, 250 events in each file, 797 files, 968.637 pb, efficiency = 0.90
/data1/mc/mc08.106071.PythiaZmumuJet_Ptcut.recon.AOD.e352_s462_r541/AOD	atlas51-53	Z->mu+mu- + jet events, r14.2.20, 250 events in each file, 791 file 968.637 pb, efficiency = 0.90
/data1/mc/mc08.106072.PythiaZtautauJet_Ptcut.recon.AOD.e352_s462_r541/AOD	atlas51-53	Z->tau+tau- + jet events, r14.2.20, 250 events in each file, 759 file 968.637 pb, efficiency = 0.90
/data1 /mc/mc08.106379.PythiaPhotonJet_AsymJetFilter.recon.AOD.e347_s462_r541/AOD	atlas51-53	250k events, gamma+jet, ckin(3)>15 GeV
/data1/mc/MC08/JS0/ESD	atlas53	also JS1, JS2,JS3,JS4,JS5,JS6,JS7 available. Talk to Belen a
/data1/mc/mc08.107141.singlepart_pi0_Et40.recon.AOD.e342_s439_r546/AOD	atlas51	200 files, r14.2.20.3, single pi0
/data1/mc/mc08.107041.singlepart_gamma_Et40.recon.AOD.e342_s439_r546/AOD	atlas51	189 files, r14.2.20.3, single gamma
/data1/mc/mc08.107680.AlpgenJimmyWenuNp0_pt20.recon.AOD.e349_a68/AOD	atlas51-53	1202 files, r14.2.20, W->e+nu+0 partons
/data1/mc/mc08.107681.AlpgenJimmyWenuNp1_pt20.recon.AOD.e349_a68/AOD	atlas51	242 files, r14.2.20, W->e+nu+1 partons
/data1/mc/mc08.107682.AlpgenJimmyWenuNp2_pt20.recon.AOD.e349_a68/AOD	atlas51	624 files, r14.2.20, W->e+nu+2 partons
/data1/mc/mc08.107683.AlpgenJimmyWenuNp3_pt20.recon.AOD.e349_a68/AOD	atlas51	165 files, r14.2.20, W->e+nu+3 partons
/data1/mc/mc08.107684.AlpgenJimmyWenuNp4_pt20.recon.AOD.e349_a68/AOD	atlas51	48 files, r14.2.20, W->e+nu+4 partons
/data1/mc/mc08.107685.AlpgenJimmyWenuNp5_pt20.recon.AOD.e349_a68/AOD	atlas51	22 files, r14.2.20, W->e+nu+5 partons

FDR2 reprocessed data: ||

/data1/mc/fdr08_run2.0052280.physics_Egamma.recon.AOD.o3_f47_r575/AOD	atlas51-53	FDR2 AOD data, release 14.2.24
/data1/mc/fdr08_run2.0052280.physics_Egamma.recon.DPD_CALOJET.o3_f47_r575/AOD	atlas51-53	FDR2 DPD data, release 14.2.24
/data1/mc/fdr08_run2.0052280.physics_Egamma.recon.DPD_EGAMMA.o3_f47_r575/AOD	atlas51-53	FDR2 DPD data, release 14.2.24
/data1/mc/fdr08_run2.0052280.physics_Egamma.recon.DPD_PHOTONJET.o3_f47_r575/AOD	atlas51-53	FDR2 DPD data, release 14.2.24
/data1/mc/fdr08_run2.0052280.physics_Jet.recon.AOD.o3_f47_r575/AOD	atlas51-53	FDR2 AOD data, release 14.2.24

Benchmarking results for 24 cores (Xeon 2.3 GHz)

Most tests done with PromptGamma package (ANL SVN)
Accessing all AOD containers + Jets/gamma/e/muons/taus/missET are written to ntuples
Data local to each CPU (3 nodes, 8 core per node, 33% of data on each box)

- **Running over AOD files**
  - 0.5M events /h
- **Fast MC simulation and on the fly analysis**
  - 1.5M events /h
- **Running over C++/ROOT ntuples**
  - 1000M events /h (1M events / min for 1 core)
- **Generating MC truth ntuples**
  - 2.5M events /h
- **AOD production (generating & reconstructing MC events)**
  - 120 events /h

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 (3 threads)

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

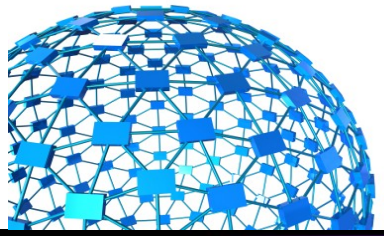
SL 5.3 TCP tune Recommended by ESnet

Brown color: at least one file has 0 size

R.Yoshida & checks by D.Benjamin for Duke's T3

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

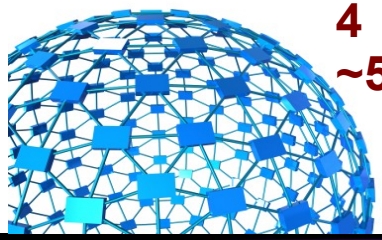
For a single thread, the network speed is < 120 Mbps (using 1 Gbps uplink!)



Getting data from Tier1/2 to ASC ANL

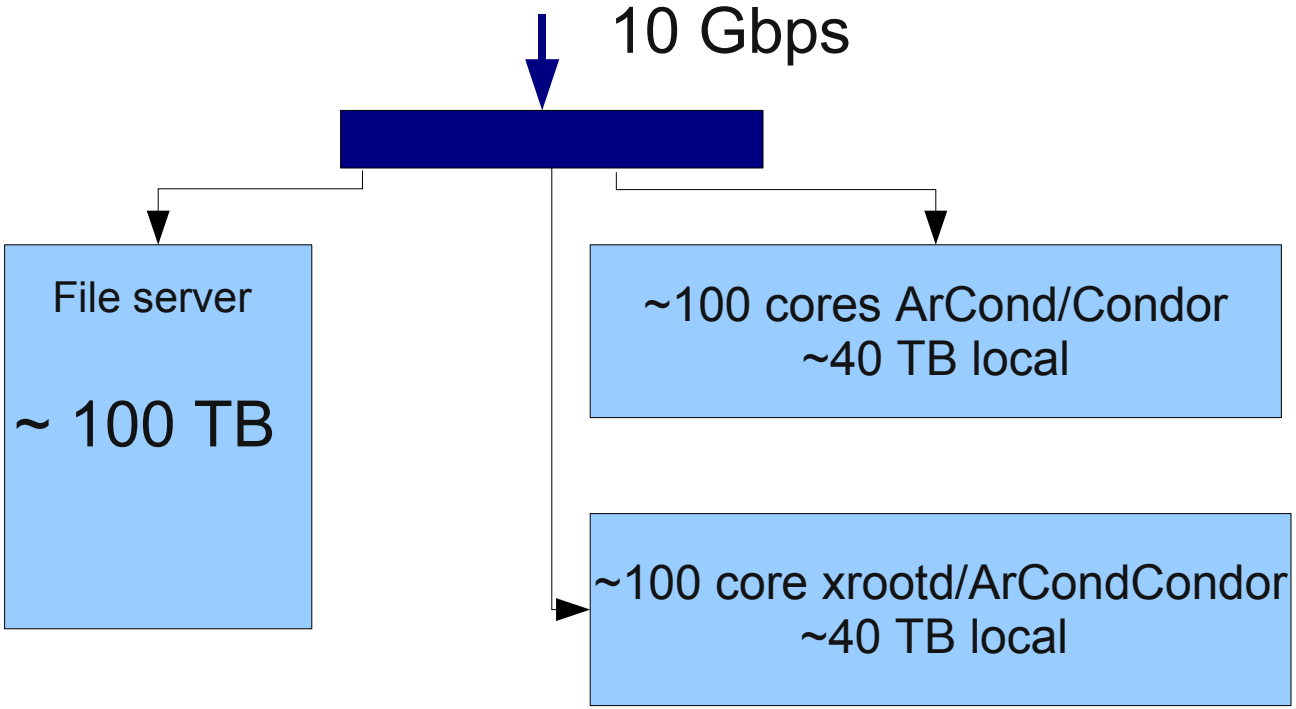
- **Even after TCP tunning, network bandwidth is ~100 Mbps for single thread download (~300 Mbps for dq2-get)**
  - Reason: packet loses in 10 Gbps →1 Gbps switches
- **Possible solution: take advantage of the PC farm design and use multiple dq2-get threads on each PC farm box**
  - Split dataset on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

- **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)



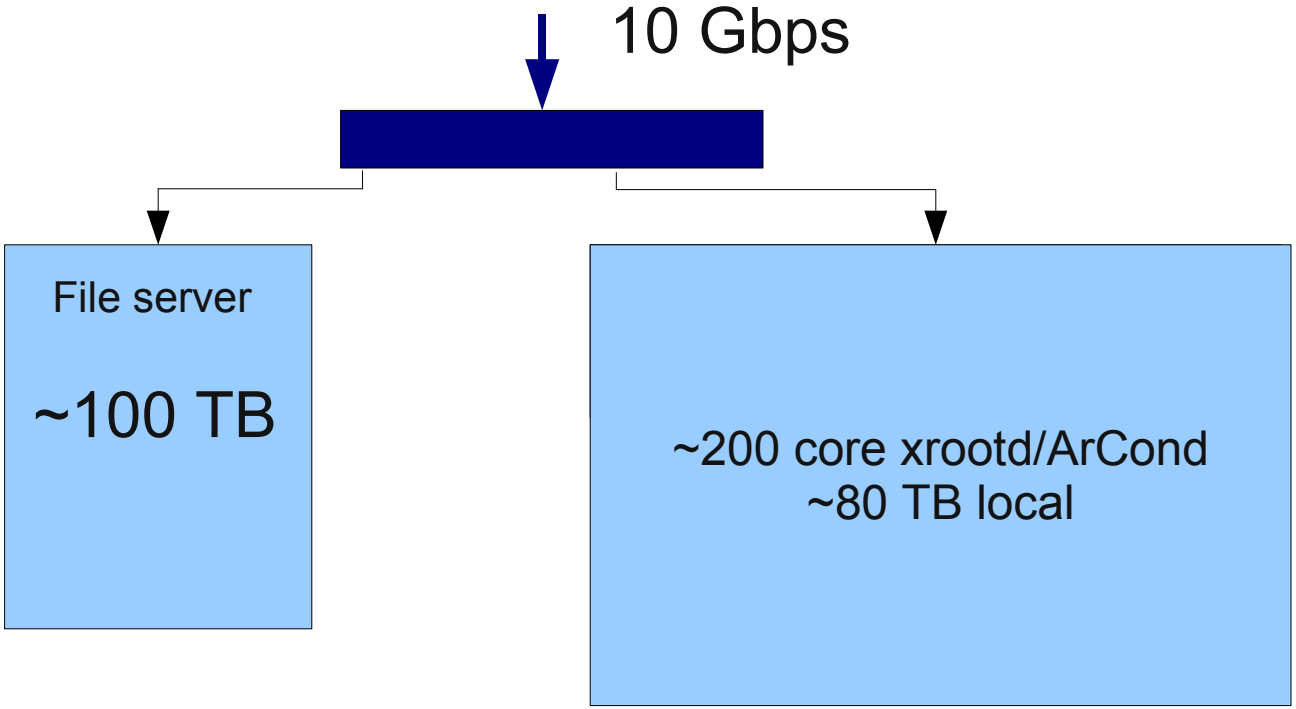
4 TB/day from BNL/SLAC achieved after using 2-3 dq2-get threads
~5 TB/day from Uchicago using a single dq2-get

ANL Tier3 future (~1 year)



PC farm nodes are based on Dell PowerEdge R710
(2 processors, Xeon 5520 (8 cores) + 2.5 GB per core + 4TB local data disks)

Future (~2 years from now)



all based on Xeon 5520 + 2GB per core

Expected performance: 24 Xeon 5404 cores (now) vs 200 Xeon 5520 (future)

Some reviews claim:
5500 (Nehalem) processors are 50%-100% faster than Harpertown Xeon (5400)
Assume 50% (benchmarks are coming):

- **Running over AOD files**
  - 0.5M events /h → **6M/h**
- **Fast MC simulation and on the fly analysis**
  - 1.5M events /h → **18M/h**
- **Running over C++/ROOT ntuples**
  - 1000M events /h (1M events / min for 1 core). **10B? I/O limit?**
- **Generating MC truth ntuples**
  - 2.5M events /h → **30M/h**
- **AOD production (generating & reconstructing MC events)**
  - 120 events /h → **1400/h**