… for a brighter future

T3g cluster with distributed file storage

S.Chekanov

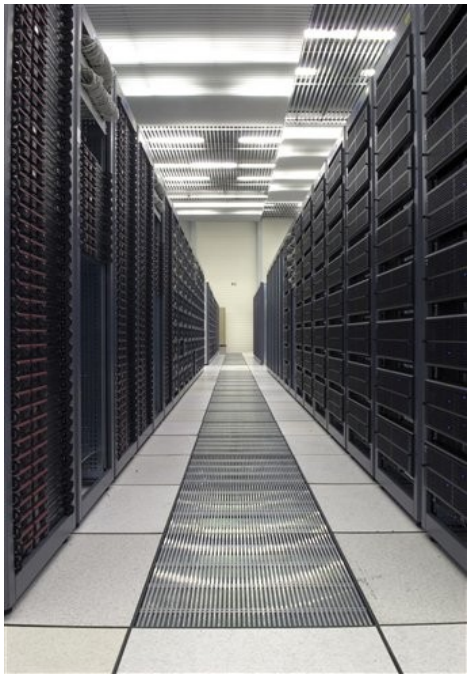
(HEP Division, ANL)

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

UChicago ▶
Argonne LLC

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

Low-cost PC farm cluster: challenges for ANL ASC and Tier3s

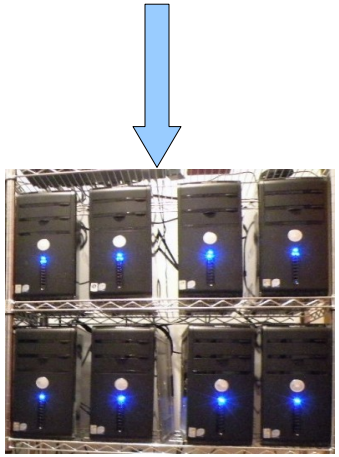


The US ATLAS Tier 3 Task Force Report of Spring 2009, concludes:

enhanced ATLAS analysis computing capabilities at home Universities of US ATLAS members are needed. Such capabilities are broadly called Tier3 computing

- essential for "chaotic" and "interactive" data analysis

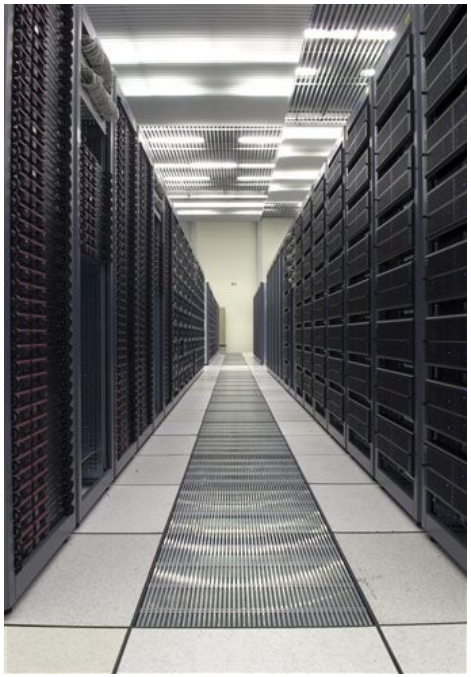
Points to the existing cluster prototype designed at ANL as a possible solution for data analysis for small or medium size HEP group (10-20 people)



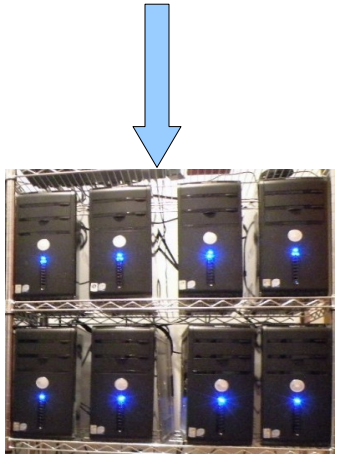
Challenges for T3 computing:

- How to build a low-cost (tens $k) cluster designed for heavy I/O (processing tens of TB /day)
- How to take advantage of 1 Gbps network bandwidth to transfer data from Tier1/2

Requirements for Tier3 cluster (T3g)



- **Interactive & chaotic analyses**
- **No resource allocation and file staging for each job execution**
  - faster data processing compared to the grid
- **Low cost: tens of $k.**
  - **~**$25k for processing power 0.5 TB/h of AOD files
- **Off-the-shelf hardware**
- **Small effort in management (0.2FTE)**
- **No special network requirement & computer room**
- **Fully scalable, no I/O bottleneck**
- **Run long jobs “by agreement”**



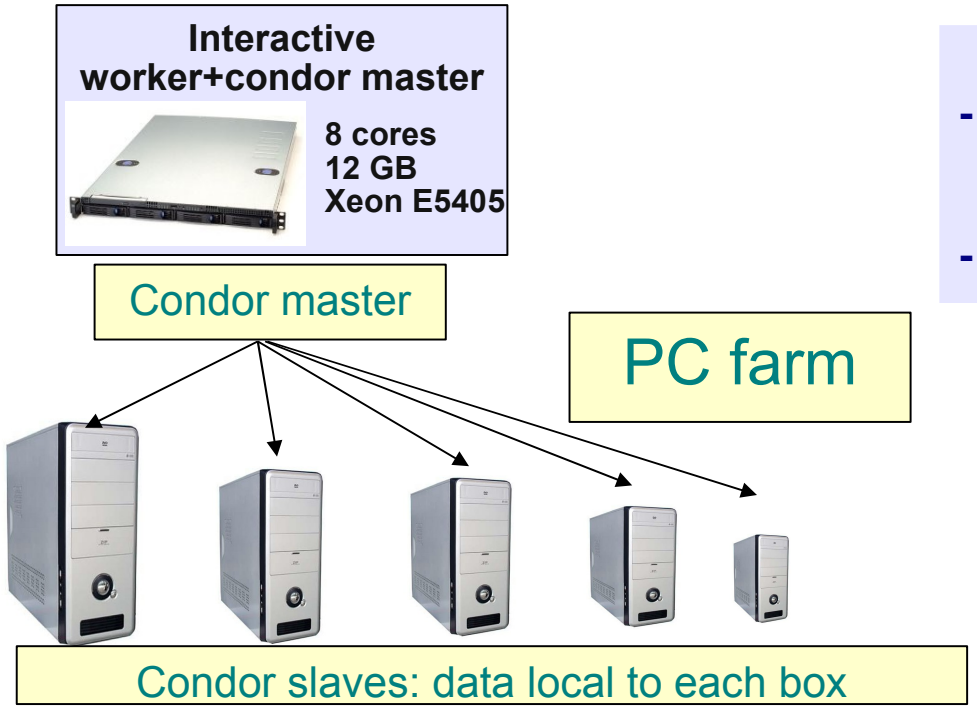
Two possible solutions for I/O intensive cluster

- **Data storage is central. Read data via NFS/AFS**
  - Good file storage is expensive
  - Load balancing is difficult - need to share file systems via NFS or other mechanisms to provide a central location for the data
  - 1 Gbps local network is not enough to support >20 CPUs accessing same data storage

- **Distributed data storage**
  - Each dataset distributed between several Linux boxes & local disks
  - No central file storage
  - No network load at runtime
  - Requires R&D

HEP ANL cluster design for T3g sites

- **Prototype was designed (24 cores) and operational since Sep. 2008**
- **Fully satisfies to the T3G requirements:**
  - Grid access
  - Commodity computers, no central file storage.
  - 2 TB/8 cores, data “pre-staged” (local to disks). No network load.
  - Low-cost: ~$25k for 80 nodes and 20 TB local storage
  - Processing power: 5-10TB ATLAS (AOD) data per day



- No single-point failure
Failed disk? Get data from the grid!

- End-user and operational simplicity

Similar setup at Duke U. and OSU

U. Duke cluster design for T3g sites (D.Benjamin)

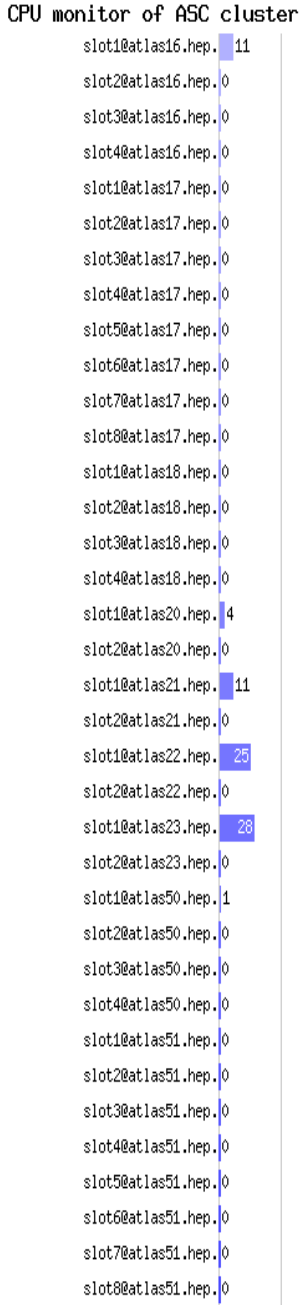
- **http://hep-atlas.phy.duke.edu/DukeTier3**
- **Based on:**
  - Condor cluster with local storage (24 nodes)
  - Central storage: Xrootd, BeStMan-Gateway, GridFTP
    - *Not yet integrated into condor farm*
  - NFS storage for ATLAS releases

Software required

- Scientific Linux (4.7)
- OSG-client (Condor)
- DQ2
- Atlas release + pathena
- Arcond (http://atlaswww.hep.anl.gov/asc/arcond/)

A Condor front-end for:

- job submission
- data discovery (using a static data base or on the fly)
- checking job status
- merging outputs
- data upload in multiple threads of dq2



Possible T3g architectures based on Condor/Arcond

Single-user workstation

- Data local to CPU
- Not scalable
- Max cores 8-16

Multi-user setup

NFS/AFS data server

- Data on NFS
- Scalable up to ~20 cores
- Require 1 Gbps network

Multi-user setup

PC farm

- Data redistributed between disks
- Fully scalable.
- No particular network requirement
- No single-point failure

Multi-user ANL setup with central interactive node

PC farm

Interactive node with ssh

Users home directories

- Data redistributed between disks
- Fully scalable
- No particular network requirement
- No single-point failure
- Interactive node with ssh
- Home directories on NFS for easy maintenance

Stored data sets

- **Since Sep. 2008, we store 15422 AOD MC files**
  - ~ 4M Monte Carlo AOD events (+ few ESD sets)
  - Corresponds to **~25%** of the total capacity of the PC farm prototype
- **Data moved to each box after using dq2-get (ArCond provides such splitter).**

/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)

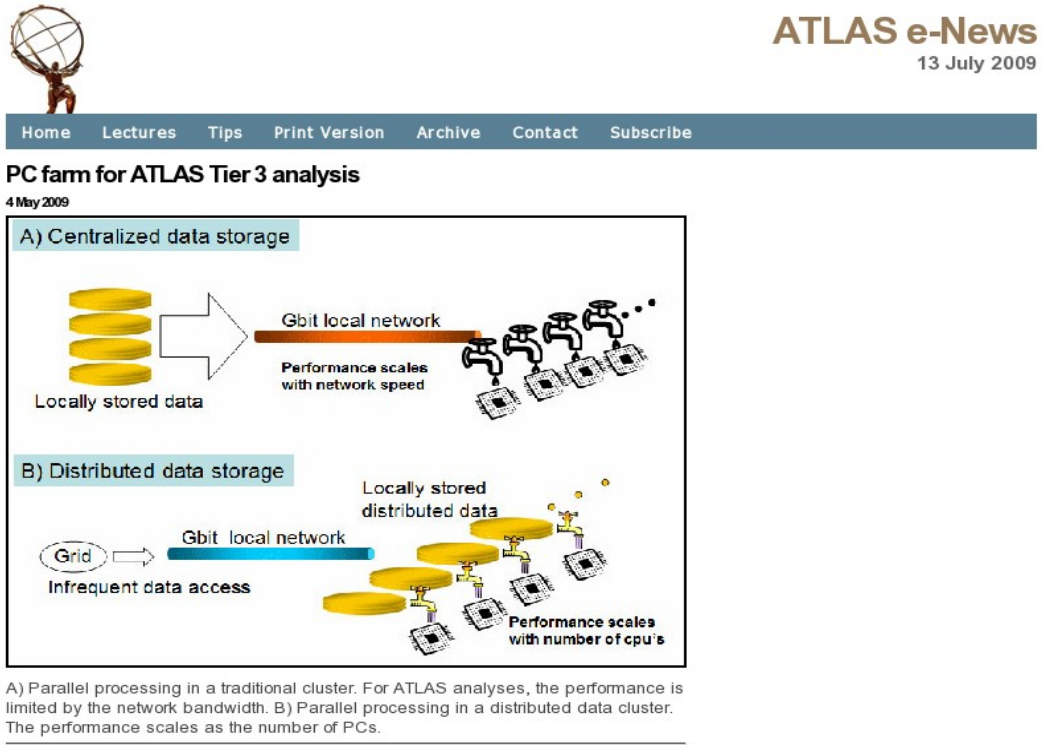
Most tests done with PromptGamma package (ANL SVN)
Accessing all AOD containers + Jets/gamma/e/muons/taus/missET are written to ntuples
Data local to each CPU (3 nodes, 8 core per node, 33% of data on each box)

- **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

PC farm challenge for T3g sites

- **A complete T3G PC farm setup is given on the ANL ASC page (atlaswww.hep.anl.gov):**

Article in ATLAS e-News



More details: "A PC farm for ATLAS Tier3 analysis" S.C., R.Yoshida, ATL-COM-GEN-2009-016

Summary

- **24-CPU PC farm prototype is fully functional**
  - $6k investment only
  - Man power: 0.5 FTE, which dropped to 0.1 FTE after the setup
  - Most of ANL results were done using the PC farm prototype (6 ATLAS notes)
- **Since Sep 1, 2008: ~300 submitted jobs (~7000 runs)**
  - no failures reported
- **T3g setup guide based on ArCond/Condor is available** (http://atlaswww.hep.anl.gov)
  - Includes hardware, software, setup and maintenance description
- **ATLAS Analysis Jamboree (Sept 9-11) at ANL:**
  - Tutorials: how to use ArCond/Condor & multiple cores for:
    - Processing ROOT ntuples using C++/ROOT
    - Fast MC simulation + on the fly analysis
    - Full MC simulation