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

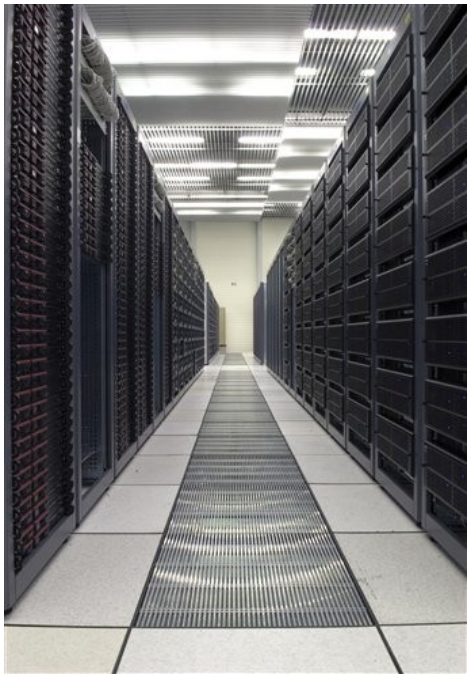
Parallel data processing at T3g

S.Chekanov

(HEP Division, ANL)

ANL ASC Jamboree
September 2009

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

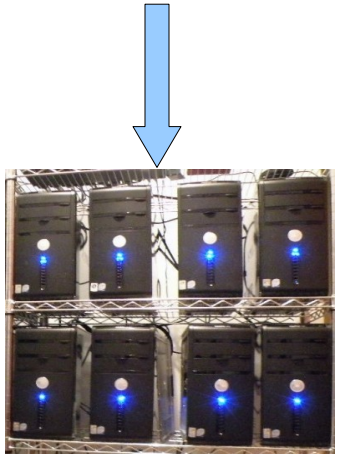


The US ATLAS Tier 3 Task Force Report of Spring 2009, concludes:

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

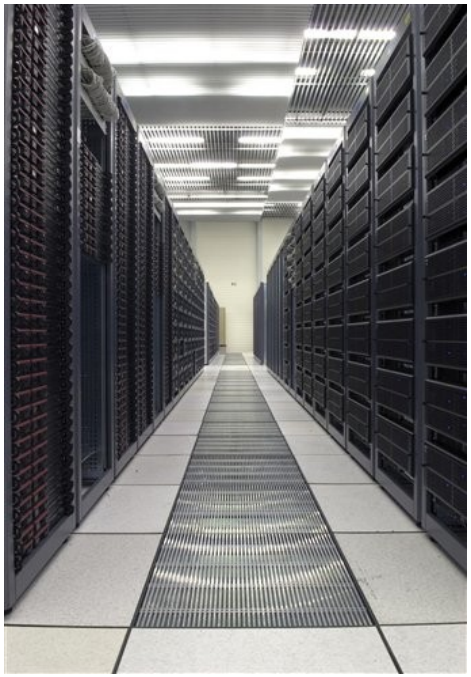
Points to the existing cluster prototype designed at ANL as a possible solution for data analysis for small or medium size HEP group (10-20 people)



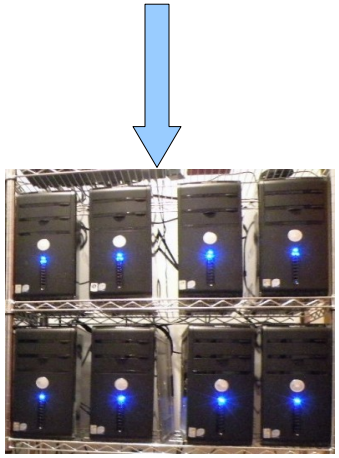
Challenges for T3 computing:

- 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)

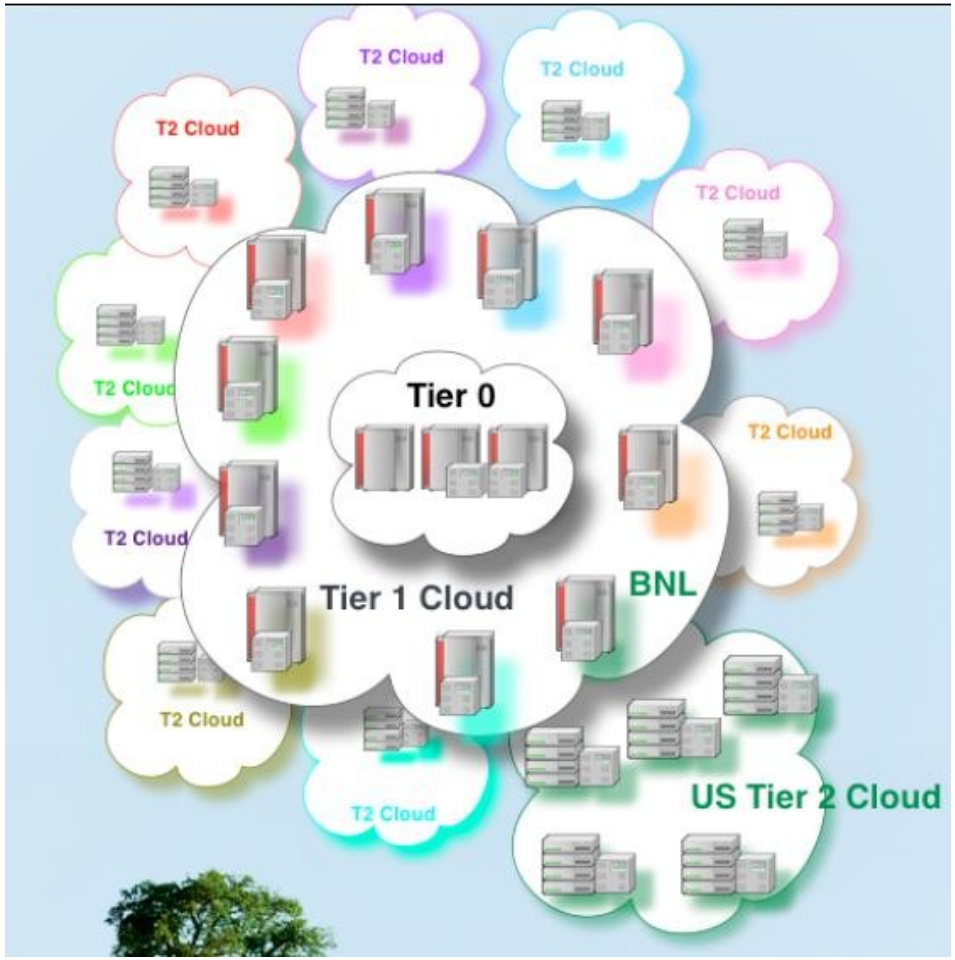


- **Interactive & chaotic analyses**
- **No resource allocation and file staging for each job execution**
  - faster data processing compared to the grid
- **Low cost: tens of $k.**
  - **~**$25k for processing power 0.5 TB/h of AOD files
- **Off-the-shelf hardware**
- **Small effort in management (0.2FTE)**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



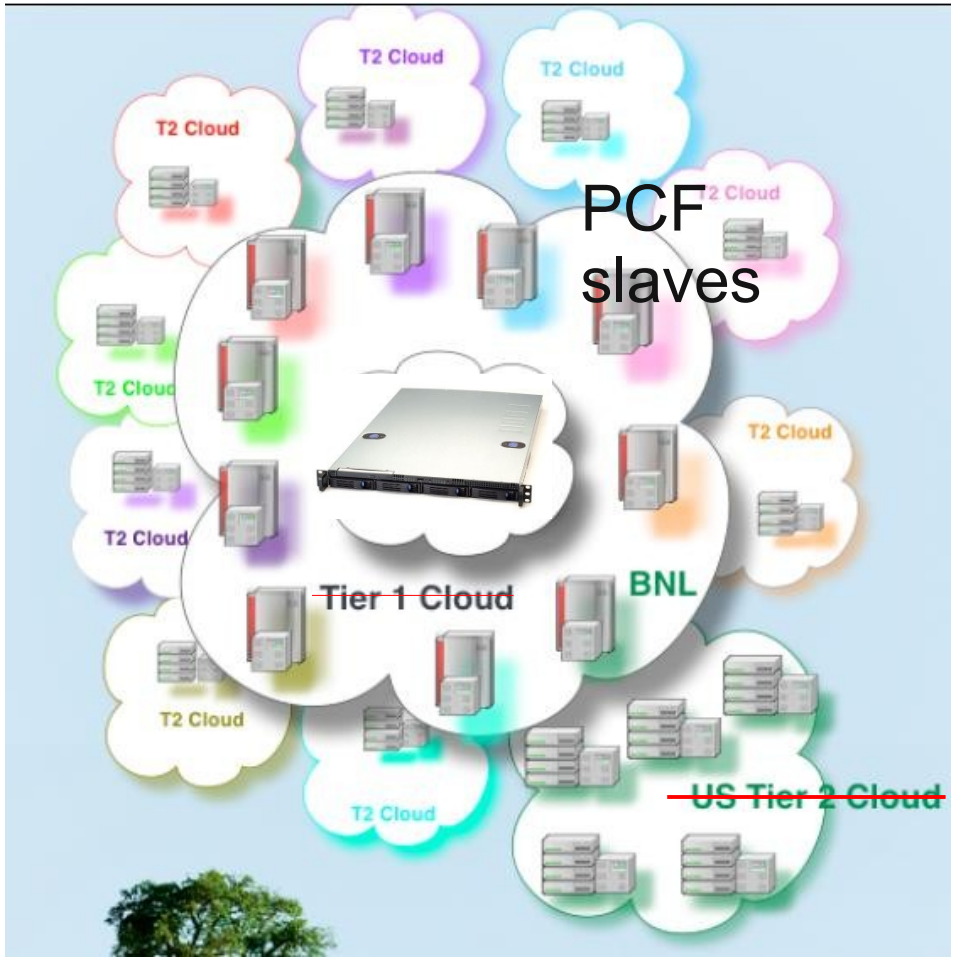
Grid is an operating system

See Rik's talk about the grid



- All you need to know is that:
- Are the data files I want to analyze in the Grid?
- Don't need to know where the data is.
- Don't need to know where your job is going to be running.

ANL PC farm as Grid-like operating system



- All you need to know is that:
- Are the data files I want to analyze in the Grid? PCF?
- Don't need to know where the data is.
- Don't need to know where your job is going to be running.

Fault rate <0.05%
(data are "pre-staged"!)

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

Possible T3g architectures based on Condor/Arcond/xrootd

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

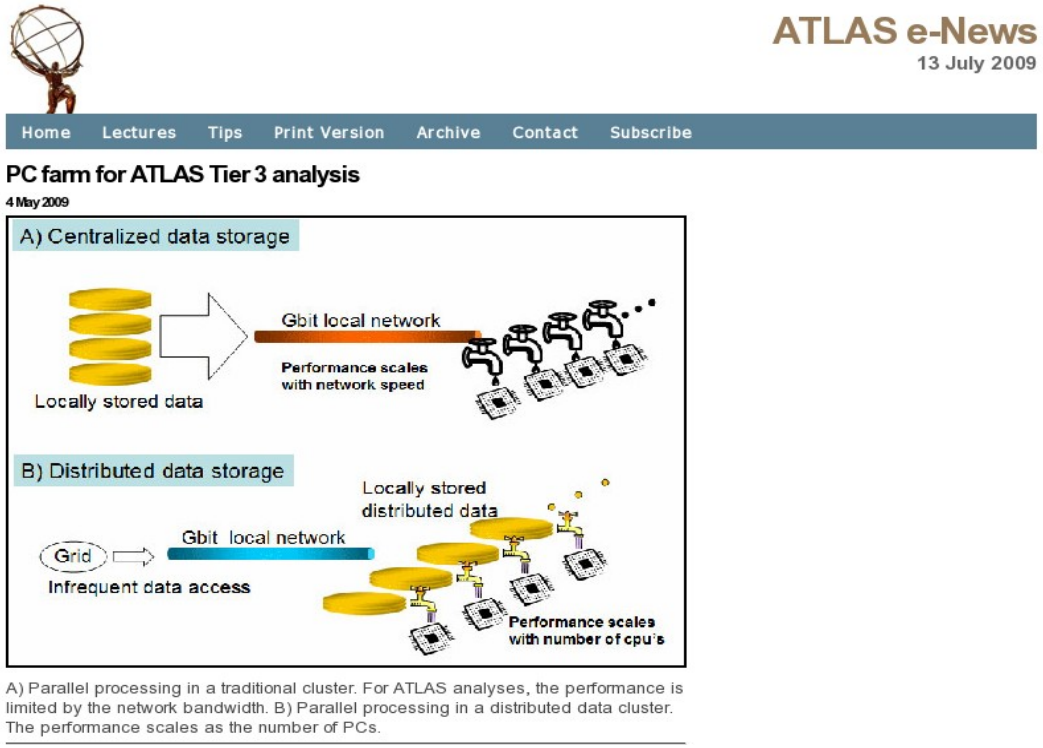
Users home directories

- 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

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

ANL T3g cluster design

- **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
- **dq2-get Stress test documentation (including log files) & Esnet tuning**
  - 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

To get started with ArCond

- **ArCond – "Argonne+Condor" for T3s computer farms:**
  - **Python front-end of Condor for: job submission, data discovery, results retrieval**
  - **Developed and supported at ANL ASC**
  - **http://atlaswww.hep.anl.gov/asc/arcond/**
- **How to get started with the ArCond:**
  - **setup atlas release:** https://atlaswww.hep.anl.gov/twiki/bin/view/Workbook/SettingUpAccount
  - **> mkdir test; cd test**
  - **> arc_setup**
  - **> arc_help (to see the commands)**
  - **> edit: arcond.conf (if needed). Pay attention to:**
    - atlas_release=15.4.0
    - events = 100
    - input_data = /data1/mc/mc08.108087.PythiaPhotonJetXXX
    - package_dir = /testarea/14.5.1/analysis/PromptGamma
    - max_jobs_per_node= -1
  - **> arcond (submit)**
  - **Check condor status as: *condor_q or condor_status***

This jamboree

- https://atlaswww.hep.anl.gov/twiki/bin/view/Jamborees/Jamboree2009SepPart3
- **How to use the PC farm:**
  - to run an athena code on a PC farm
  - to analyze ROOT ntuples using a PC farm
  - to run full Monte Carlo simulation and reconstruction