Argonne NATIONAL LABORATORY

… for a brighter future

U.S. Department of Energy

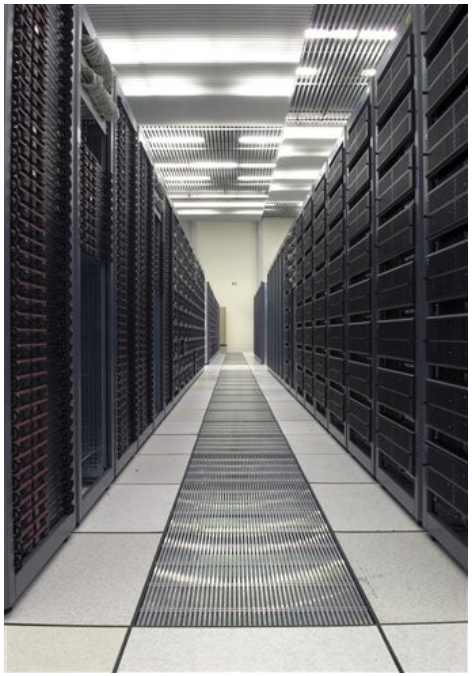
UChicago ▶ Argonne LLC

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

# *T3G example model*

***Sergei Chekanov*, *Rik Yoshida***

*ANL*

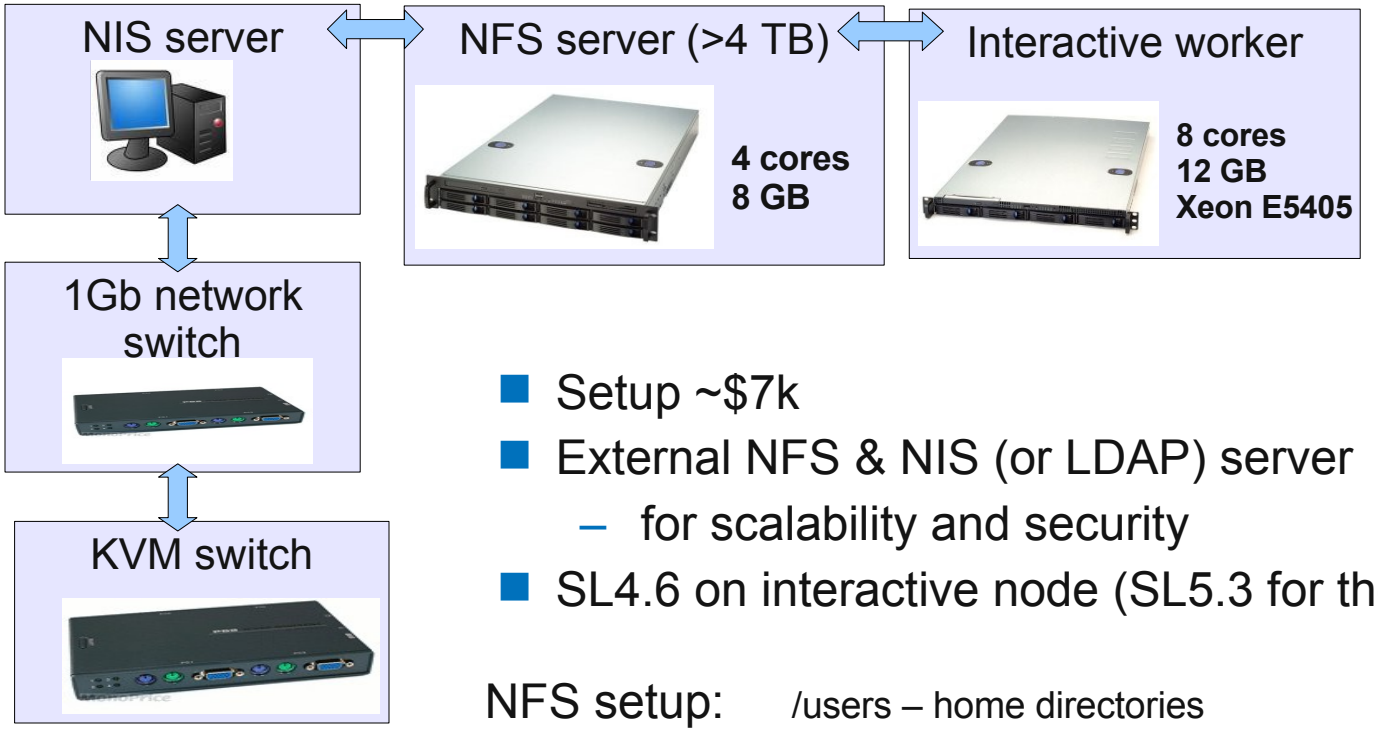


More details: “A PC farm for ATLAS Tier3 analysis” S.C., R.Yoshida, ATL-COM-GEN-2009-016

# *Requirements of T3G cluster*

- Allows “data staging” and “chaotic” job execution
- Allows interactive analyses (including PROOF)
- No resource allocation and file staging for each execution.
- 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)
- 1 Gb network
- Fully scalable, no I/O bottleneck

# Initial setup. Scalable interactive worker node



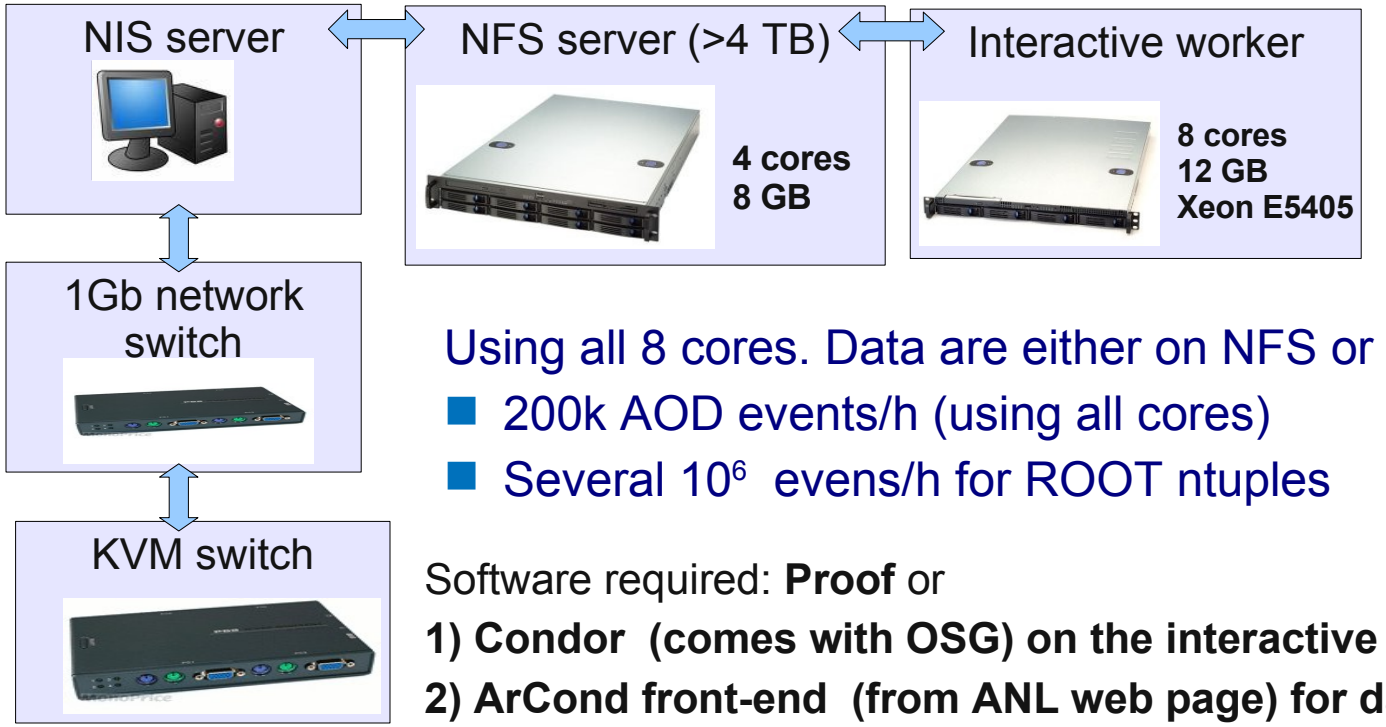

- Setup ~$7k
- External NFS & NIS (or LDAP) server
  - for scalability and security
- SL4.6 on interactive node (SL5.3 for the rest)

NFS setup:

/users – home directories
/data - data files
/share - ATLAS release + OSG

What can be done with such setup?

# *Initial setup. Scalable interactive worker node*




Using all 8 cores. Data are either on NFS or Interactive node:

- 200k AOD events/h (using all cores)
- Several $10^6$ evens/h for ROOT ntuples

Software required: **Proof** or

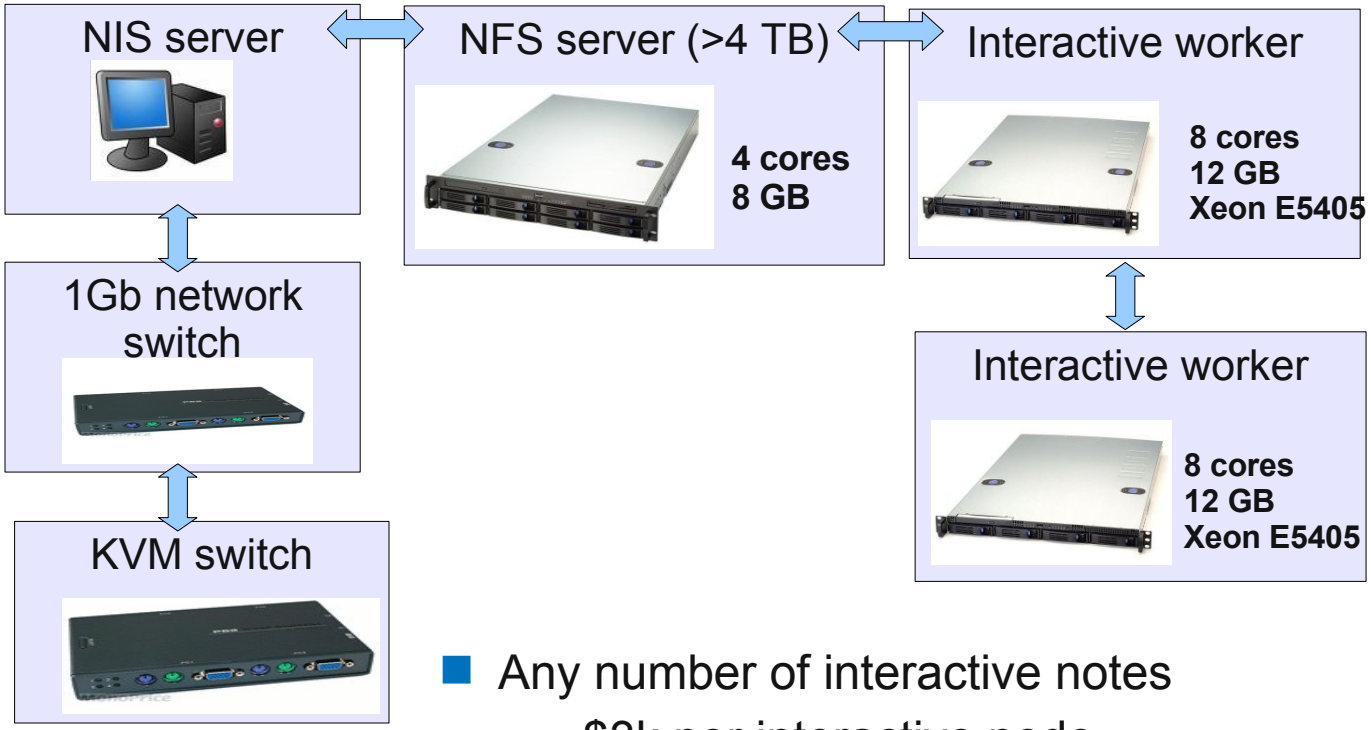
**1) Condor (comes with OSG) on the interactive node**

**2) ArCond front-end (from ANL web page) for data discovery**

Why Condor? Probably the most reliable and best known cluster software (not only in HEP):
See tutorials: Linux.com “Building a Linux cluster on a budget” 2005
LunuxJournal: “Getting Started with Condor”

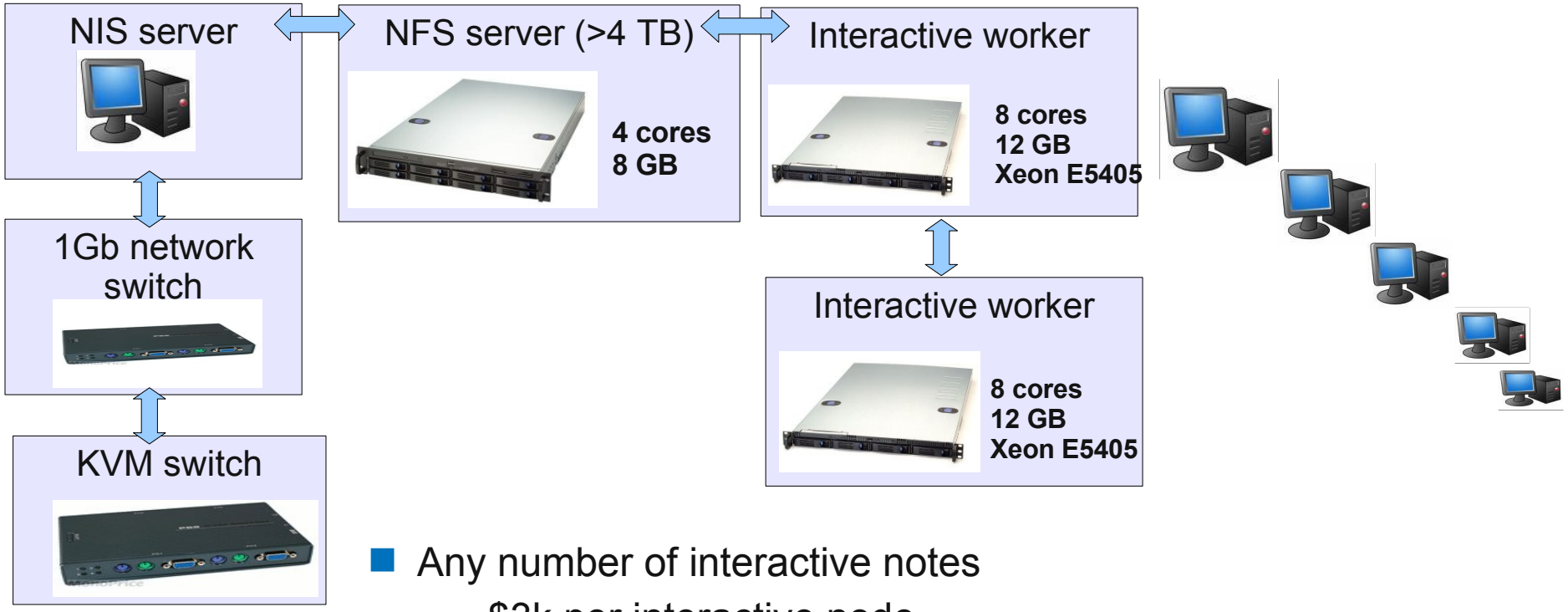
Why ArCond?: Data discovery, splitting input files, submission athena jobs (or any other)

# Expanding to a multi-user system



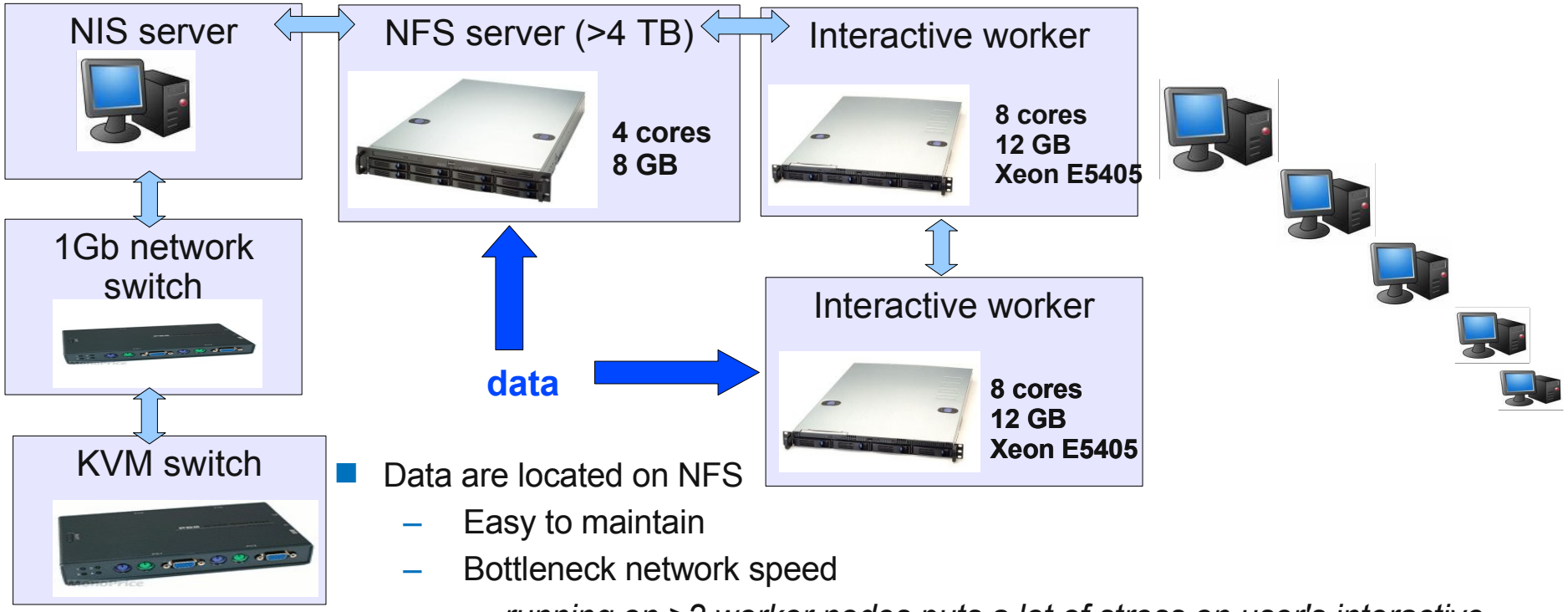

- Any number of interactive notes
  - $3k per interactive node
  - Condor & Arcond to use cores from all interactive nodes
- Any number of desktop PCs for user login
  - $500 per user PC

# Expanding to a multi-user system



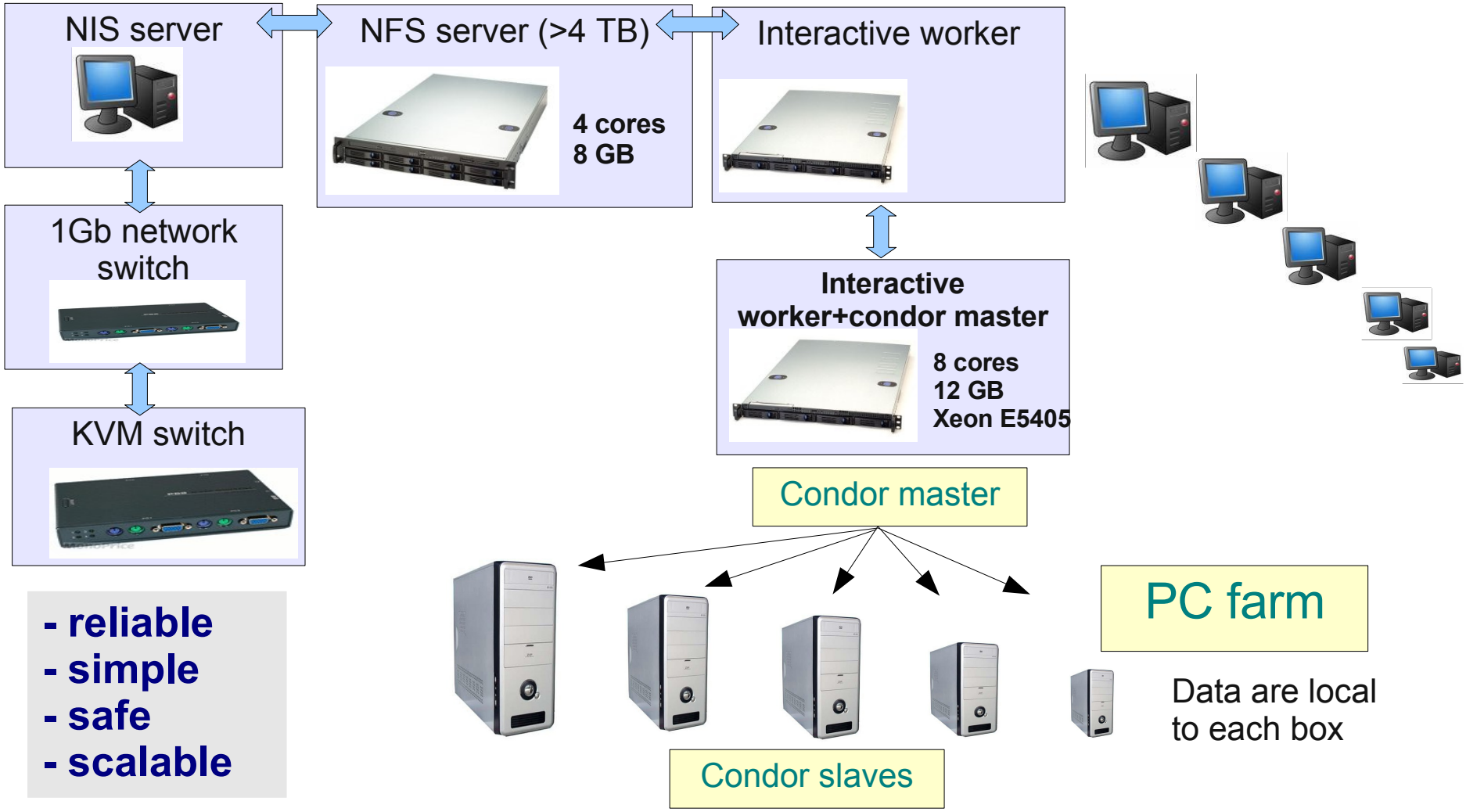

- Any number of interactive notes
  - $3k per interactive node
  - Condor & Arcond to use cores from all interactive nodes
- Any number of desktop PCs for user login
  - $500 per user PC

# Two ways to run over data: NFS vs Local storage




- Data are located on NFS
  - Easy to maintain
  - Bottleneck network speed
    - *running on >2 worker nodes puts a lot of stress on user's interactive applications!*
- Data are located on worker nodes
  - "Distributed" data model
  - ArCond will help to handle submissions
  - Bottleneck: put stress on CPUs for user's interactive jobs

# Interactive worker nodes with a PC farm



# *Tier3 PC farm at ANL*

- **Jobs run 2 orders of magnitude faster compare to standard desktops (2-4 cores)**
  - + takes out load from Tier1-2 by enabling high performance at Tier3
- **Can deal with tens of TBs of data**
- **No resource allocation and file staging for each execution.**
  - **Can be faster than Tier1-Tier2s for multiple runs:**
    - ***3-4 factor faster for ATLAS input files (~100 GB)***
    - ***>10 faster for small jobs (ROOT ntuples)***
- **Better interactivity and full local control of processing of large datasets**
- **Generating large MC sample & CPU-consuming NLO predictions**

**Characteristics:**

- **Cost effective - tens of $k, preferably commodity PCs**
- **Low maintenance – max 0.5 FTE**
- **Scalability**
- **Low network load (assume commodity 1 Gb networking)**
- **Extension of the desktop rather than Tier2**

**ANL PC farm fulfills all these characteristics**

# Hardware configuration for condor slave PC

| | | | | | |
|---|---|---|---|---|---|
| 1 | CUSTOMSERSERIAL | CYBERTRONPC | | CUSTOM CONFIGURED SERVER | 1 |
| 2 | PRC-INT-X5410RA | INTEL | BX80574E5410A | XEON E5410 C4 2.33GHZ 771 RET | 2 |
| 3 | MBD-SPM-X7DVLE | SUPERMICRO | X7DVL-E BULK | 771 V X8 6D2 2GL R DUAL XEON | 1 |
| 4 | MEM-GEN-2FB667 | SUPERTALENT | MEM-SAM-2GEBA | 2GB FB DDR2 ECC FB PC5400/667 MHZ | 4 |
| 5 | | | | TOTAL 8 GB | |
| 6 | HDR-WDG-25AAKS | WESTERN DIGITAL | WD2500AAKS | 250GB S2 7200 16MB | 1 |
| 7 | | | | SYSTEM DISK | |
| 8 | HDR-SGT-1TBS2B | SEAGATE | ST31000340NS | 1TB SATA2 7200RPM 32MB RAID EDITION | 2 |
| 9 | | | | DATA DISKS | |
| 10 | CDR-LIT-16XDVDB | LITEON | DH-16D2P | 16X DVD IDE BLACK | 1 |
| 11 | FLD-ALPS-144MBB | ALPS | DF35 | 1.44 MB FLOPPY DISK DRIVE BLACK | 1 |
| 12 | AD-VID-NOUPGRD | | | ONBOARD VIDEO | 1 |
| 13 | AD-NET6 | | | DUAL 10/100/1000 GIGABIT NETWORK | 1 |
| 14 | SUP-CHN-BRKTNC | CHENBRO | 84H312410-022 | NACONA BRACKET SET OF 2 | 1 |
| 15 | CAS-CHN-SR105 | CHENBRO | SR105-BK(10569-BLACK) | SERVER TOWER BLACK | 1 |
| 16 | POW-SPK-460W | SPARKLE | FSP460-60PFN | 460 WATT INTEL XEON CFT24 PIN | 1 |
| 17 | WARR-EXTENDED1 | | | 1 YEAR WARRANTY ON LABOR & PARTS | 1 |
| 18 | | | | LIFETIME U.S. BASED TECHNICAL SUPPORT | |
| 19 | SHIPPFREE | | | FREE GROUND SHIPPING | 1 |

**CybetronPC quote: $2000 per box (Jan 2009 update)**

Summary:
8 Xeon 2.33 GHz cores
8 GB RAM,
2 TB disks+ 1 system disk

Time to bring to a full operational mode ~ ½ day :
- SL4.6 installation
- starting necessary services (NIS, Condor, etc)
- configure condor home directory + iptables

# *Example performance for AOD*

- **mc08.106070.PythiaZeeJet_Ptcut.recon.AOD.e352_s462_r541**
  - Release 14.2.21
  - 200k events. 800 AOD files. 266/per box, 33GB
  - Lumi=230 pb-1
  - Data equally distributed among 3 PC slaves
- **Program accessing:**
  - Jets,Photons,Muons,Electrons
  - Same for the truth level
  - 100 histograms + fill a ntuple with all objects
- **Processing time: 30 min + 4 min (compilation) on 24 cores** (110 ev/sec, 5 ev. sec).

**10 fb-1 data: ~1 day of running on 24 cores, 6h on 80 cores**
**Data storage: 1.4 TB for data**
**x4 MC = 6-7 TB for MC and data**

If ATLAS release and data located on NFS (ReadyNAS), a low performance due to I/O bottleneck is observed:

- about 10 min to setup ATLAS release (24 cores hit NFS at the same time)
- factor ~2-3 slower during reading AOD events stored on NFS
- poor performance of desktops with NFS-based user home directories

# *PC farm prototype performance for AOD*

- **Estimates for 10 fb-1 assuming 80 cores + 20 TB (skimmed AOD)**
  - Inclusive jets (PT>400-500 GeV)
  - Dijets (PT>200 GeV)
  - Z+jet, PT(jet)>40 GeV
  - .. all other processes with lower x-section (H->gg, etc..)
    - Inclusive direct photon analysis (PT(gamma)>50 GeV, signal ~ background)

  In all cases it is assumed that analysis data set consists of:

  - Data and MC are in form of AODs or DPDs
    - for worst- case scenario when DPDs size = AOD size
  - Monte Carlo samples have 4 times larger statistics than (signal) data

**Estimates for 10 $fb^{-1}$:**

- **did not hit the limit ~10-20 TB for a single analysis**
- **processing time < 1 day for 80 cores in all cases**

# *Analyzing ntuples*

- **200k events from the previous example analyzed using a compiled C++**
- **Ntuple structure and size:**
  - **Storing TLorentzVectors for:**
    - Photons, Muons, Cone4Jets, 10 vectors with doubles (PID for photons)
    - Same for MC truth
  - **Ntuple size: 75MB**

- **Processing 200k events takes 10 sec on one Xeon 2.33 CPU**
  - Filling ~10 histograms with invariant masses (jet-jet, γ-jet, γ-γ)
- **Similar checks where done for Z+jet analysis**

**Estimates for 10 fb-1:**

- requires 3GB file storage
- processing time:
  - 7 min on one core
  - ~ 20 sec on 24 cores
    (assuming no I/O bottleneck)

# *ArCond (Argonne's Condor)*

http://atlaswww.hep.anl.gov/asc/arcond/

- **A Condor front-end:**
  - job submission
  - data discovery
  - checking job status
  - merging outputs
- **Does not require installation & Atlas release**
- **No maintenance or extra service**
  - 1 cron job to build a static database with files (optional!)
- **Minimum requirement: OSG-client (for condor) and standalone ROOT**
- **Designed for analysis of data flatly distributed over multiple PCs**
  - **Example:**

/data1/GammaJet/AOD1.root - 33% of data on atlas1.cern.ch
/data1/GammaJet/AOD2.root - 33% of data on atlas2.cern.ch
/data1/GammaJet/AOD3.root - 33% of data on atlas3.cern.ch

Dataset name (used as a “metadata”)

# 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

1) Data moved to each box after using dq2_get (ArCond provides such splitter).

2) "dq2_get" front-end is ready to get data directly on each box from Tier1/Tier2

| | | |
|---|---|---|
| /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 |

# *Getting data on a PC farm*

- **Data can be copied using using ArCond directly to each PC box:**
  - Calls dq2-ls, splits a list with files, calls dq2-get in parallel
- **Performance: for 1 Gb connection between ANL and UChicago (Tier2):**
  - 15 MB/sec using single-box download
  - 45 MB/sec using 3 parallel downloads directly on 3 PC farm boxes
  - 50 MB/sec using 5 parallel downloads

**~4 TB/day download rate**

# ArCond PC farm submission

- **Pure python & bash. Does not need installation. Requires OSG-client (Condor)**
  - **> arcond**
    - Reads a configuration file (with atlas release version, input directory with AOD files on all boxes, package athena name)
    - Splits jobs to be run in parallel: N=N(PC boxes) x N(cores)
    - Data discovery using local storage. Builds a database with input files and associates each AOD file with specific box
    - Splits data list, prepares shell scripts for submission. Can include:
      - Compilation statement “make from /cmt”
      - Multiple “athena” executions or anything
    - Submits scripts, runs jobs using local condor home directory
    - When jobs are ready, outputs are copied to the user submission directory
      - optional, depends what do you put in shell script
    - Output root files merged automatically (arc_add command)

# Running arcond

- Before submitting a job, prepare a configuration file (“ arcond.conf”)

```
atlas_release=14.5.1

# events to process in each job
events = -1

# dir with input AOD files.
input_data =
/data1/mc/mc08.105802.JF17_pythia_jet_filter.recon.AOD.e347_s462_r541/AOD

# package directory on NFS
package_dir = /users/chakanau/testarea/14.2.21/analysis/PromptGamma
```


- Check data availability as:
  - `arc_ls <dataset>`

Ready to submit!

```
chakanau@atlas16:submit$ ./arcond
################## ARCOND v1.0 ######################
##                ANL ASC                  ##
#####################################################
 Input configuration=arcond.conf
---> Input data located at =
/data1/mc/mc08.105802.JF17_pythia_jet_filter.recon.AOD.e347_s462_r541/AOD
---> Checking computing cores
     -->1 PC node=atlas51.hep.anl.gov with=8 cores found
     -->2 PC node=atlas52.hep.anl.gov with=8 cores found
     -->3 PC node=atlas53.hep.anl.gov with=8 cores found
---> Total number of found cores= 24
Start data ArCond data discovery tool?
-> To discover data on-fly, type "f"
-> To discover data using ArCond static database created every 24h
-> Do not discover data, say "n"
---> Checking claimed CPUs
---> Total number of claimed CPU cores= 0
---> Building the database on all nodes with input AOD/DPD files
---> Checking for duplicate input data files
    --> PC node= atlas53.hep.anl.gov   has  1987  input files
    --> PC node= atlas51.hep.anl.gov   has  1964  input files
    --> PC node= atlas52.hep.anl.gov   has  1722  input files
    --> ## SUMMARY: Total number of input files = 5673
Project file:/users/chakanau/work/submit/Job/PromptGamma.tgz was found.
Do you want to rebuild it (y/n)? y
---> Package submission file = Job/PromptGamma.tgz
---> Package submission log file = Job/PromptGamma.log
---> Number of events in one job = -1
---> Atlas release = 14.5.1
---> 24  jobs will be submitted to = 3 PC boxes
Do you want to prepare the submission scripts (y/n)? y
Submit all prepared jobs to the PC farm? (y/n)
```

only for first submission! (see next slide)

it was found since I've sent this package before

To run ArCond in silent mode use: “arcond -allyes”

# Data discovery

**PC farm users have several choices for data discovery:**

- **"s" - to discover data using a small flat-file database**
  - Updated every night
  - Implementation: Each slave note runs a cron job
    - (based on **find "/data1/ -type f > /users/condor/$date.txt")**
    - for 10000 AOD files, run time is 3-5 sec.
  - Copied and stored on NFS
  - When a user runs "./arcond", always the latest database is used
  - Also can be used to recover data when PC box fails (do not have experience yet)
- **"f" - to discover data "on-fly"**
  - If data have been copied recently, the database may not exists
  - Then arcond sends a small script on each PC boxes and brings data list back
  - Usually takes ~20-30 sec (assuming that Condor is not busy)
- **"n" if the user selected "s" and "f" from previous runs, there is no need to discover data (previous data list will be used)**

**Simple and robust. So far required no attention from admin.**

# Checking and getting jobs back

- Run condor commands: condor_status or condor_q
- Your jobs are in “idle” state?
  - check who is running on the farm as:
    - condor_status -submitters (OR) condor_q -global
- Check output files as: arc_check
- If arc_check tells that all output files “Analysis.root” are ready, combine output files to one file using arc_add. This creates “Analysis_all.root”
- To debug program and check errors:
  - ./Job/runXXX/Analysis.log - athena log file
  - ./Job/runN_atlasXXX/Job.ShellScript.atlasXXX/job.local.out - Condor log file

# Summary

- **24-CPU PC farm prototype is fully functional**
  - ~$6k investment last year
  - Man power: 0.3 FTE, which dropped to 0.1 FTE after the setup
- **Since Sep 1, ~5000 jobs completed**
  - ~ 200x24-core completed jobs.
  - Most of ANL results were done using the PC farm prototype
- **No any failures reported:**
  - Small problem if Condor master is busy (at present runs on the worker PC)
    - 1-2 cores are not identified correctly by Condor → lower efficiency
    - a dedicated Condor master should be installed
- **With extra $14k investment, the PC farm could be extended to ~T3g**
    - **Goal:** 80 CPUs with 20 TB data storage

**Tutorials: How to use PC farm for athena and ROOT-ntuple type of jobs:**

**~chakanau/public/2009_jamb_may/15.1.0/Tutorials/arcond_athena**
**~chakanau/public/2009_jamb_may/15.1.0/Tutorials/arcond_ntuple**