Distributed analysis with PROOF in ATLAS

Sergey Panitkin, Michael Ernst, Hironori Ito, Tadashi Maeno, Stephanie Majewski, Ofer Rind, Torre Wenaus, Shuwei Ye (Brookhaven National Lab), Doug Benjamin (Duke), Kyle Cranmer, Fabien Tarrade, Akira Shibata (NYU), German Carrillo, Wen Guan, Bruce Mellado, Neng Xu (UWM)

CHEP09
Prague, Czech Republic
March 24, 2009

BROOKHAVEN
NATIONAL LABORATORY

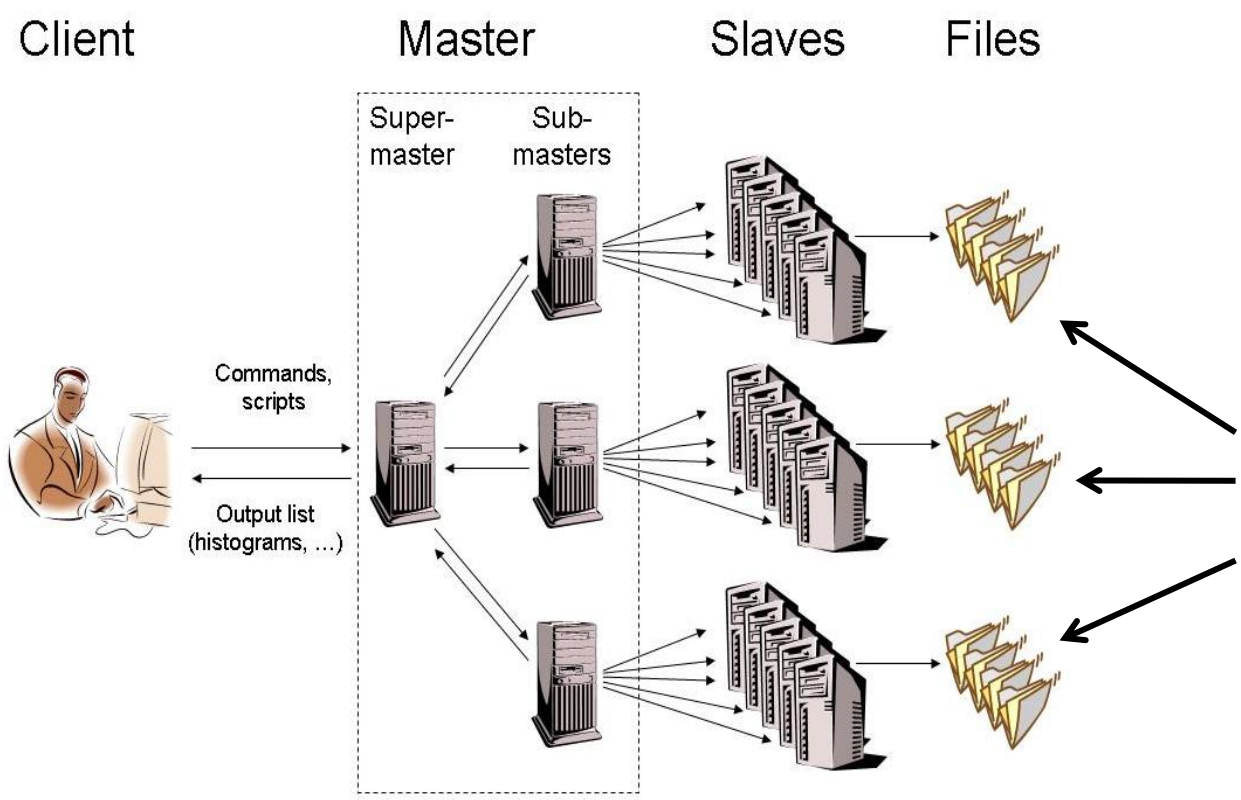
Introduction

- Large datasets will be a basic feature of Atlas physics analysis
  - Expect ~2x10E9 events per year
- Most of the analysis will be distributed
  - Large fraction of analysis is expected to be done on Atlas Grid –T1, T2 centers
  - Some (growing) number of T3 centers are envisaged, to facilitate end user analysis
- Most of Atlas analysis data (ESD, AOD, D1PD, D2PD) are written in POOL/**ROOT** files
- D^3PD will be written as plain ROOT trees
- AthenaRoot Access (ARA) provides tools for accessing POOL root data – AOD, DPDs - directly in ROOT, without Athena framework
- Large fraction of user analysis is expected to be done in ROOT
- How to analyze ~10E9 DPD events efficiently in ROOT, on Tier 3?
- **Use PROOF!**

Introduction to PROOF

- The Parallel ROOt Facility -PROOF -is ROOT's extension for parallel data processing
- Integral part of ROOT framework. Distributed with ROOT. Supported by the ROOT team
- Speed up the query processing by employing inherent parallelism in event data
- Allows interactive and batch analysis modes
- Allows access to remote distributed data from ROOT prompt
- Can efficiently run on commodity, heterogeneous hardware.
- Well suited for distributed local storage model
- Scales well from a dual core laptop to clusters with hundreds of nodes
- Can federate geographically distributed farms
- Used by several experiments at LHC and elsewhere

PROOF Architecture



Adapts to wide area *virtual* clusters

Geographically separated domains, heterogeneous machines

Super master is users' single point of entry. System complexity is hidden
Automatic data discovery and job matching with local data
Can be optimize for data locality or high bandwidth data server access

PROOF sites in Atlas

- University of Wisconsin, Madison
  - 200 cores, 100 TB, RAID5
  - Data analysis (Higgs searches)
  - I/O perfomance tests w/ multi-RAID, integration with Atlas DDM
  - PROOF-Condor integration, Analysis Facility prototype
  - ~20 registered users
- Brookhaven National Lab
  - Prod. :40 cores, 20 TB HDD, Test: 72 cores, 25 TB HDD, 192 GB SSD
  - Data analysis, I/O perfomance tests with SSD, RAID, DDM development
  - ~25 registered users
- Munich LMU/LRZ
  - 10 AMD Dual CPU / dual Core Processors 2.7 GHz, 8 GB RAM
  - Data analysis, I/O performance and scalability tests
- PROOF test farms in Madrid, UT Arlington, Duke University

Atlas PROOF farm at BNL

- Atlas PROOF test farm was set up at BNL about a year and a half ago
- Part of Atlas Computing Facility (ACF) at BNL. Co-located with Atlas T1.
- Resides inside ACF T1 security perimeter.
- Accessible from ACF's interactive and batch nodes
- Connected to dCache via Xrotd dooor on dCache
- Ganglia monitoring page: http://www.atlasgrid.bnl.gov/ganglia/?c=ATLAS%20Xrootd%20Testbed
- Xrdmon monitoring page: https://network.racf.bnl.gov/xrdmon

Current BNL Farm Configuration

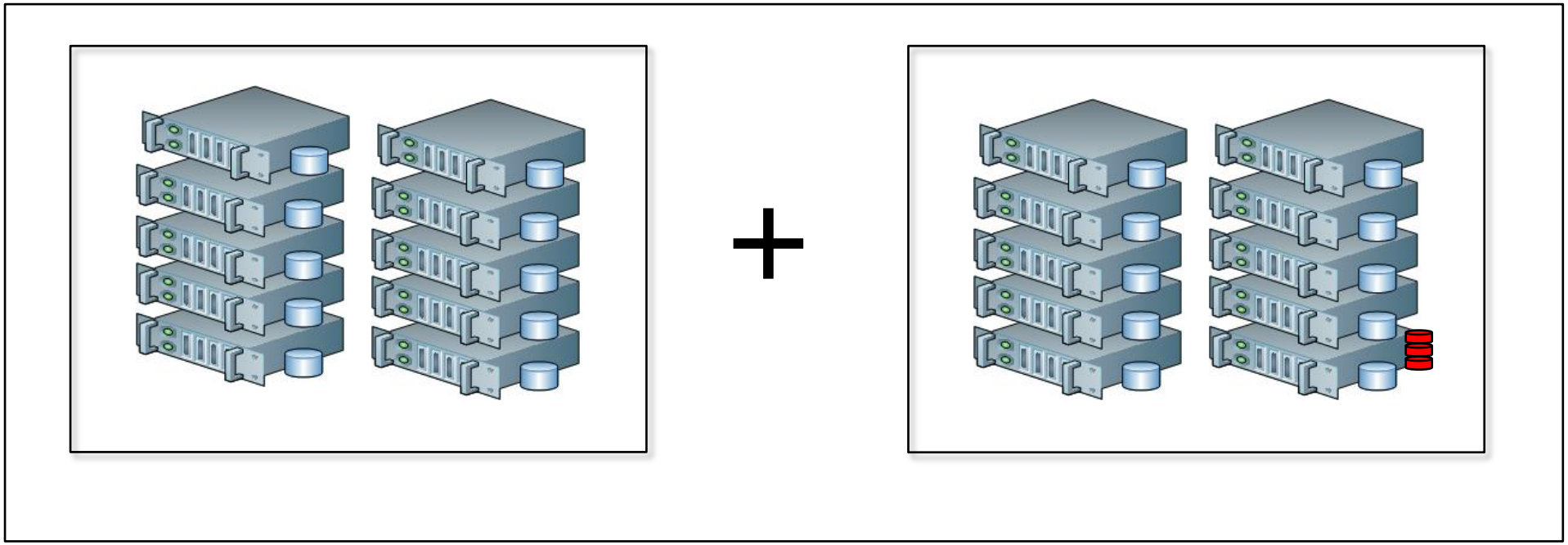
"Old farm, acas0420" –open for users

- 10 nodes – 8 GB RAM each
- 40 cores: 1.8 GHz Opterons
- 20 TB of HDD space (10x4x500 GB)
- Root v5.21 for rel. 14 compatibility
- ARA test, FDR 1 and 2 data analysis

Extension, acas0601 – test mode

- 9 nodes - 16 GB RAM each
- 9x8 cores: 2.0 GHz Kentsfields
- 25 TB of HDD space (9x2.25 TB)
- Root v5.20, ARA tests
- +1 node for SSDs testing (3 SSDs)

As of March, 2009



Data flow at BNL PROOF farm

Atlas user

dq2_ls –fp –s BNLXRDHDD1 “my_dataset”

analysis

acas0420.usatlas.bnl.gov

/data

BNLXRDHDD1

Xrootd/PROOF Farm

DQ2

T0

Panda

Grid transfers

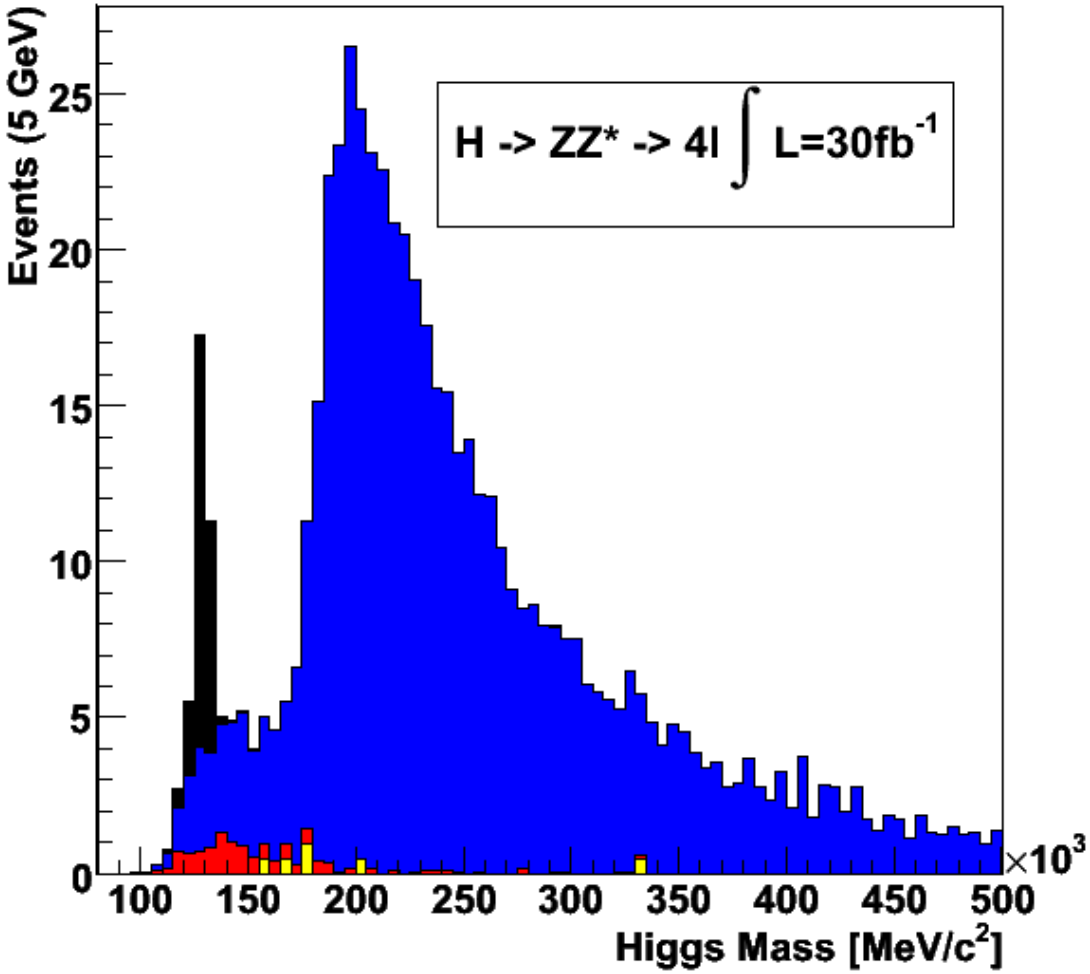
dCache

xrdcp with dq2 registration

PROOF in Atlas Analysis

Summary plot of H->4l analysis done by Wisconsin group on UWM PROOF cluster

German Carrillo, Bruce Mellado

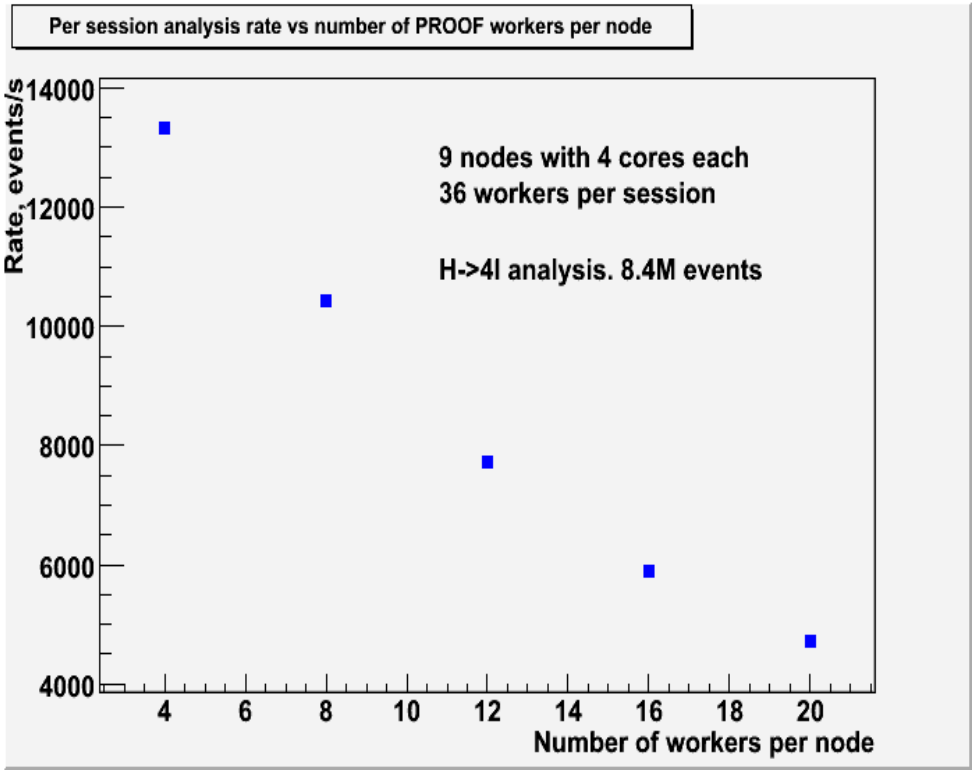


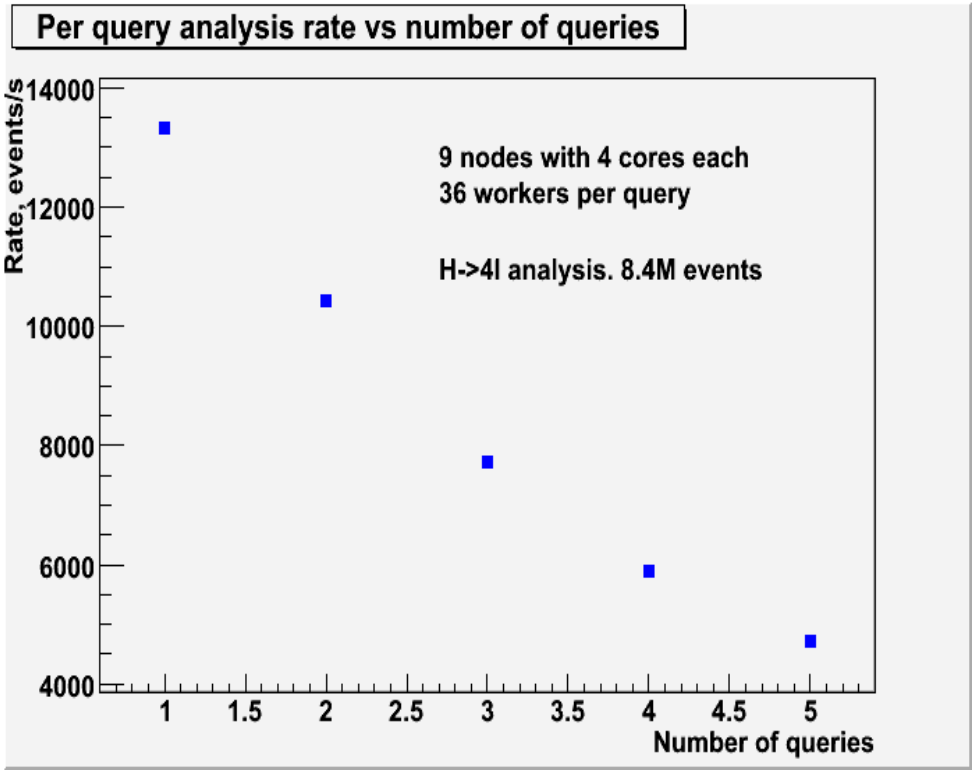
4.5M simulated events
~ 68GB of data

CPU and I/O intensive analysis

Used for PROOF tests at UWM and BNL

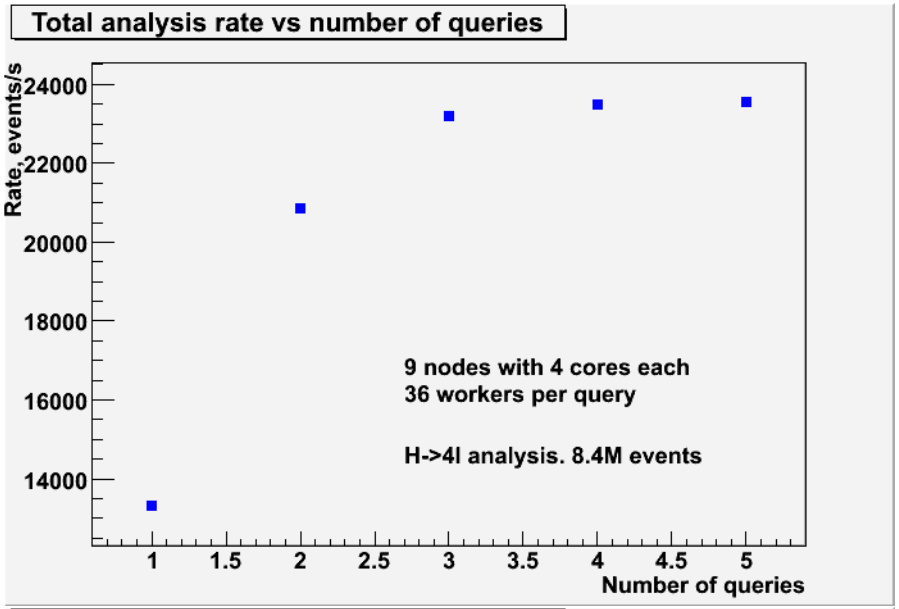
Multisession performance

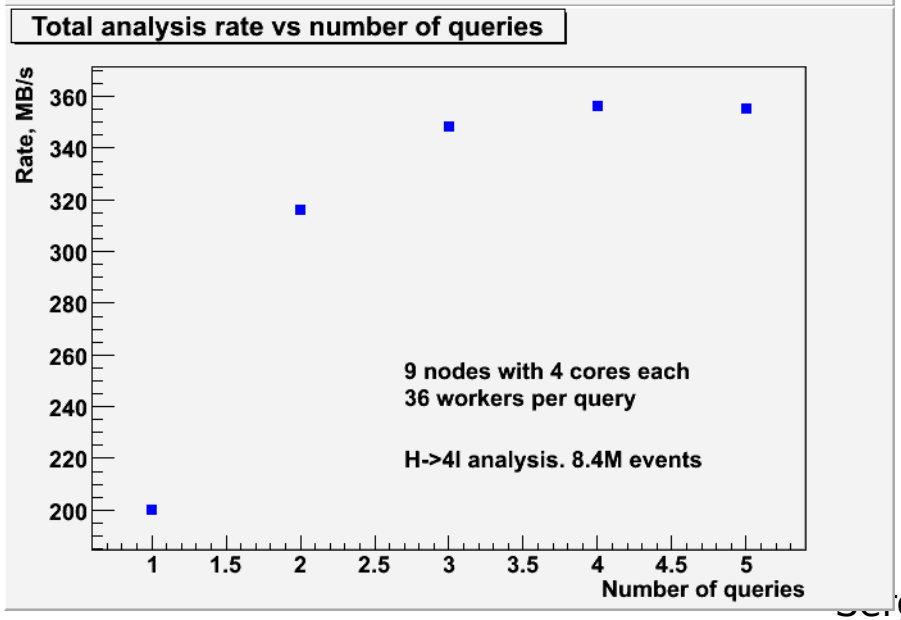


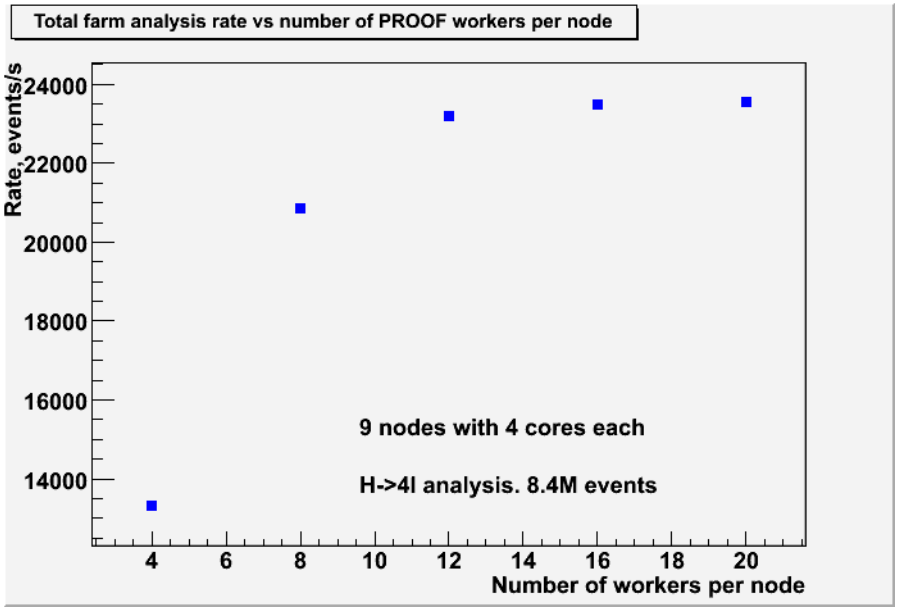


As expected per query performance drops as number of queries increases.
Resource sharing between jobs with equal priorities.

Analysis rate scaling







- Aggregate analysis rate saturates at about 3 (full load) queries
- Max analysis rate is about 360 MB/s for a given analysis type

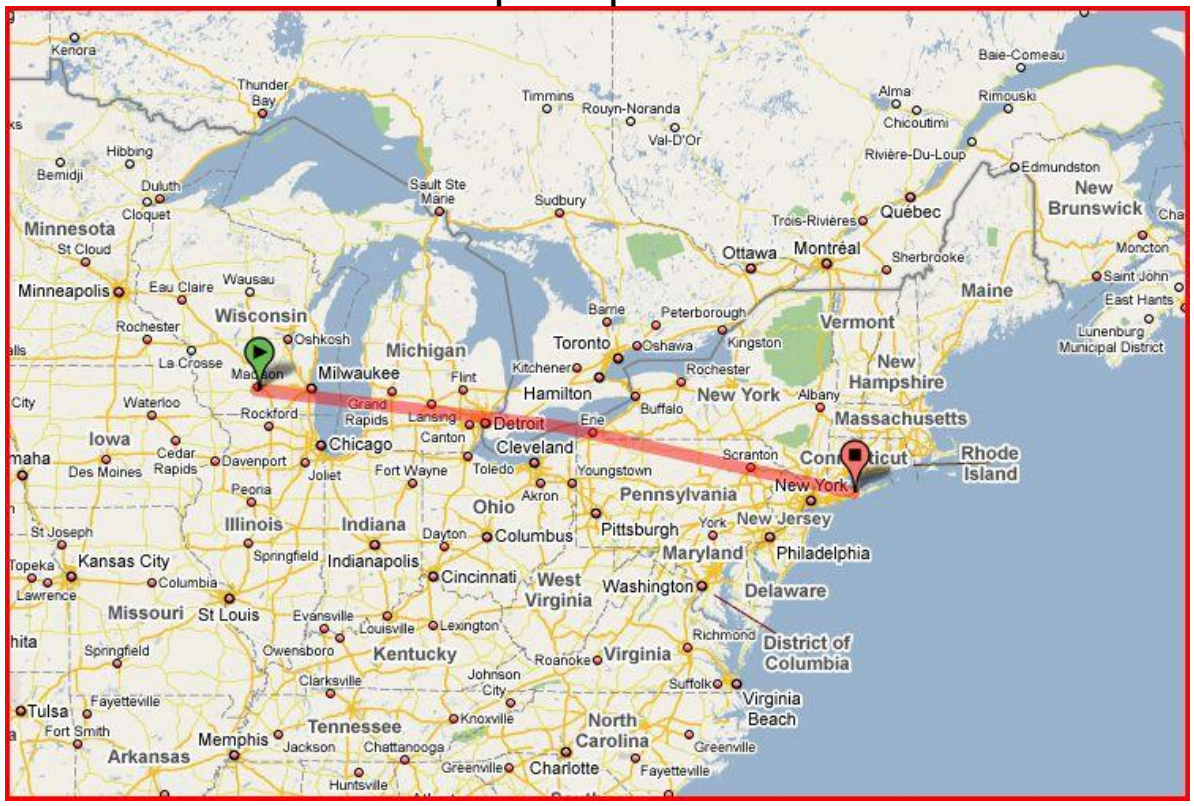
It makes sense to run PROOF farm at optimal number of queries

PROOF Cluster federation tests

- In principle Xrootd/PROOF design supports federation of geographically distributed clusters
  - Setup instructions at: http://root.cern.ch/twiki/bin/view/ROOT/XpdMultiMaster
- Interesting capability for T3 applications
  - Pool together distributed local resources: disk, CPU
    - *Collaborative dataset storage!*
  - Relatively easy to implement:
    - Requires only configuration files changes
    - Can have different configurations
  - Transient partnerships, with changing partners, depending on task
  - Transparent for users
    - Single “name space” – may require some planning
    - Single entry point for analysis
- New and untested capability

BNL-Wisconsin PROOF Federation

Proof of principle tests



G. Ganis
S. Panitkin
N. Xu

- 3 local PROOF clusters, 2 at BNL and 1 in Wisconsin were successfully federated into one “super-cluster”.
- Main issue – how to deal with firewalls at BNL .
- Current solution - ssh tunnels. Dual homed master nodes may be a better solution – investigating.
- Tests are underway

Support and documentation

- Main PROOF Page at CERN, PROOF worldwide forum
  - http://root.cern.ch/twiki/bin/view/ROOT/PROOF
- USAtlas Wiki PROOF page
  - http://www.usatlas.bnl.gov/twiki/bin/view/ProofXrootd/WebHome
- Web page/TWIKI at BNL with general farm information, help, examples, tips, talks, links to Ganglia page, etc.
  - http://www.usatlas.bnl.gov/twiki/bin/view/AtlasSoftware/ProofTestBed
- Hypernews forum for Atlas PROOF users created:

  hn-atlas-proof-xrootd@cern.ch

  https://hypernews.cern.ch/HyperNews/Atlas/get/proofXrootd.html
- Several PROOF tutorials were given at Atlas Analysis Jamborees
  - First PROOFLight tutorial was given in December 08

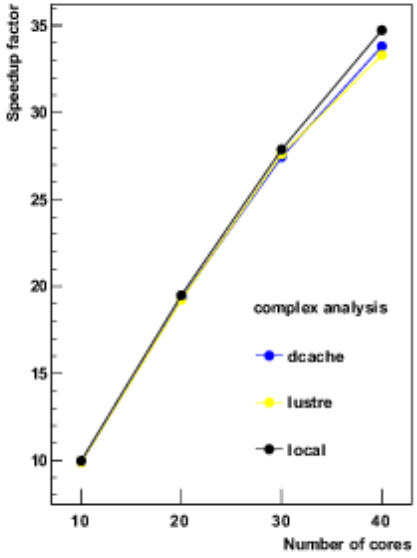
PROOF tests in LRZ Munich

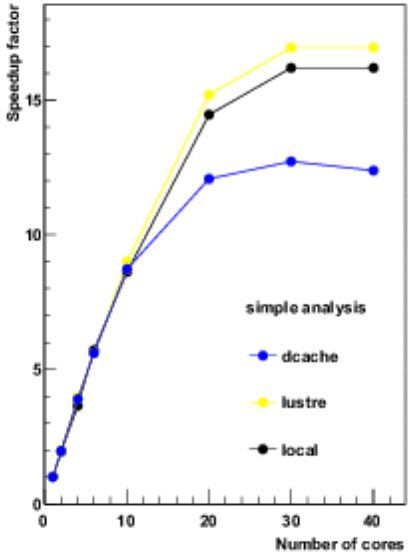
Comparison of storage strategies (P. Calfayan, poster for CHEP09)

Tests are carried out at LRZ Munich. A PROOF cluster of 10 Opteron nodes with 4 cores and 8Gb RAM per node is utilized.

- → **local** disks: data is stored on each local node
- → **dcache**: data access via client/server connections, RAID6, 10GB switch
- → **lustre**: filesystem optimized for parallel computing, all nodes can access the data without a dedicated server

- Test analyis: Z boson reconstruction plus control histograms.
  Complex variant includes 200000 tanh operations per event.
- Input data files: D^3PD, 1.6 million of events, nearly 4kB per event.
- Using ROOT v5.20





PROOF tests at LRZ Munich

Performance of PROOF with pool files (P. Calfayan, poster for CHEP09)

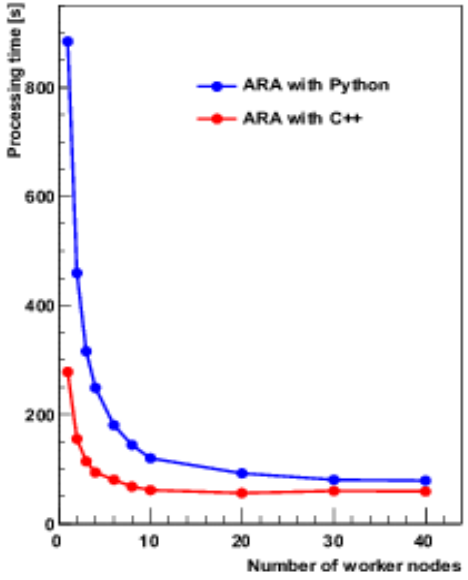
(using the same setup as for the comparison of storage strategies)

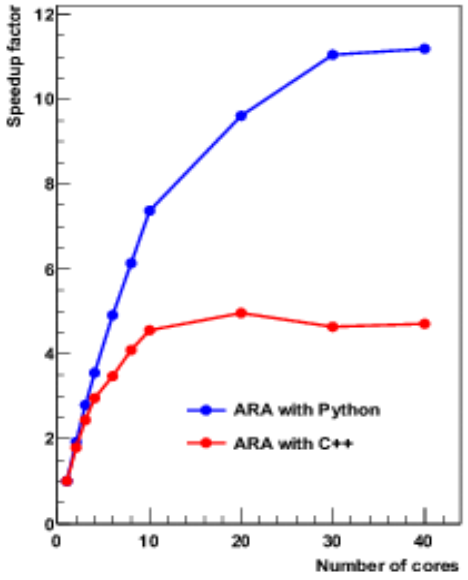
→ We use AthenaROOTAccess to read AODs (persistant tree → ROOT transient tree).

→ Processing AOD pool input files with PROOF and a compiled C++ analysis is not possible with CINT dictionaries, because of CINT limitations.

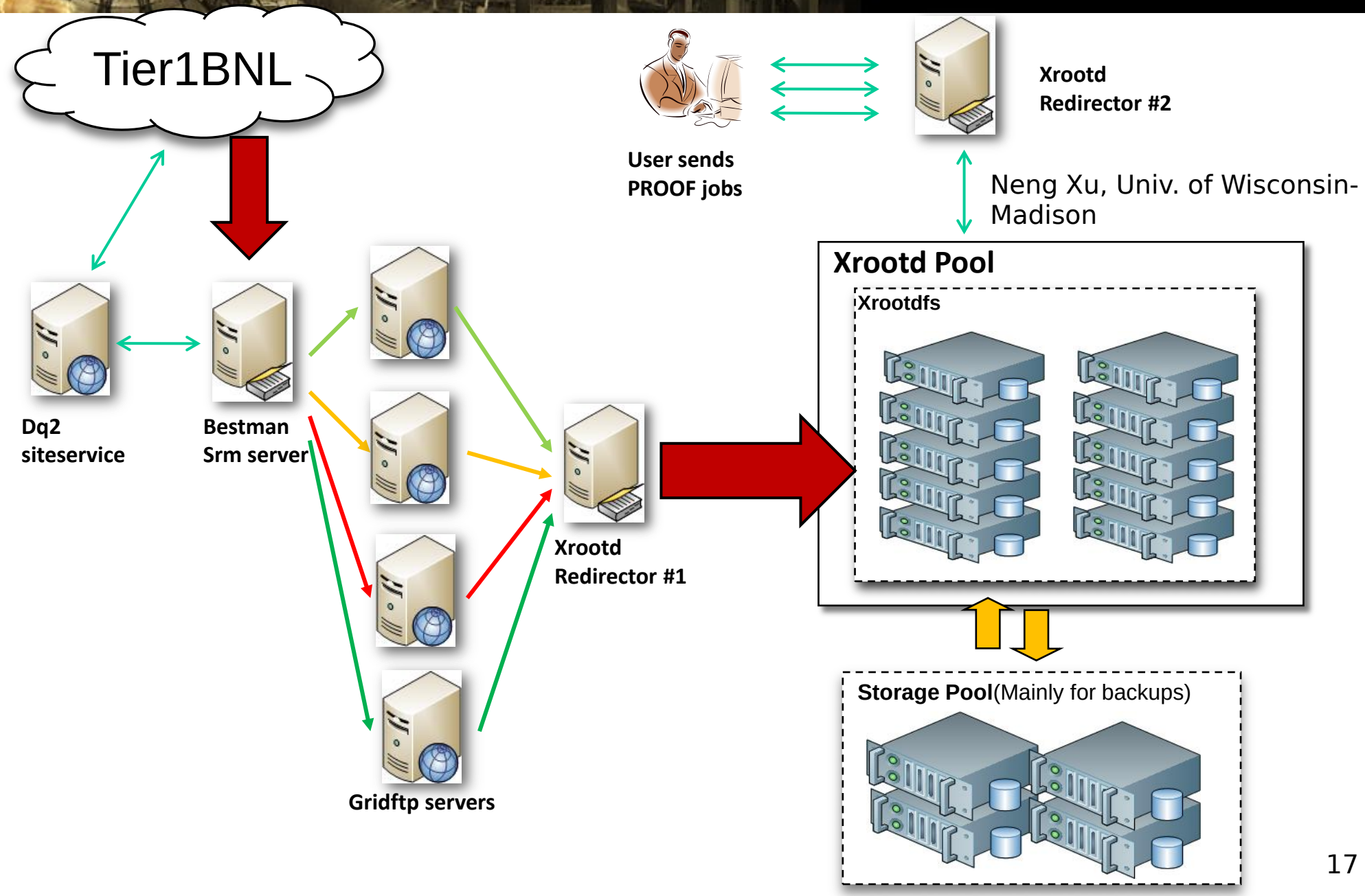
→ We compile the analysis loop (TSelector) in a CMT package with Athena 14.2.23, and use a REFLEX dictionary.

- Transient tree read in 2 ways: compiled C++ or Python (via TPython in a compiled TSelector).
- Test analyis: W transverse mass calculated 10k times plus control histograms, nearly 12500 $W \to \mu\nu$ events (10 TeV, Athena 14.2.20), files stored with Lustre.

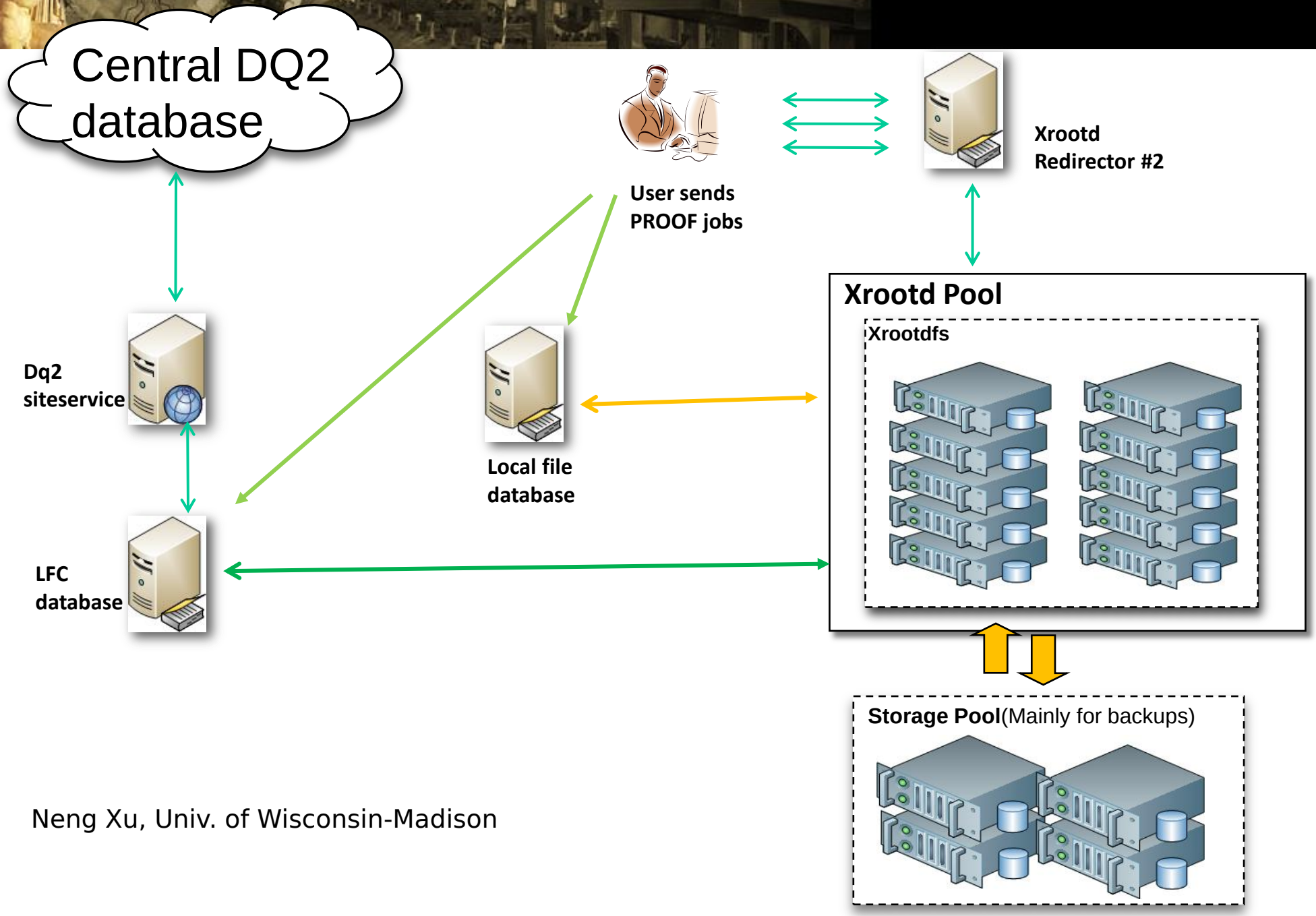




DDM system for UWM T3 PROOF farm



File registration

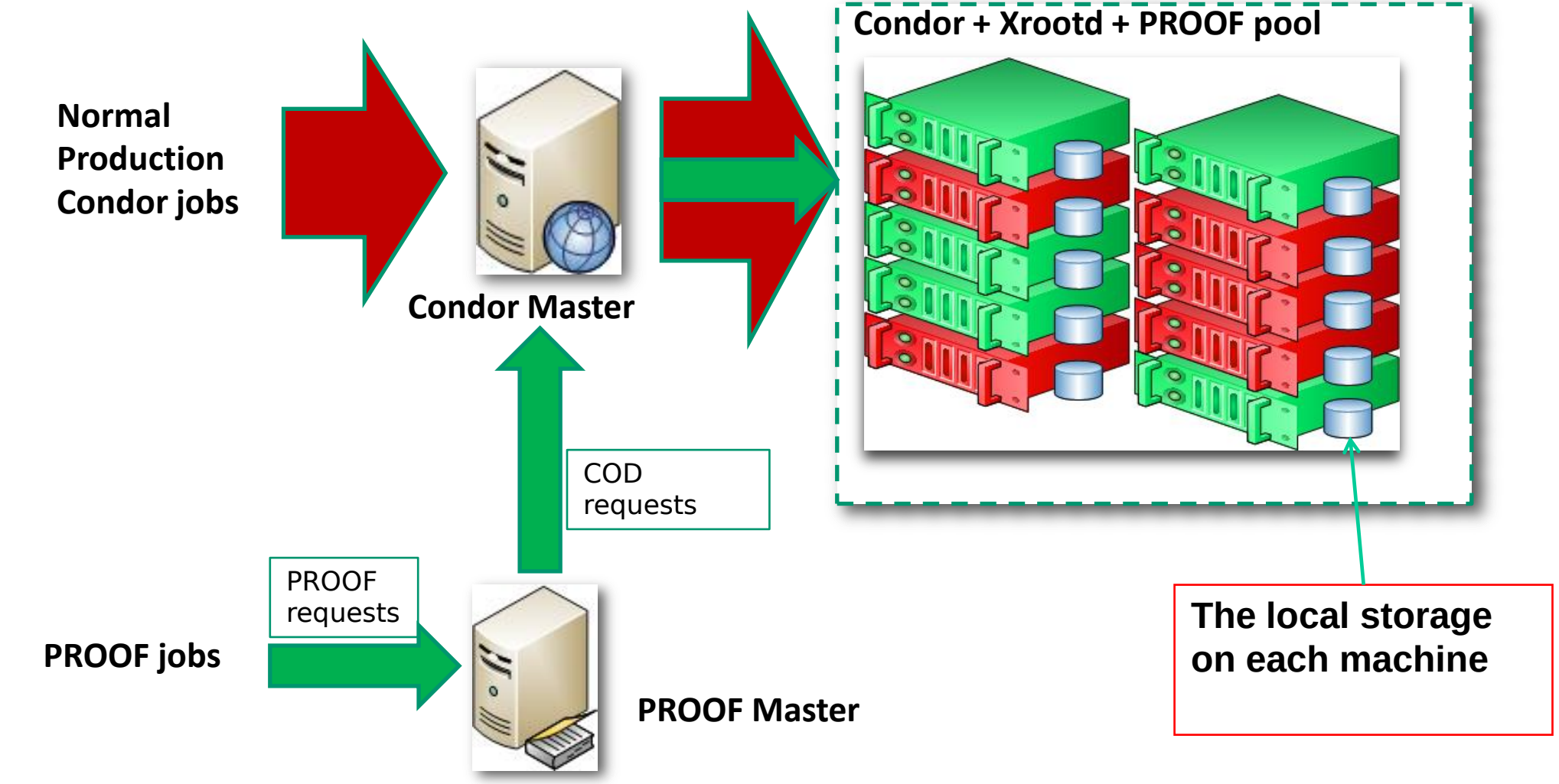


Neng Xu, Univ. of Wisconsin-Madison

Multipurpose Tier 3 site model

- Combination of PROOF pool and batch pool seems to be a good solution for small T3 sites
- No empty CPU cycles.
- Production/batch jobs won’t be affected by PROOF.
- PROOF jobs get immediate CPU resources.
- Transparent to PROOF and batch users.
- UWM pioneered this approach in Atlas .

The basic PROOF+COD Model



Summary

- PROOF/Xrootd is an attractive technology for Atlas, especially for T3 centres
- Several PROOF test farms are operational in Atlas (BNL, Madrid, Munich, Wisconsin)
- Significant experience with PROOF was gained
  - Several Atlas analysis scenarios were tested, with good results
  - AthenaRootAccess was shown to work on PROOF. Used during FDR1 and FDR2
  - Improved integration with Atlas DDM was demonstrated
  - PROOF farms are used for analysis by many Atlas physicists
  - Working prototypes/examples of farm management and monitoring setup exist.
  - Federation of geographically distributed PROOF clusters was demonstrated
  - Several bugs in PROOF/Xrootd were discovered, reported to developers and fixed
- Wiki pages is available for Atlas PROOF users with examples, etc
- Several PROOF tutorials were given at Atlas Analysis Jamborees
- Integration with Condor is being actively explored by Atlas Wisconsin group.