Atlas Tier 3 Virtualization Project

Doug Benjamin
Duke University

What is a Tier3?

- Working definition
  - "Non pledged resources"
  - "Analysis facilities" at a University/Institute/...
- Tier 3 level
  - The name suggests that it is another layer continuing the hierarchy after Tier0, Tier1s, Tier2s...
    - Probably truly misleading...
    - Qualitative difference here:
      - **Final analysis vs simulation and reconstruction**
      - **Local control vs ATLAS central control**
      - **Operation load more on local resources (i.e. people) than on the central team (i.e. other people)**

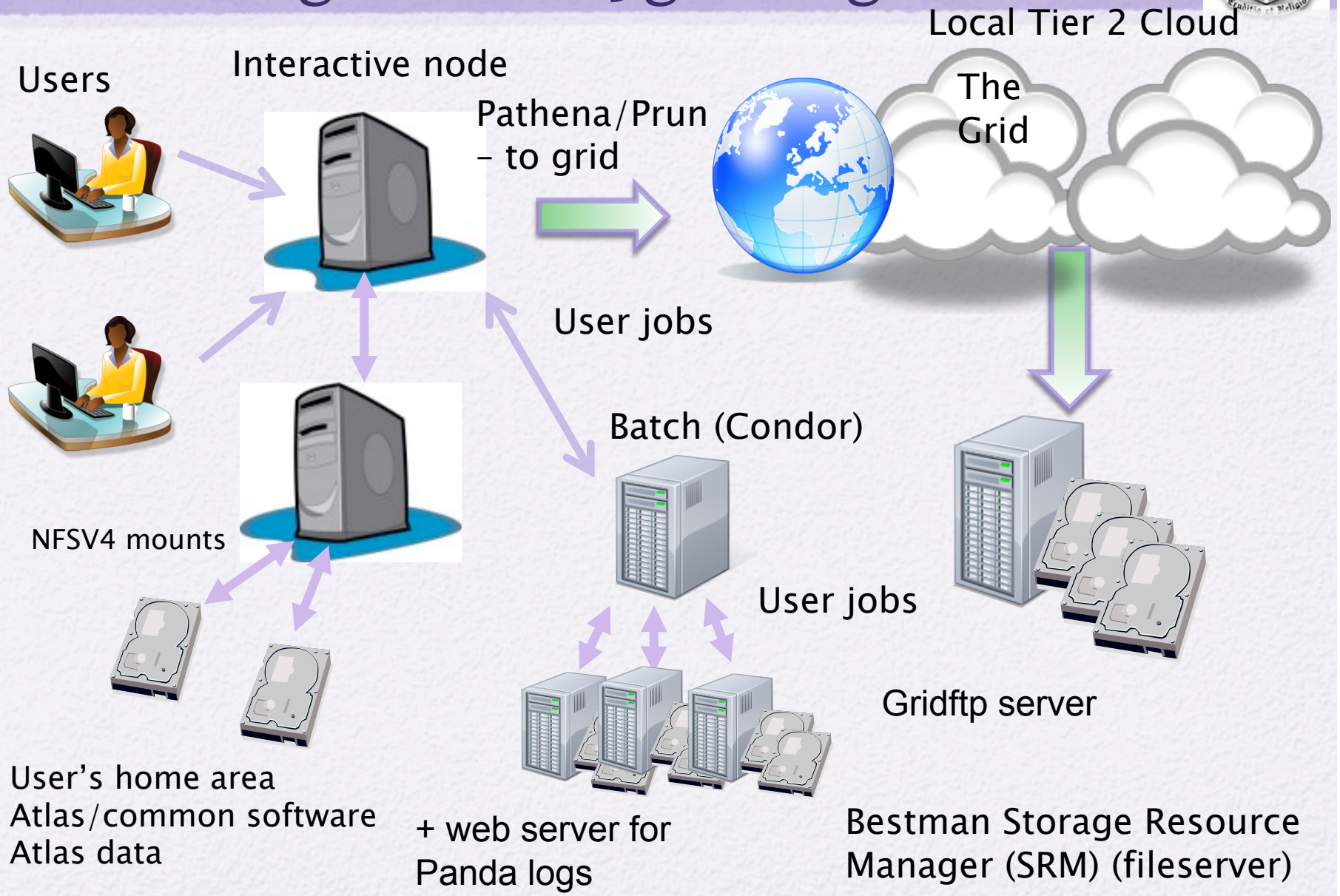
What is a Tier3?

- Comments:
  - No concept of size (small Tier3 vs big Tier2...)
  - Tier3s can serve (and be controlled by) a subset of the ATLAS collaboration (local or regional users).
- *Non-pledged* resources does not mean *uncontrolled or incoherent*
  - Need to provide a **coherent** model (across ATLAS)
    - Small set of template to be followed while setting up a Tier3 for ATLAS users.
  - Coherent because:
    - Guarantee no negative repercussions on the ATLAS Grid (service overload, additional complex manual support load) by the proliferation of these sites

Tier3: interesting features

- Key characteristics (issues, interesting problems)
  - Operations
    - Must be simple (for the local team) (**< 0.25 FTE** *Physicist*)
    - Must not affect the rest of the system (hence central operations)
  - Data management
    - Again simplicity
    - Different access pattern (analysis)
      - I/O bound, iterative/interactive
      - More ROOT-based analysis (PROOF?)
      - Truly local usage
    - "Performances"
      - Reliability (successful jobs / total)
      - Efficiency (CPU/elapsed) → events read per second

Non grid - Tier 3g configuration



Non-grid site Tier 3 design/Philosophy

- Design a system to be flexible and simple to setup (1 person < 1 week)
- Simple to operate - < 0.25 FTE to maintain
- Scalable with Data volumes
- Fast - Process 1 TB of data over night
- Relatively inexpensive
  - Run only the needed services/process
  - Devote most resources to CPU's and Disk
- Using common tools will make it easier for all
  - Easier to develop a self supporting community.

ATLAS Tier 3

- Currently in EU – Tier 2/3 combined sites
- Most US Atlas institutions received funds from the funding agencies to support Tier 3 computing at their home institutions.
- The funds are set to flow shortly (1-2 months).
- Expect to see ~30 new (or greatly enhanced) Tier 3 sites in the US before the end of the year.
- Amost all of new Atlas sites will be a non-grid Tier 3

Transformative technologies

- By their operational requirements non-grid Tier 3 sites will require transformative ideas and solutions
- CernVM FS (*needs long term support*)
  - Minimize effort for Atlas software releases
  - Conditions DB
- Xrootd
  - Allows for straight forward storage agregation
  - Wide area data clustering will helps groups during analysis (couple xrootd cluster of desktops at CERN with home institution xrootd cluster)

Transformative technologies(2)

- Robust client tool to fetch experiment data (Dq2-get with fts data transfer) (no SRM required – *simplification*)
- Proof
  - Efficient data analysis
  - Tools can be used for data management at Tier 3
- Virtualization / cloud computing
  - Sharing of physical resources
  - Better isolation of critical services
  - Straight forward standardization (easier support)

Virtualization in Tier 3

- Initial Virtualization occurred during the design phase
  - Created a "Tier 3 in box"
  - Intel i7-920 (4 core w/ HT) 12 GB
  - Contained Xen VM's - batch head node, interactive nodes, squid, nfs file server , 2 worker nodes
  - Used to prototyping the configuration
- Virtualization in the current design (production)
  - head node services
    - Squid Cache, ldap server , Condor Collector/ Negotiator , Puppet Server

Virtualization of services

- Atlas Tier 3 head nodes contains several services
  - Squid Proxy Cache, Ldap Server, Condor Collector/Negotiator, Puppet Master, Xrootd Redirector, Proof Master
- Each service does not need its own physical machine
- Initially used XEN VM's (SL 5.3)
- Switched to KVM with SL 5.5
  - (due to RH long term support of KVM)
  - Talking with Predrag to use specialized CernVM's
- Not all services work well in VM's
  - Xrootd redirector had erratic behaviour in XEN VM

What about worker nodes?

- Some Atlas Tier 3 sites have virtual machines running on worker nodes
  - Need virtualization to run old OS (SL 4) on newer hardware (Intel i7)
- Conflicting performance results prevent wider adoption
  - Need to better understand Cost/Benefit for worker node virtualization
  - Need better tests that simulate analysis load in Tier 3's

PROOF-Lite tests details

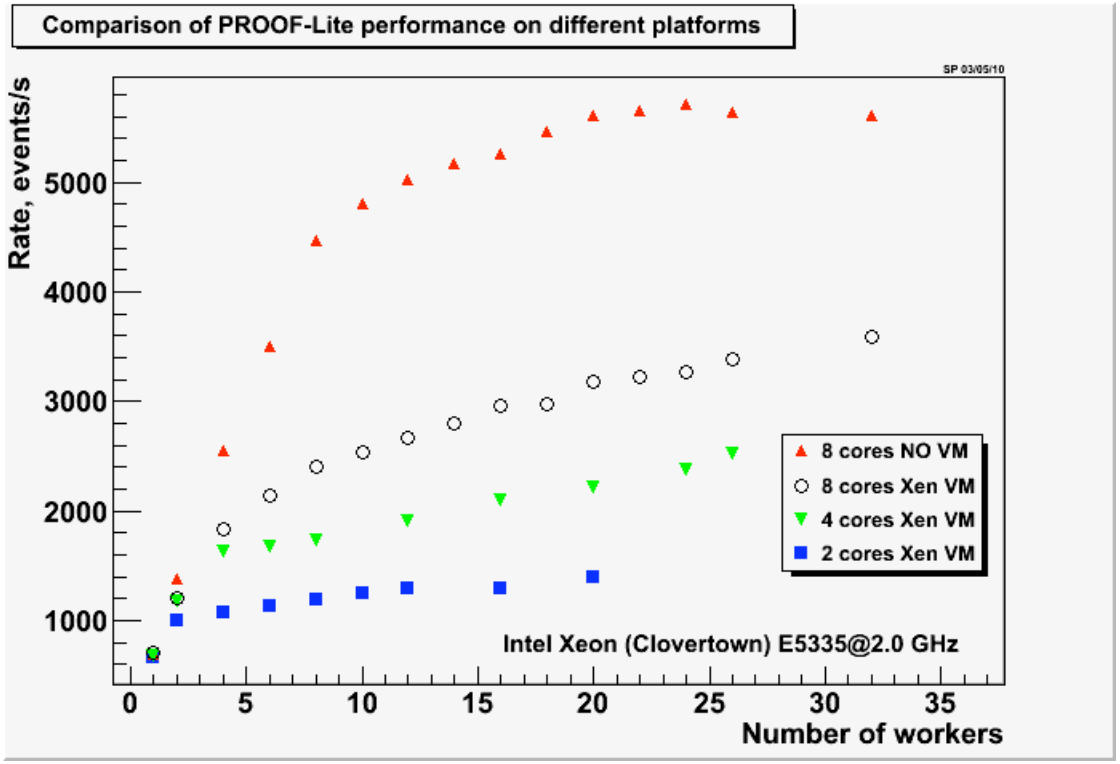
- Proof-Lite was tested on different virtual and physical machines
- VM tests were done mostly on acas0010 interactive login node
  - It has Intel Xeon CPU E5335 running at 2 GHz
  - with 8 GB of RAM and 1GB network interface
  - VM was configured with 4 and 8 cores for comparisons
  - It runs 2.6.18-164.2.1.el5xen x86_64 kernel

- The test analysis jobs were run on 900 GeV data.
- (DATA900_MinBias_esd_r998_V17)
- 208 files, ESD on D3PD format
- 1103478 events, ~19 GB in size (19,254 MB), ~17.4 kB/event

- Files were located on one of the ACF's high-performance NFS server (/usatlas/groups/top), Model BlueArc ……..
- Test jobs were examples of Atlas min-bias analysis code

- For OS level performance monitoring we used Ganglia setup at ACF

PROOF-Lite on different platforms



It looks like PROOF-Lite running in a VM suffers severe performance penalty
~x2 worse performance than on “bare metal”

Tier 3 - Virtualization Future

- Tier 3 networking configured so that the worker nodes are on a private network
  - Private network space very large – can accommodate many VM per physical box
- Tier 3 design includes configuration management (Puppet) that can be used for real and virtual machines
- Choice of KVM hypervisor allows for long term support in OS
- XRootD data server can run on physical hardware and server data via virtual network (to be tested)

Conclusions

- Tier 3 computing important for data analysis in Atlas
- A coherent Atlas wide effort has begun in earnest
- Tier 3's are being designed according to the needs of the local research groups
- Striving for a design that requires minimal effort to setup and successfully run.
- Technologies for the Tier 3's are being chosen and evaluated based on performance and stability for data analysis
- Virtualization plays an important role

Backup Slides

Head node services

- Head node installation via kick start
- Private network configuration
- Virtual machine installation for Puppet Master
- VM installation for LDAP server
- VM installation for Squid server
- VM installation for Condor Collector/Negotiator
- Configuration of following services on head node
  - Dnsmasq (provides dns services to machines on private network)
  - firewall (for public interface)
  - Add local user accounts
  - NFSV4 mounts for configuration files and AtlaslocalRootBase
  - Ganglia Client
  - Creation of atlasadmin account in LDAP
- Test Head node configuration

NFS node services

- Kick start file installation of OS configuration of disks
- Private network configuration (including NAT) and testing
- Installation of local accounts
- Creation of NFS V4 exported file systems
- Installation of Ganglia server and web server
- Installation LDAP client authentication
- Firewall configuration
- Installation/Configuration of Condor config files

Interactive node (installation)

- Node installation via kickstart file
- Private network configuration
- Local accounts
- Addition of extra libraries need for Atlas software
- Installation of manageTier3SW by atlasadmin account on NFS mounted disks
- Installation of CernVM-FS for Atlas software and conditions DB
- Simple Atlas Test analysis jobs (to be written)

Batch Node installation

- Installation via kickstart files initially
- Static IP address on Private network
- Condor software installation and configuration
- Installation of required additional libraries for Atlas software
- CernVM-FS installation/configuration
- Condor system testing

Proof-Lite Performance scaling in Xen VM

Sergey Panitkin

BNL

BROOKHAVEN
NATIONAL LABORATORY

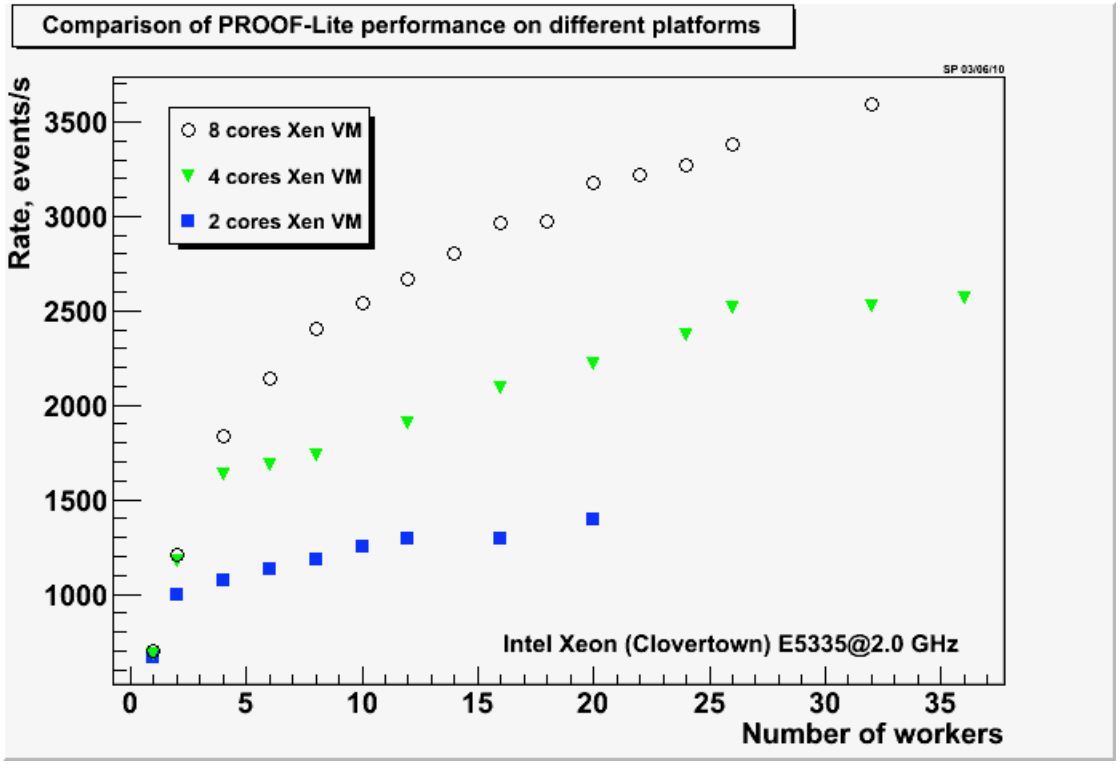
Introduction

- An interesting use case emerged when physicists at BNL started to analyze LHC data in 2010 at the Atlas Computing Facility (ACF) at BNL
- They used PROOF-Lite based analysis on interactive login nodes
- Questions about analysis performance arose
- This prompted a study of PROOF-Lite performance in virtual environment

General Information

- A set of interactive login machines at Atlas Computing Facility (ACF) at Brookhaven Lab are available for user code development and interactive analysis
- Currently there are 12 interactive login machines available
- They are Xen virtual machines (VMs) in two different configurations
- 8 machines - acas000[1-8] are 2 core VMs mapped to 2 physical cores
- 4 machines - acas00[09-12] are 4 core VMs mapped to 4 physical cores
- All VMs have similar underlying hardware, hypervisor and OS
  - Each node has dual quad core Intel Xeon CPUs (E5335) running at 2 GHz and 1Gbs network interface
  - VMs with 2 cores are configured with 3GB RAM
  - VMs with 4 cores are configured with 8GB RAM
  - Currently all VMs run 2.6.18-164.2.1.el5xen x86_64 kernel
- Typically remaining physical cores on the same node are used by another VM which is configured to be a part of a Condor batch queue (production). For our tests the “neighboring” batch VM was disabled to reduce interference with measurements

PROOF-Lite on different platforms



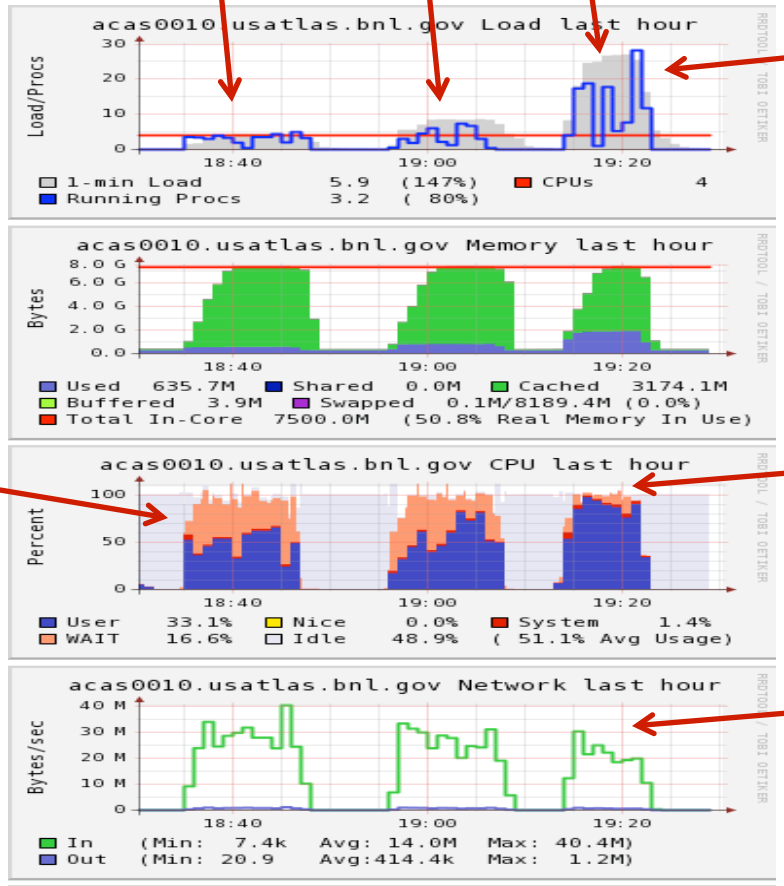
Dependence on the number of cores in VM.
More cores in a VM leads to a better performance. Hardly a surprise!

Performance curve has two slopes
Performance keep growing with increased number of P-L workers. Why?

Hardware monitoring plots I

4 cores Xen VM

#workers 4 8 32



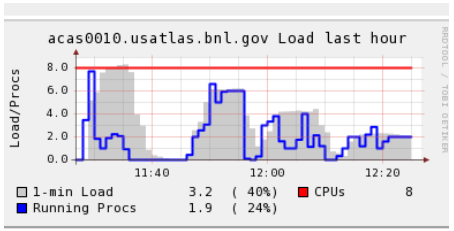
Note that number of simultaneously running processes does not exceed 30. Some processes are waiting for data

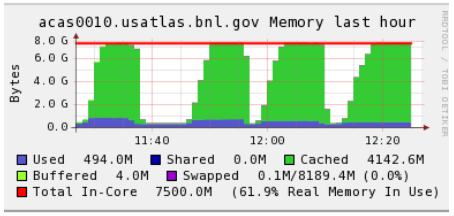
Large Wait IO fraction
User load ~50%

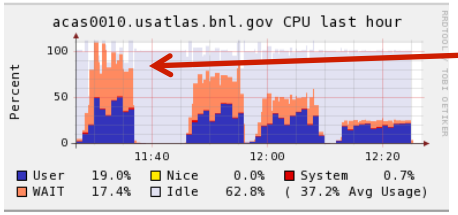
Wait IO fraction is diminished for large number of workers
User load close to 100%

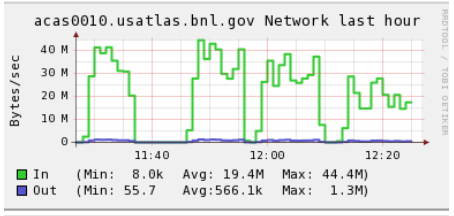
Network traffic does not exceed 40 MB/s

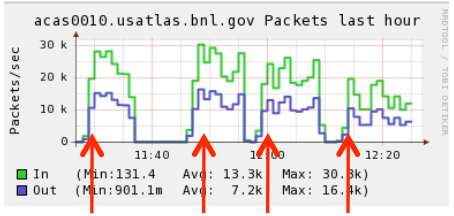
Hardware monitoring plots II

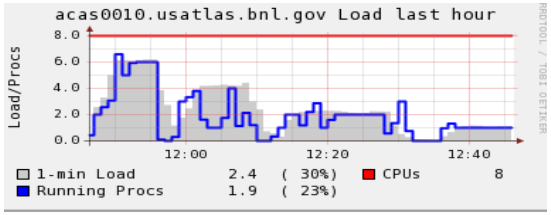


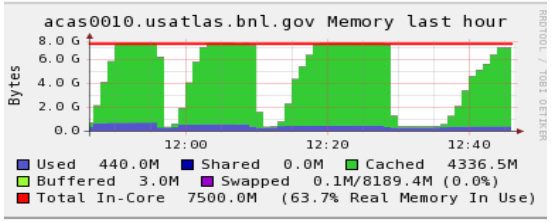


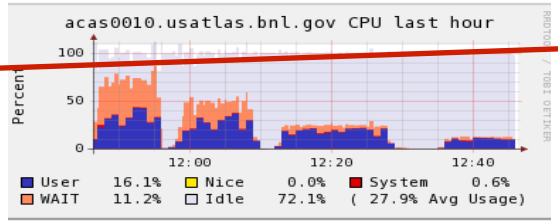


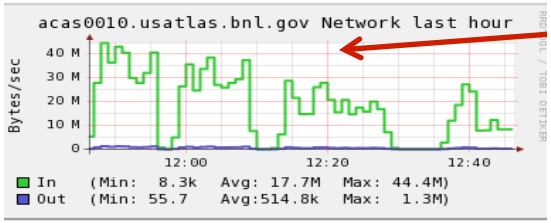


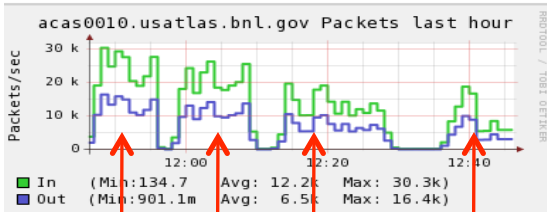












8 6 4 2

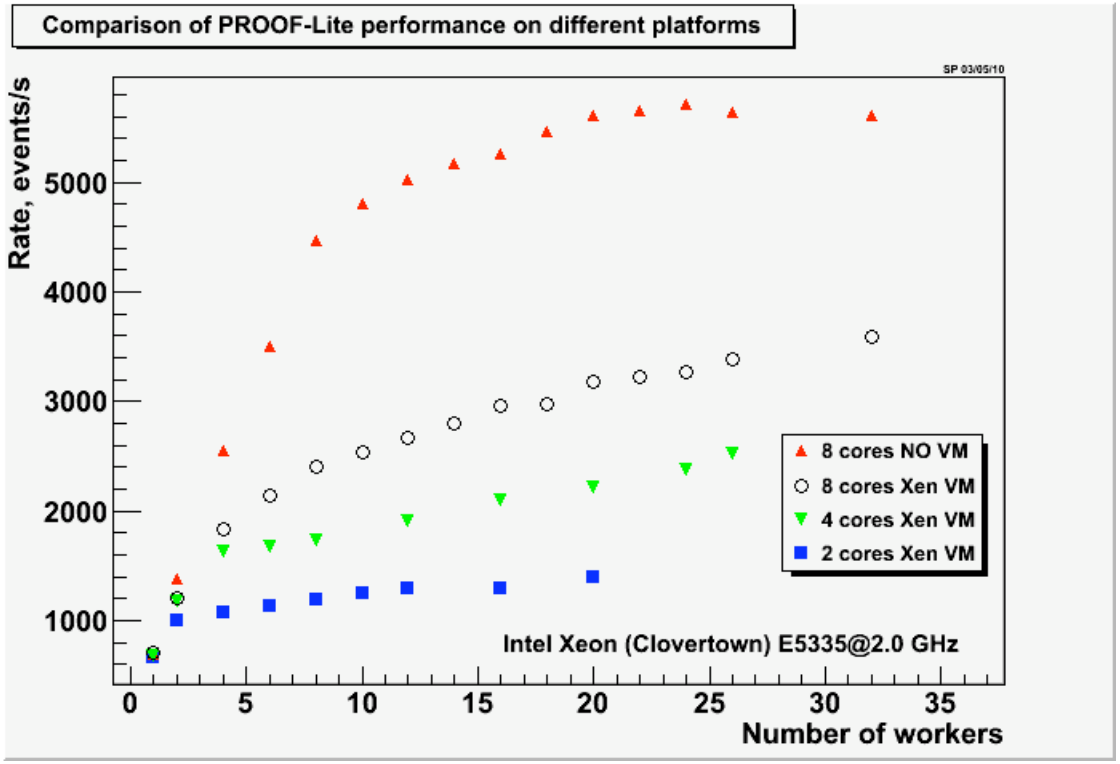
6 4 2 1

8 cores Xen VM

Large Wait IO fraction
User load ~50%

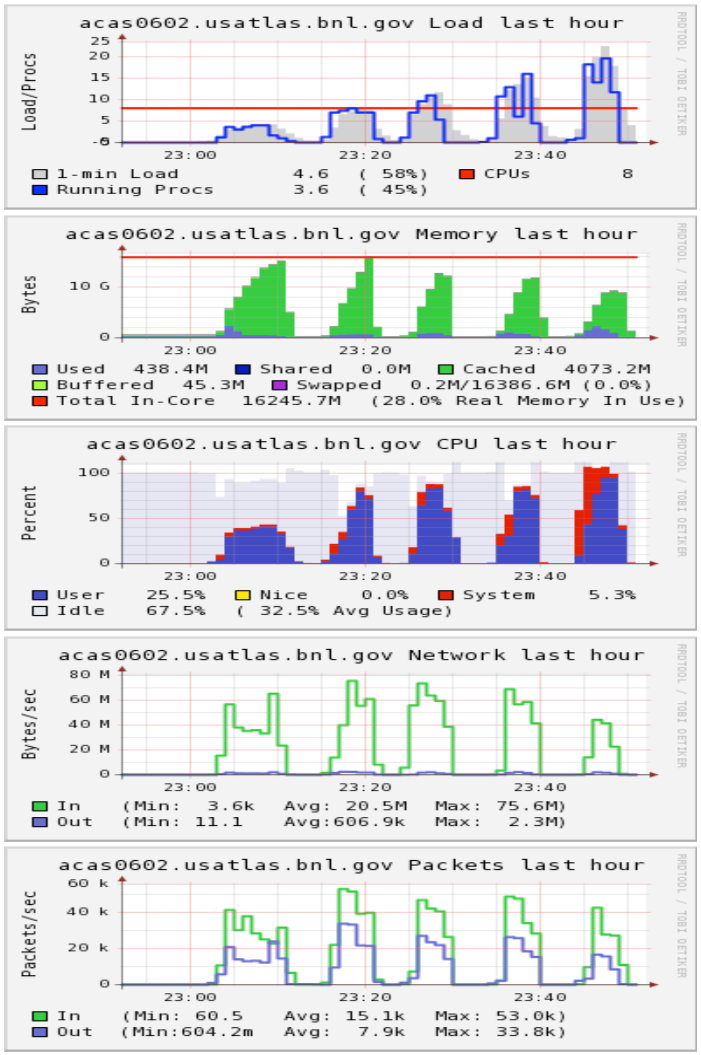
Network traffic does not exceed ~40 MB/s

PROOF-Lite on different platforms



It looks like PROOF-Lite running in a VM suffers severe performance penalty
~x2 worse performance than on “bare metal”

Hardware monitoring plots III



8 cores “bare metal”

#workers	4	8	12	16	24