UNIVERSITY OF OSLO

ATLAS

Performance of an ARC-enabled computing grid

for ATLAS/LHC physics analysis and Monte Carlo production under realistic conditions

SAMSET, Bjørn H. (Fysisk institutt - University of Oslo), for the ATLAS collaboration

with

CAMERON, David (*NDGF*)
ELLERT, Mattias (*Uppsala Universitet*)
GRONAGER, Michael (*NDGF*)
MAFFIOLETTI, Sergio (*Swiss National Supercomputing Centre*)
PAJCHEL, Katarina (*Fysisk institutt - University of Oslo*),
TAGA, Adrian (*Fysisk institutt - University of Oslo*)
FILIPCIC, Andrej (*Institut "Jožef Stefan"*)
KLEIST, Josva (*NDGF*)
OULD-SAADA, Farid (*Fysisk institutt - University of Oslo*),
READ, Alexander Lincoln (*Fysisk institutt - University of Oslo*)

Data access on ARC-enabled sites - outline:

- ARC-enabled sites in ATLAS, and the ARC middleware
- ATLAS MC production on ARC-enabled sites
- Features of user analysis vs official production
- Readiness of ARC and ARC-enabled sites to cope with user analysis within and outside ATLAS
- Lessons from large scale testing of analysis jobs

- Core issue:
  - User analysis is different, more random, than central production
  - Performance for MC prod + user analysis will show robustness of grid sites
  - The user rush is coming with LHC data, and the sites can't hide

ATLAS distributed computing

10 clouds, 3 grids, 3 middlewares, 1 production system

Open Science Grid

VDT

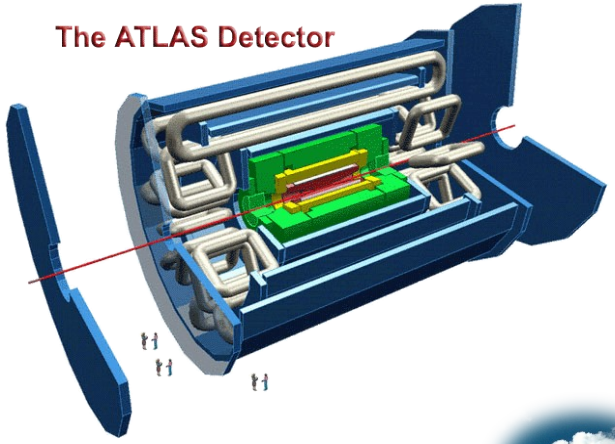
ES

CA

CERN

US

The ATLAS Detector



DE

NDGF

FR

NORDUGRID

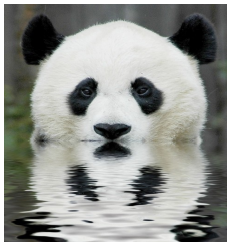
Grid Solution for Wide Area Computing and Data Handling

ARC

UK

TW

IT



The ARC middleware

- Lightweight, non-intrusive
- Sites advertise what they have available, not forced to install any specific system
- For ATLAS: Must provide the active ATLAS software kit releases, details left to the site
- Grid manager:
  - Manages job brokering, load balancing
  - *Stages in input files before the job runs*
  - Moves output to grid storage if requested
- ARC storage:
  - Distributed storage within the cloud (SRM+dCache for LHC)
  - Caching system: files pulled from local cache – or source if not found in cache
- Very lightweight user client: <20MB

ARC-enabled ATLAS sites: 13 sites, >11k CPUs

ATLAS Grid Monitor

2009-03-11 CET 12:07:08

Processes: Grid Local

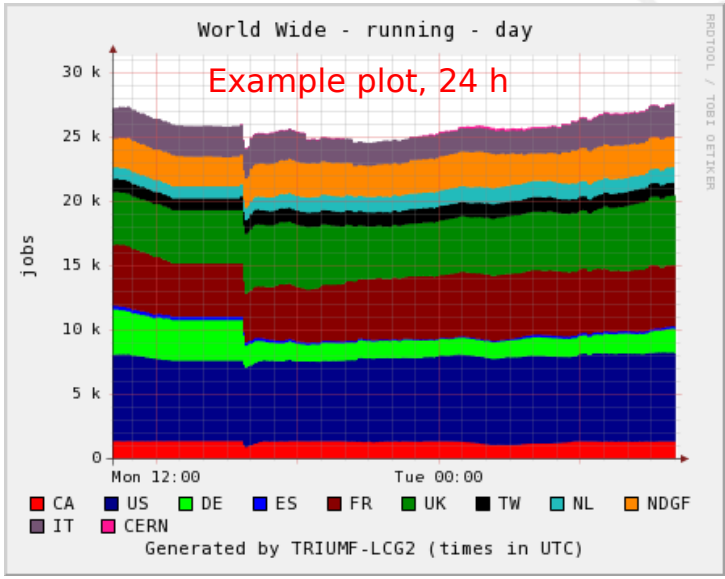
Country	Site	CPUs	Load (processes: Grid+local)	Queueing
Denmark	Steno (DCSC/KU)	2324	873+1084	**207**+3
Norway	EPF (UiO/FI)	17	1+5	**2**+0
	Titan A (UiO/USIT)	3956	843+2248	**404**+-70
Slovenia	SiGNET	586	577+0	**164**+1
Sweden	Grad (SweGrid, Uppmax)	512	318+60	**41**+13
	Ritsem (SweGrid, HPC2>)	432	408+0	**110**+0
	Ruth (SweGrid, PDC)	392	392+0	**119**+-1
	Siri (SweGrid, Lunarc)	496	188+229	**52**+20
	Smokerings (NSC)	528	506+14	**154**+-3
Switzerland	Bern ATLAS T3 Cluster	14	14+0	**26**+0
	Bern UBELIX T3 Cluster	1024	0+463 (queue inactive)	**0**+1
	Geneva ATLAS T3	148	0+120	**0**+120
	Manno PHOENIX T2	960	182+56	**144**+0
TOTAL	*13 sites*	*11389*	*4302 + 4279*	*1423 + 84*

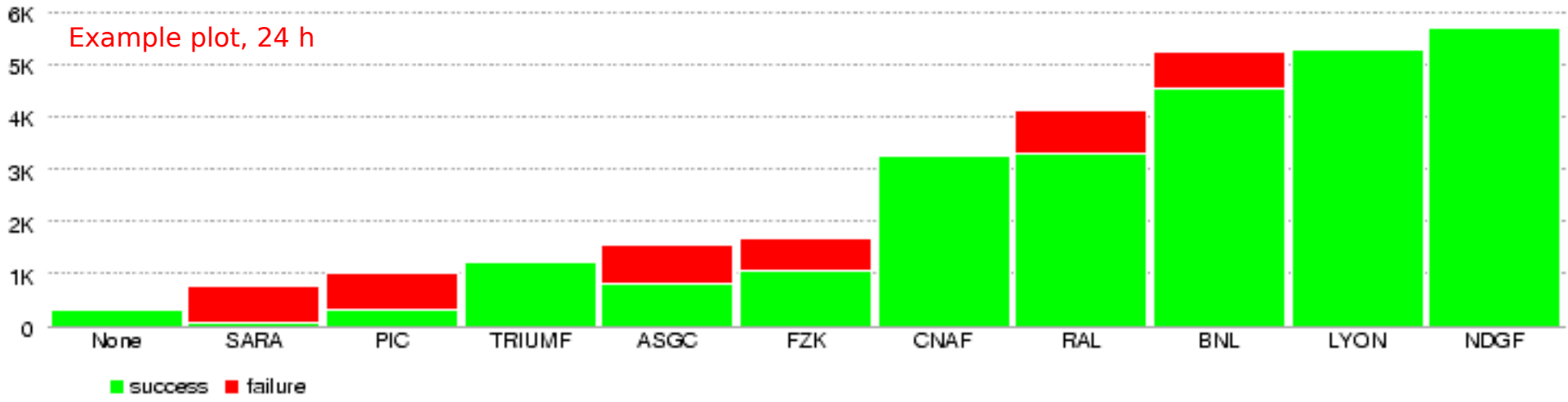
- Inhomogeneous cloud – several architectures, operating systems
- **Shared resources** between
  - ATLAS MC production
  - User analysis
  - External (non-HEP) usage
  - Local site users

NORDUGRID
Grid Solution for Wide Area Computing and Data Handling

ATLAS MC production on NDGF sites

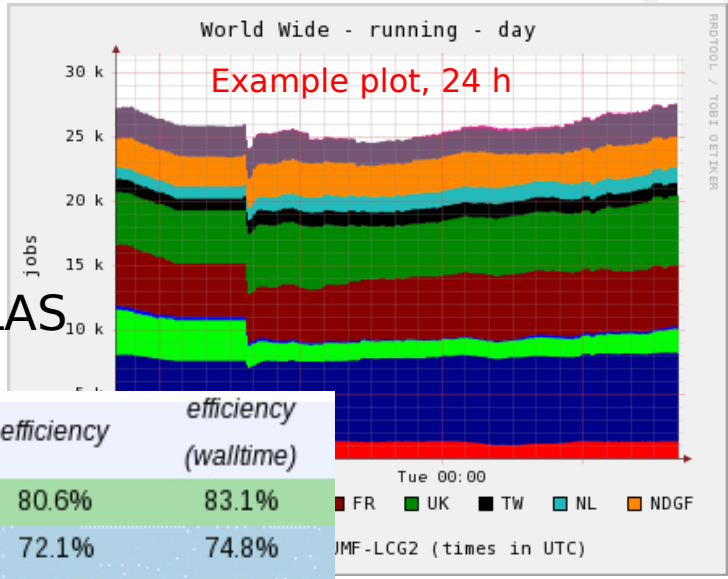
- ATLAS production system accesses ARC-enabled resources through the same interface as other ATLAS grids
- NDGF cloud shows stable performance, good efficiency, no scaling problems up to present size



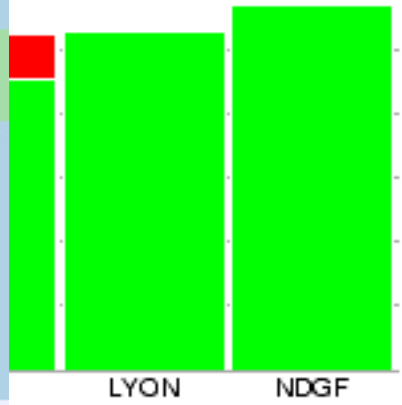


ATLAS MC production on NDGF sites

- ATLAS production system accesses ARC-enabled resources through the same interface as other ATLAS grids
- NDGF has highest annual efficiencies in ATLAS. Numbers for 2008:



cloud	success	failure	success (walltime)	failure (walltime)	efficiency	efficiency (walltime)
BNL	3493728	841093	65063597186	13219670831	80.6%	83.1%
LYON	1636582	632042	28756889477	9689269549	72.1%	74.8%
RAL	1181696	465182	28491890089	9835768394	71.8%	74.3%
FZK	1069164	285734	19789918222	3093157257	78.9%	86.5%
NDGF	1172710	95704	20349144540	640866720	92.5%	96.9%
TRIUMF	1067293	121087	17925637777	2137756729	89.8%	89.3%
ASGC	580902	322712	10350042275	1405728318	64.3%	88%
CNAF	548069	266693	10374576563	2414506380	67.3%	81.1%
None	363269	358981	6870094636	1811546252	50.3%	79.1%
PIC	514311	168786	8213783532	1508124912	75.3%	84.5%
SARA	469153	213281	8097905169	1741793495	68.7%	82.3%
CERN	34260	15575	222659148	83117009	68.7%	72.8%
total	*12131137*	*3786870*	*2.24506138614e+11*	*4.7581305846e+10*	*76.2%*	*82.5%*



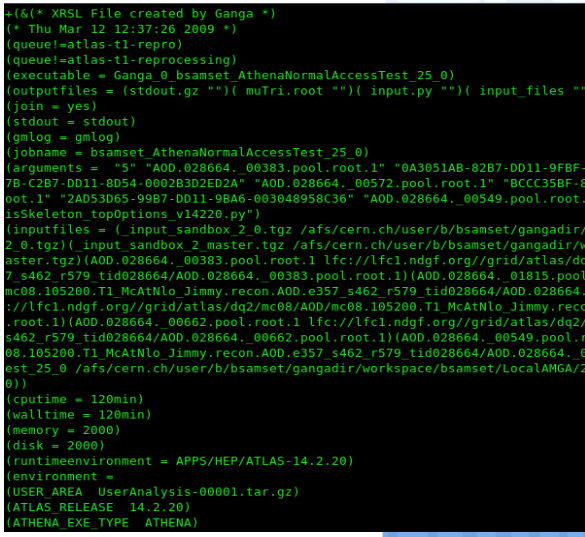
User Analysis – how it differs from MC production

- Official production/reduction:
  - Sets realistic requirements for sites
  - Well tested before submission
  - Uses official, validated software
  - Average I/O ratio of 1-10 (input 1-10 times the size of the output)
    - Gives good CPU utilisation, disk IO not a bottleneck
- User analysis
  - Users don't know what load their jobs will put on the systems
  - Software is often private, untested, buggy
  - May need to be compiled, set up on worker node
  - I/O ratio can be up to 10^6 (the case for test jobs used for this talk)
    - Reading from storage becomes bottleneck, leading to low CPU usage
  - Random access (e.g. every ~1000^{th} event) may be required
- Performance for MC prod + user analysis will show robustness of sites!

User analysis in the NDGF ATLAS cloud

- Low level access:
  - Users can write xrsl job description files
  - All resources available, very flexible
  - Not user-friendly...
- High level access:
  - ARC user client is integrated and shipped with GANGA

    Ganga: User-friendly Grid job submission and management tool for LHC and beyond by Dr. Daniel VAN DER STER (CERN) Monday 14:40 - 15:00
  - Allows user access across ATLAS, i.e. can submit to all clouds and middlewares
  - Also interesting for non-ATLAS ARC users
- User status:
  - A number of casual and power users from all corners of ATLAS, ~all using ganga
  - In general good feedback, users stay and don't complain




```
#   73   completed   bsamset_AthenaNorm
#   74   submitted   bsamset_AthenaNorm
#   75   submitted   bsamset_AthenaNorm
#   76   submitted   bsamset_AthenaNorm
#   77   completed   bsamset_AthenaNorm
#   78   completed   bsamset_AthenaNorm
#   79      failed   bsamset_AthenaNorm
#   80      failed   bsamset_AthenaNorm
#   81  submitting   bsamset_AthenaNorm

In [3]:
```

Scattered experiences before systematic testing

- Experience from:
  - Large scale ATLAS user analyses by single users
  - Testing during user interface development
- Good features:
  - Fast, stable input access
  - Very low rate of grid-related errors
  - Environment announcement ensures availability of requested software
  - Quick turnaround
  - Access to good monitoring system, incl. live log file peeking
- Issues:
  - Varying underlying architectures makes compiling user code challenging
  - Integration of services developed for other systems, like ATLAS-specific dataset handling

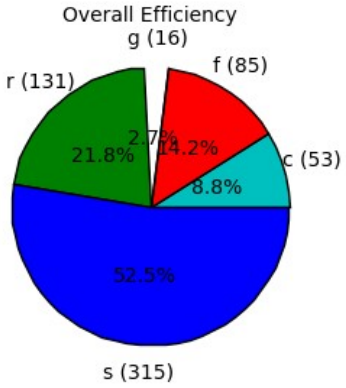
Cloud-wide tests: HammerCloud – hammering the cloud with analysis jobs

Functional and Large-Scale Testing of the ATLAS Distributed Analysis Facilities with Ganga
by Daniel Colin VAN DER STER (CERN) Thursday 15:20 - 15:40

- First test choked entire system, due to huge input files
  - File staging is different in ARC than for other ATLAS grids
  - Almost a designed failure
- Lessons:
  - We have to limit the input files per job
  - Users will not think to do this, i.e. A single user can take down a site. Not desirable... Must be implemented in the code.
  - Later fixed
  - File access methods are crucial for user analysis
  - Compilation must be sorted
    - Solution found

Overall Efficiency

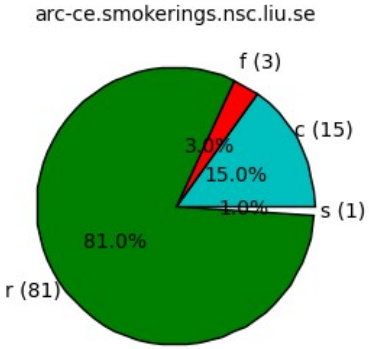
Legend: s: submitted, r: running, g: completing, c: completed, f: failed

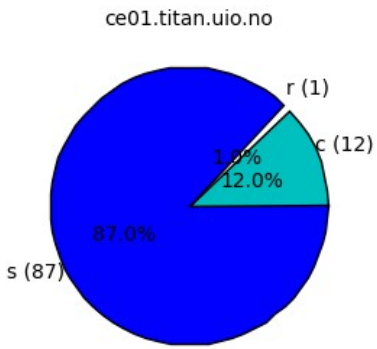


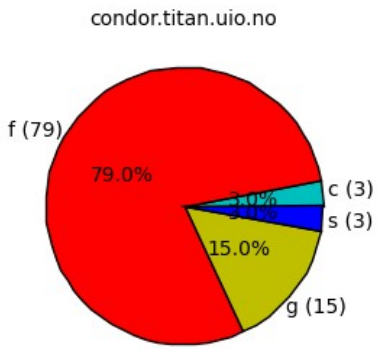
NDGF HammerCloud 1: Status after 24 hours

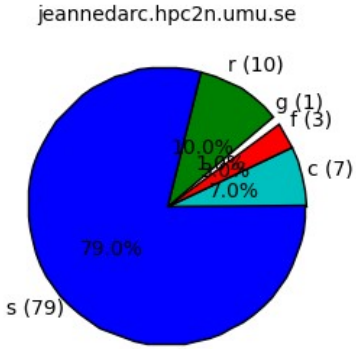
- This raised awareness of user analysis demands and features

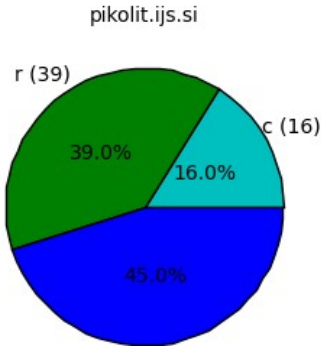
Site Efficiency

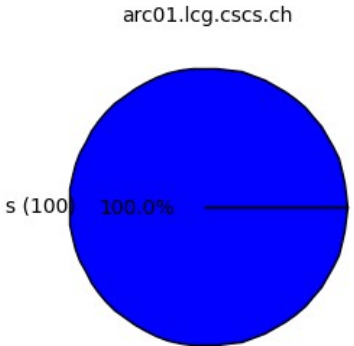






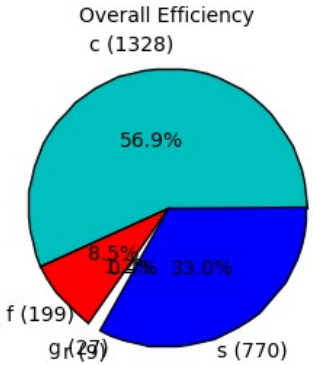






Overall Efficiency

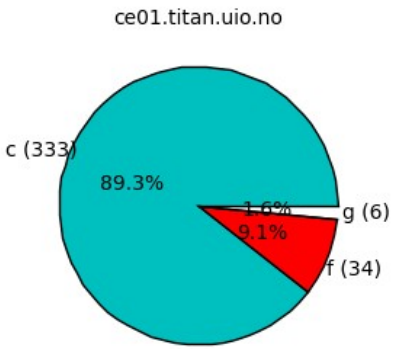
Legend: s: submitted, r: running, g: completing, c: completed, f: failed

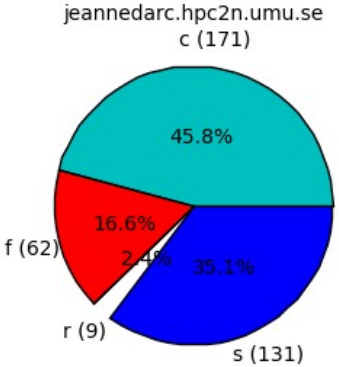


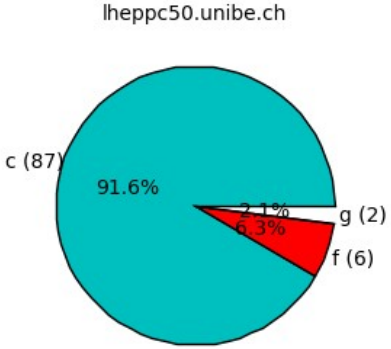
NDGF HammerCloud 2: Status after 48 hours

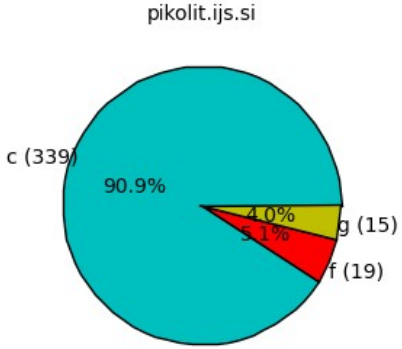
- Reasonable inputs (<5GB)
- Sites were not warned
- Some known issues, like an unaccessible SRM pool and a middleware bug (later fixed)

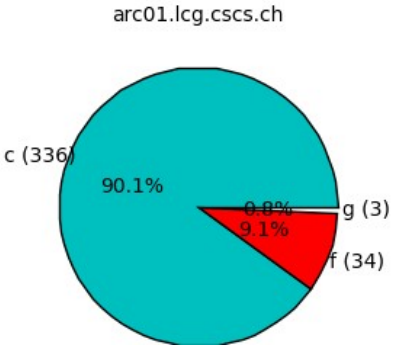
Site Efficiency

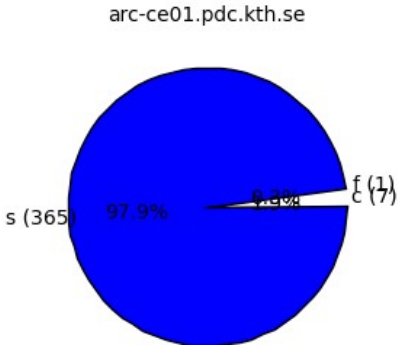


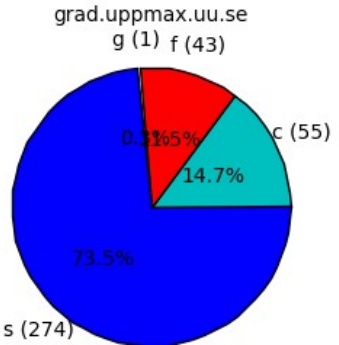








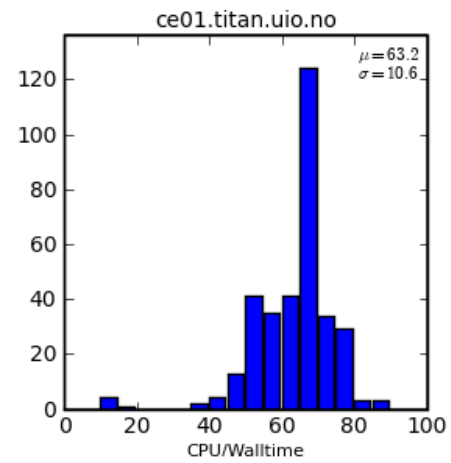


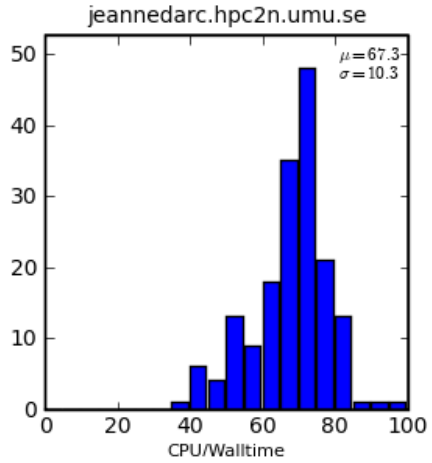


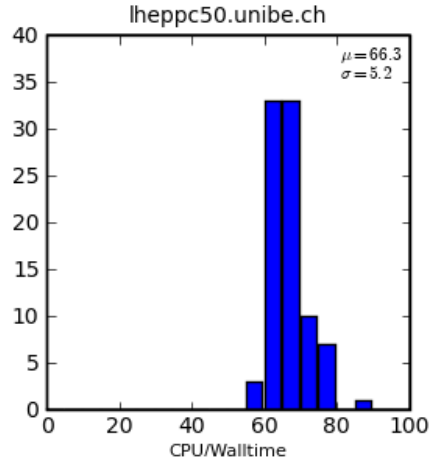
CPU usage for user jobs

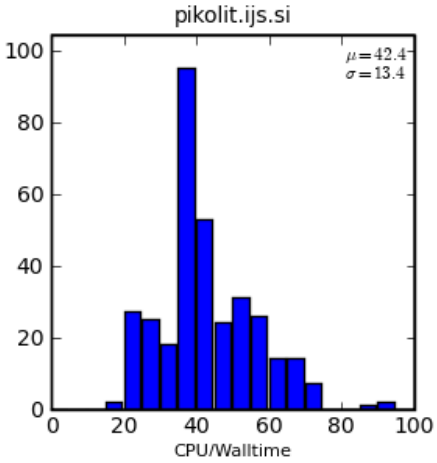
- 70-80% is considered good
- NDGF sites performed well
- Pre-staging of inputs is one reason for this

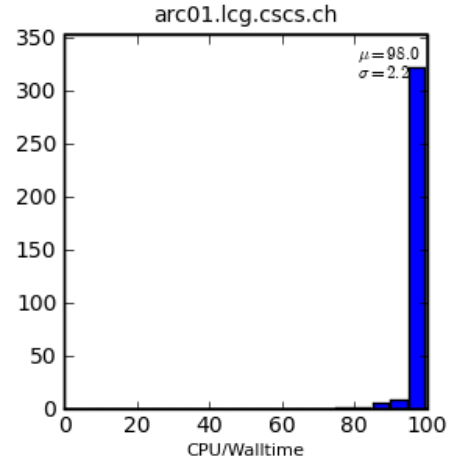
Site CPU/Walltime

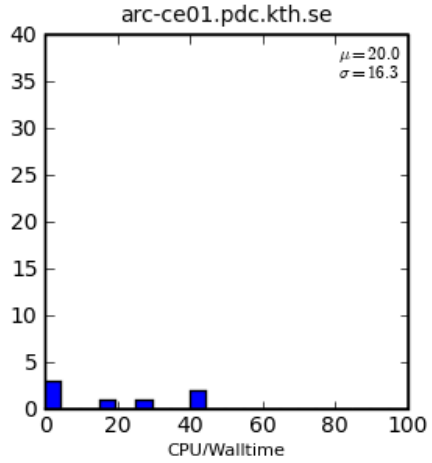


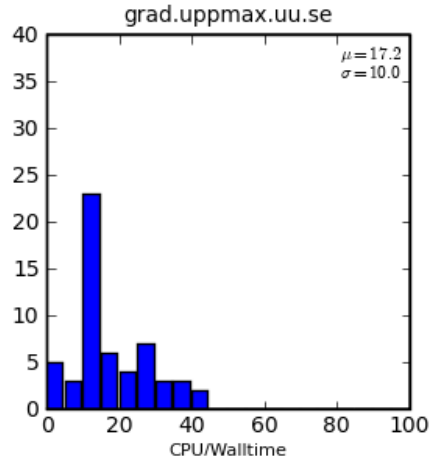




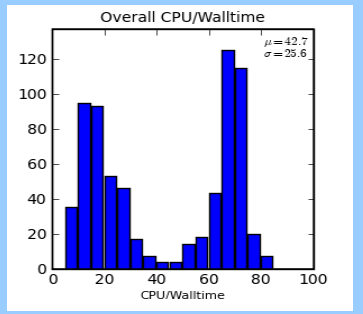








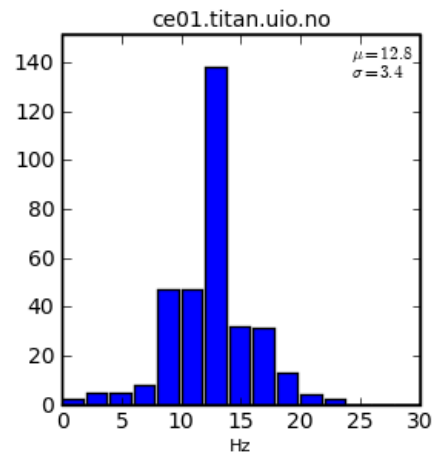
Comparison:
Test on Tokyo T1
(all sites)

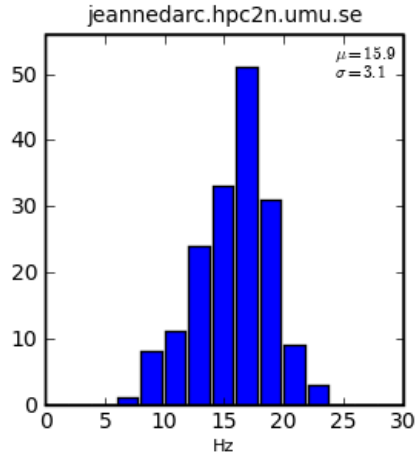


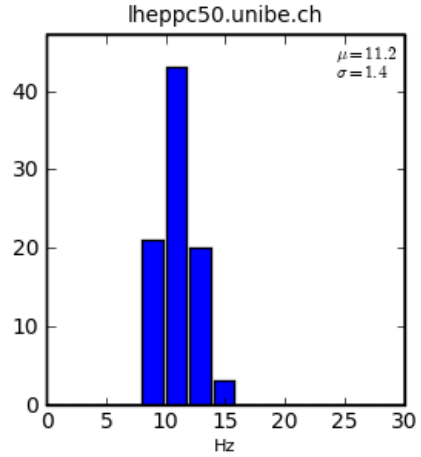
Events per second analyzed

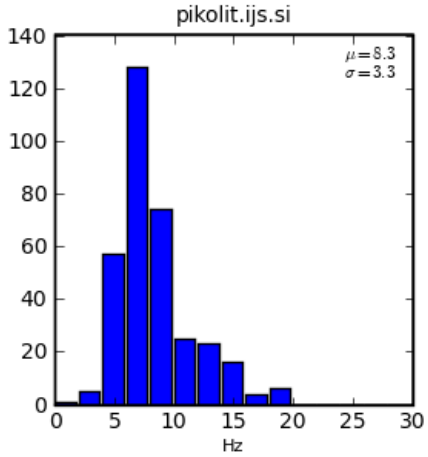
- 15Hz is nominal rate (analysis dependent)
- Most NDGF sites average at this value
- Some slow sites visible, various reasons

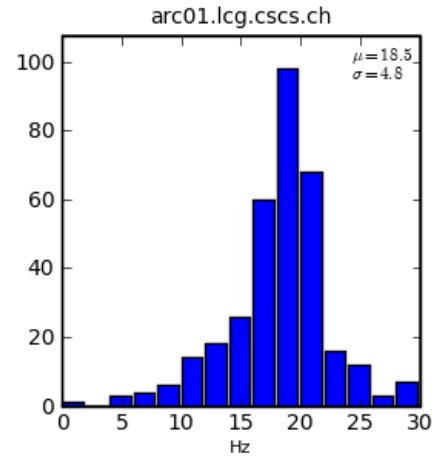
Site Events/Second

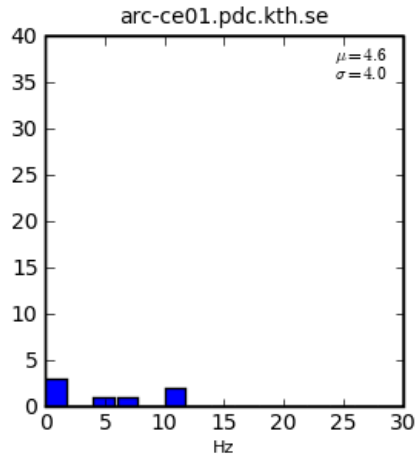


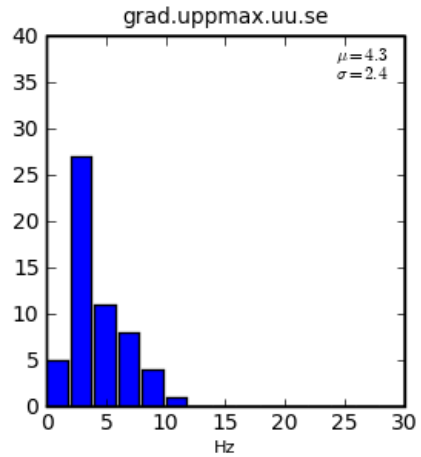




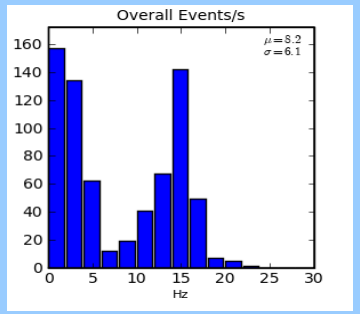








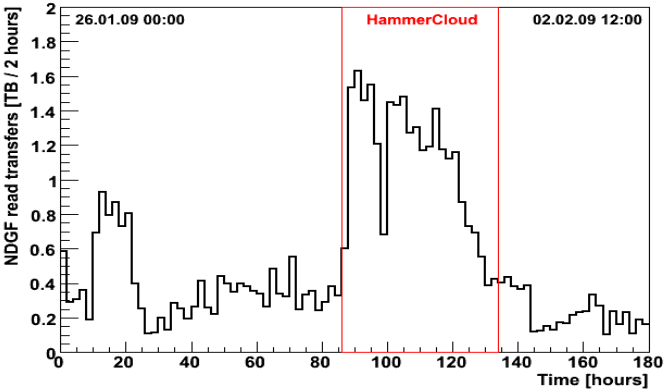
Comparison:
Test on Tokyo T1
(all sites)

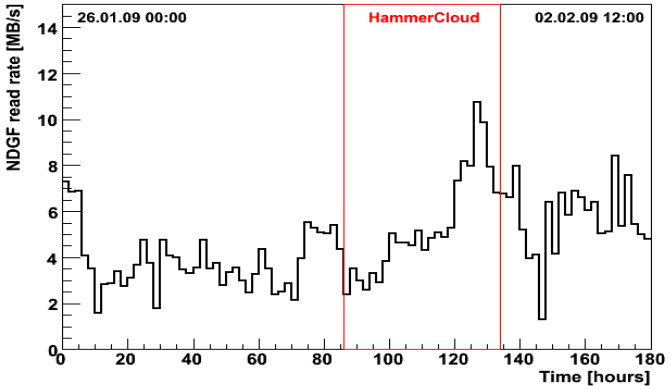


NDGF SE performance during HammerCloud test

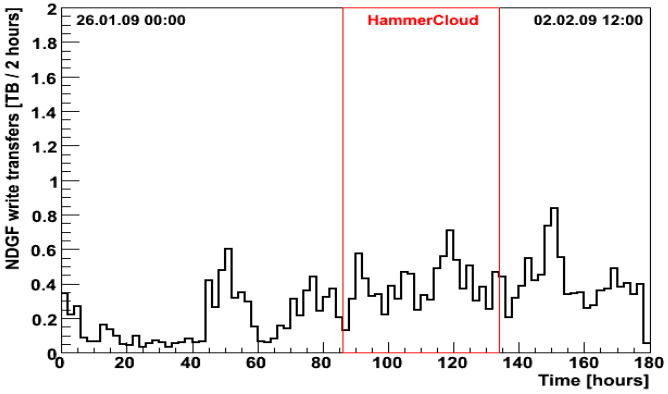
- All read/write transactions to NDGF dCache
- Test clearly visible in total amount of data read
- Not visible in writing (jobs have small outputs)
- No adverse effect on internal transfer rates

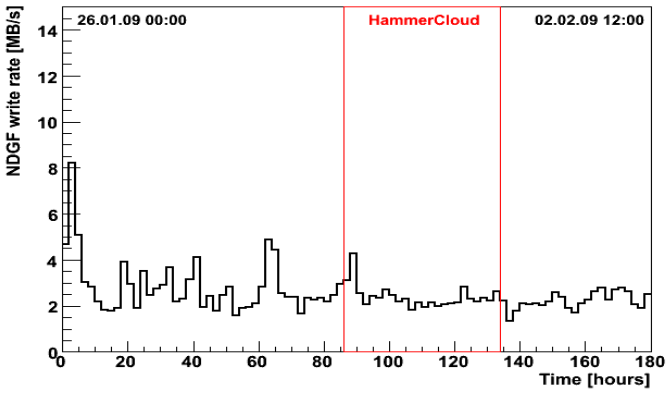
Read





Write



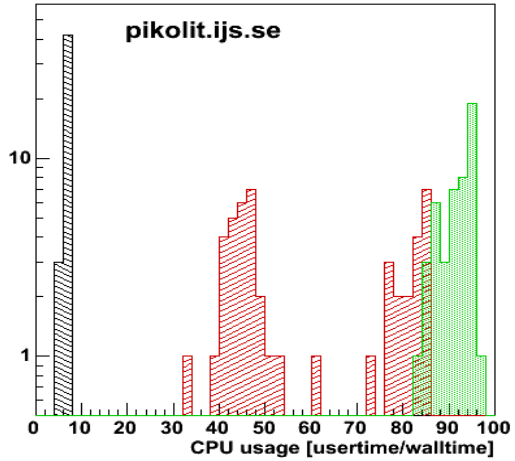


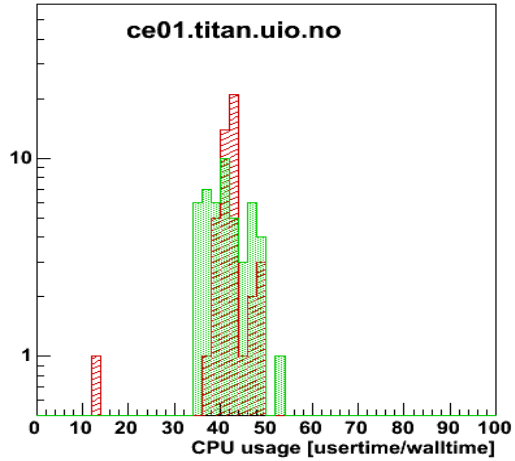
Going beyond HammerCloud: Data access solutions

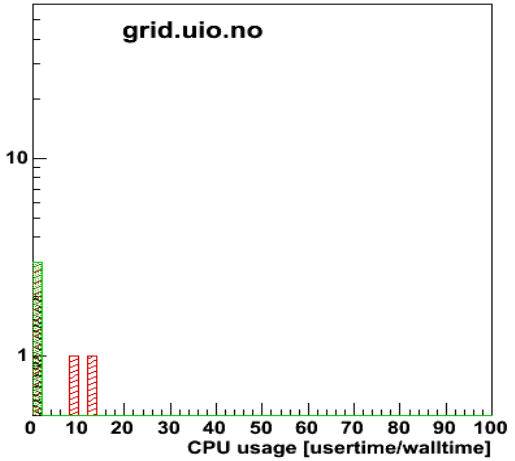
- Some limitations on user jobs have to be put in on
  - Site level: Actively use job requirements on e.g. input size
  - User tool level: Split large jobs automatically, warn about unreasonable requirements
- Beyond this we can try to optimize e.g. file access:
  - "Pure ARC" - let the grid manager handle file transfers
  - Local disk access – move files from local shared file systems to grid job working directory
  - Gsidcap – read straight from dCache pool via gsidcap
  - xrootd

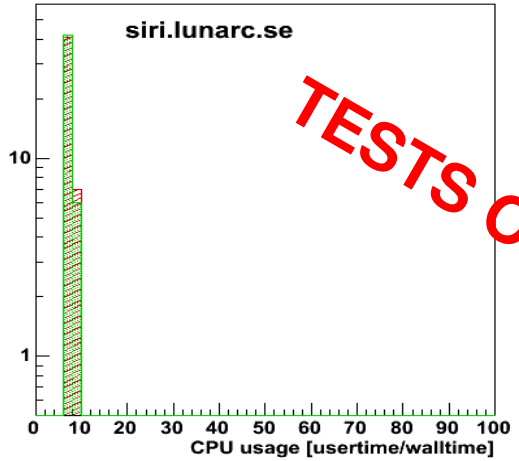
File access comparisons

- Five selected sites, at one random time (several factors influencing performance), “nasty” jobs with high IO ratio
- gsicap only on pikolit. Not performing well...
- ARC/local approx. similar – interesting site patterns

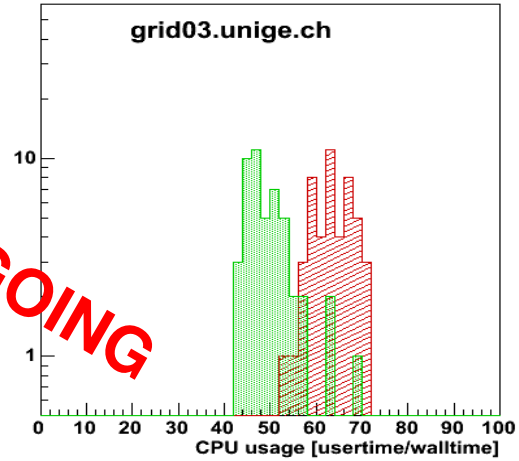


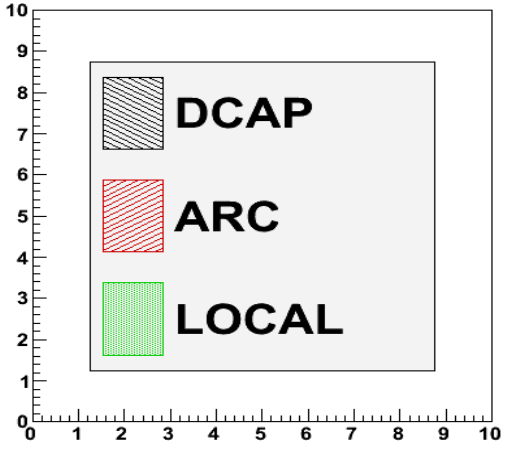






TESTS ONGOING





Other observations from user analysis on ARC sites

- Single user brought one site to standstill by writing to SE in another cloud
  - This must be disallowed at the grid user tool level
- User data on grid storage must be regularily cleaned
  - In progress now
  - Corrollary: Users must have an easy way to move their data from grid to local storage
- Sysadmins must be aware of user jobs
  - Used to the behaviour of MC production jobs
  - User jobs are different – if they're too different, sysadmins are quick to ban the user (has happened in three known cases)

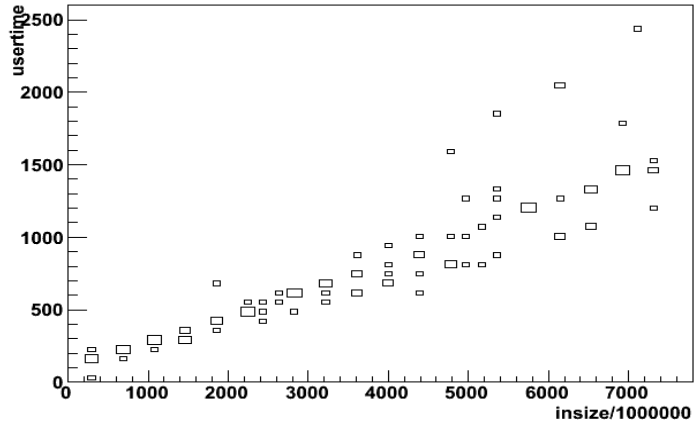
The NDGF cloud performs well for ATLAS analysis

- User job I/O ratio makes file staging, file access important parameters
- ARC-enabled sites are able to efficiently run both official, tuned MC production and user analysis jobs, even simultaneously on the same resources
- Users are happy, though some level of technical skill is still required to run analysis jobs
- To avoid creating bottlenecks, some limitations must be enforced on both site and user level
- A hetereogenous cloud must take greater care in ensuring a good local environment for user jobs
- The NDGF cloud in ATLAS will not be killed by users when the LHC data arrive. It looks set to be a nice place to do one's analysis.

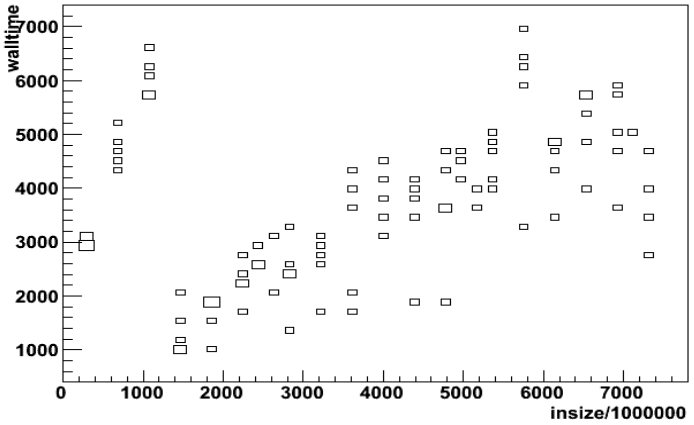
BACKUPS

Sanity check: Performance vs input size on pikolit.ijs.se

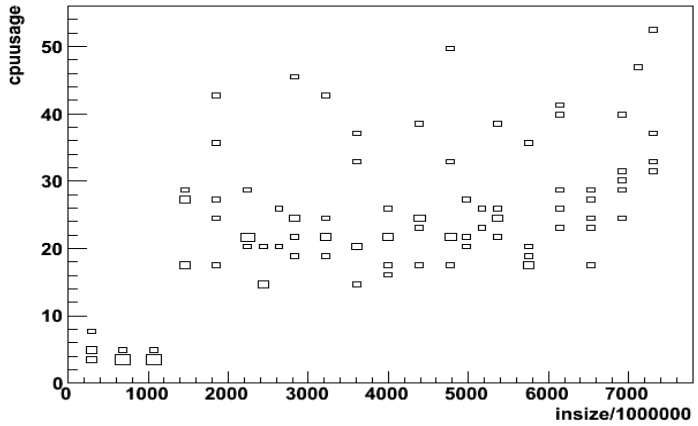
usertime:insize/1000000



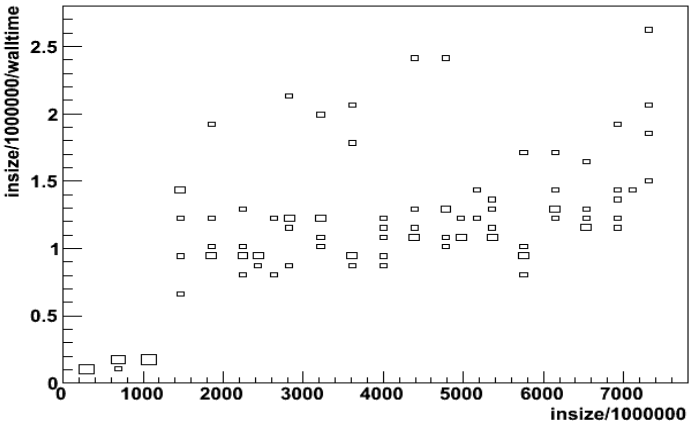
walltime:insize/1000000



cpuusage:insize/1000000



insize/1000000/walltime:insize/1000000



Comparisons on other site (siri, grid?) To be put in for comparison