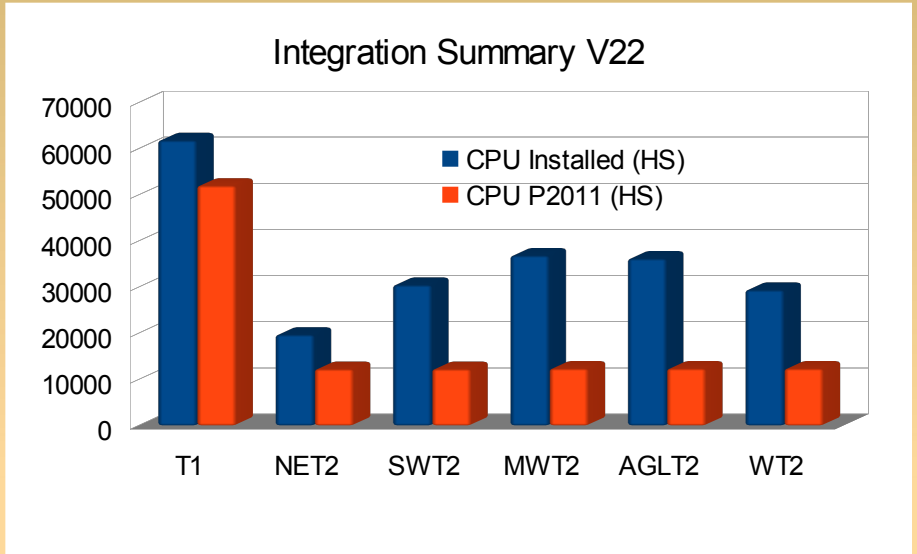
Above Pledge Resources

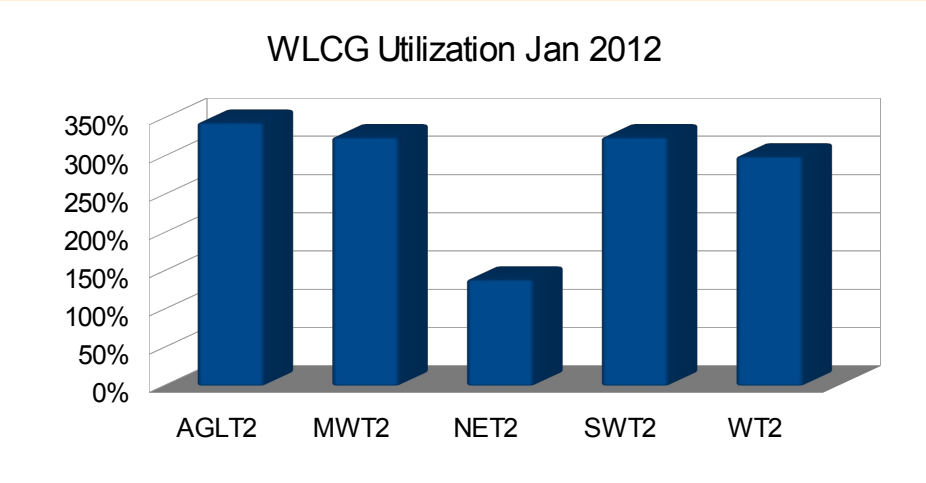
Kaushik De
Univ. of Texas at Arlington

US ATLAS Facilities, Lincoln
March 19, 2012

US Resource Usage

- US ATLAS sites provide more CPU than WLCG pledge
  - Justified since utilization is very high
  - Would like to directly benefit US users
- Storage situation is different – not so much extra capacity





But PanDA is Designed for Speed not Pledge

- PanDA automatically & optimally distributes work among available computing resources
  - Designed primarily for fast execution of jobs
  - Brokerage module in PanDA server
    - Distributes tasks among clouds
    - Assigns jobs to sites (start data transfer if necessary)
  - Dispatcher module - feeds available work to pilots
  - PD2P manages data distribution for user analysis
  - All sites are treated equally irrespective of user VO

Recall - Site Selection in PanDA

- Site selection is based on a variety of factors:
  - Different algorithm for production & user jobs
  - Production task and job assignment
    - Initial ask assignment loosely based on MoU shares
    - Other cloud assignment based on location of input files
    - Site is chosen for fastest execution
  - User analysis job assignment
    - Jobs go to data – input must be available at site
    - Site is chosen for fastest execution

Recall - Job Execution in PanDA

- Pilot asks PanDA server for a job
- Dispatcher sends job
  - Highest priority job among those assigned to site
  - Special handling for user analysis
    - Priority is boosted for local users by site policy
    - Priority is boosted for production managers
    - Priority is lowered based on fair share algorithm
  - Recently shares were implemented for task types
    - Production jobs only, MC vs Group vs Reprocessing etc

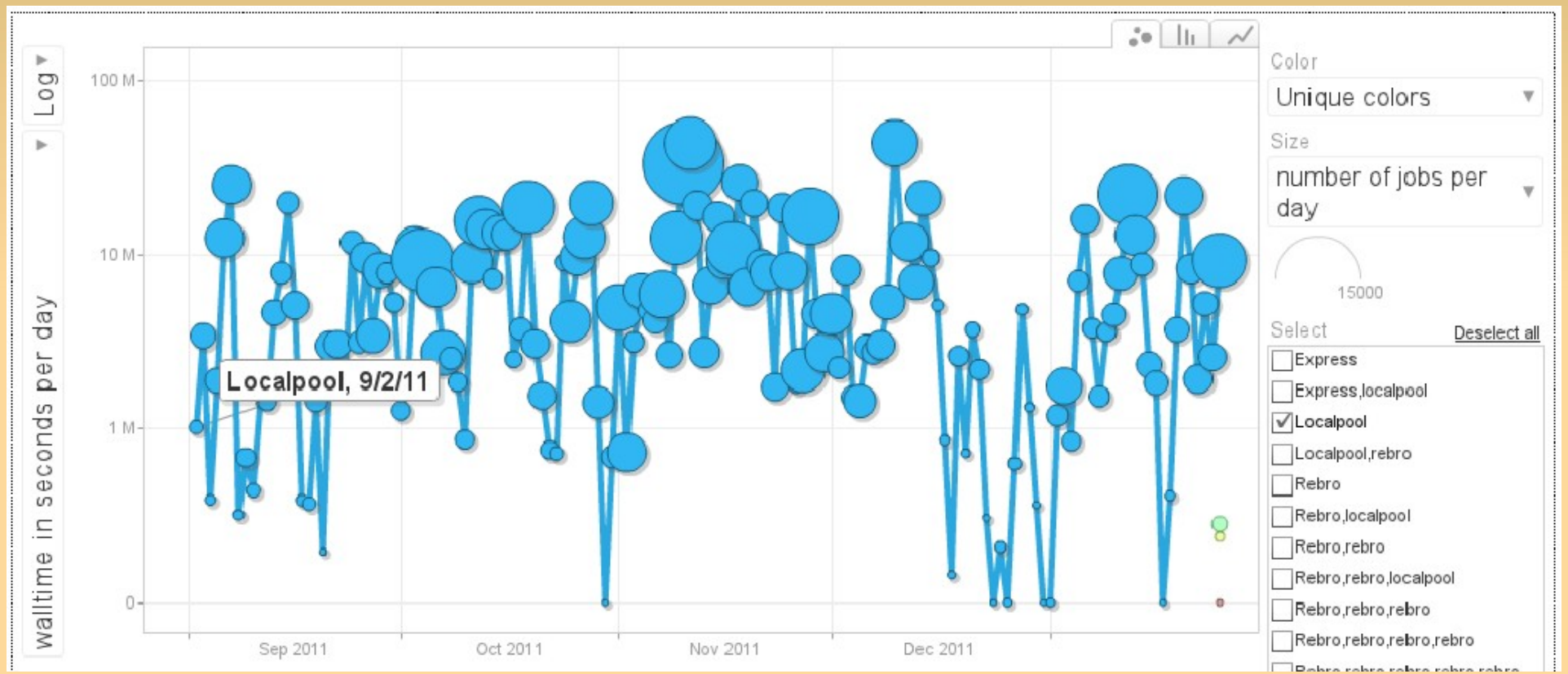
Above Pledge – Phase I

- We have gradually added special handling and special procedures so that US users can benefit from the above pledge CPU
- Procedure for additional production – Phase I
  - US users can request production tasks to run in US
  - Requests made through CREM
  - Once approved, tasks are submitted by central production team, to run in US only
  - Recently re-added a flag for accounting

Above Pledge – Phase II

- Since we already have priority boosting & priority lowering mechanism in PanDA dispatcher, this is the natural place to start automatic VO matching
  - We defined a new user analysis tag for jobs called localpool (special handling field in DB)
  - For each US site we store excess CPU capacity
  - Fraction of jobs in proportion to capacity get priority boost - only USATLAS user jobs at US sites
  - Of course, this is a small change, but good start

Egg Plot


Log
walltime in seconds per day
100 M
10 M
1 M
0
Sep 2011
Oct 2011
Nov 2011
Dec 2011
Localpool, 9/2/11
Color
Unique colors
Size
number of jobs per day
15000
Select
Deselect all
Express
Express,localpool
Localpool
Localpool,rebro
Rebro
Rebro,localpool
Rebro,rebro
Rebro,rebro,localpool
Rebro,rebro,rebro
Rebro,rebro,rebro,rebro

Above Pledge – Phase III

- Dispatcher can only raise priority for jobs already assigned to a site
  - Job assignment is VO neutral in PanDA
- Brokering is needed to preferentially assign US user jobs to above pledge US resources
  - We do not want to restrict US users to US sites, or DE users to DE sites …
  - We do not want to slow down execution of user analysis jobs

Phase III - Implementation

- Added extra weight in brokering in proportion to the above pledge resources available
- All other brokering factors – data locality, site availibility, site load etc are still used
- Turned on this feature about 5 weeks ago
- No plots yet – but spot checks show no adverse effects
- Beneficial effects for US users still to be proven

Localpool in PanDA Monitor

Analysis job summary at 03-19 14:08, last 12 hours (Details: errors, nodes) pathena analysis queue status
Processing types: ganga(8809) gangarobot(14688) hammercloud(1936) pathena(223018) prun(74923) usermerge(311)
Working groups: det-indet(1242) det-muon(48) det-slhc(1326) perf-jets(1077) perf-tau(4613) phys-beauty(105) phys-higgs(600) phys-susy(97)
Special Handling: express(21) localpool(1433) rebro(1730) rebro,rebro(1699) rebro,rebro,rebro(1407)
Pilot counts are for the last 3 hours. Error rates above 20% are shown in red.

Cloud	Pilots	Latest	pending	defined	waiting	assigned	activated	sent	starting	running	holding	transferring	finished	failed	cancelled
ALL			0	12299	0	0	72545	1	445	19499	2363	21	132904	21121	62487
CA	648	03-19 14:05	0	45	0	0	8781	0	0	658	135	0	8594	857	2411
CERN	968	03-19 14:05	0	3	0	0	669	0	0	340	20	0	3666	6709	2430
DE	1771	03-19 14:05	0	362	0	0	9040	0	0	2303	403	0	25568	3265	3917
ES	360	03-19 14:05	0	44	0	0	3521	0	0	326	83	0	4723	101	1534
FR	2322	03-19 14:05	0	329	0	0	24985	0	1	2883	355	0	16484	1341	19552
IT	1033	03-19 14:05	0	85	0	0	9748	0	0	1325	260	0	13409	1695	2168
ND	1406	03-19 14:05	0	30	0	0	9	0	442	596	47	21	4424	969	711
NL	1110	03-19 14:05	0	17	0	0	3282	0	1	800	181	0	13383	464	164
TW	200	03-19 14:05	0	4872	0	0	3005	0	0	697	9	0	588	2986	2
UK	2003	03-19 14:05	0	4607	0	0	3650	0	0	3470	349	0	27233	1265	10840
US	3774	03-19 14:05	0	1905	0	0	5855	1	1	6101	521	0	14832	1469	18758

What Next

- Need better monitoring of localpool!
- Need better monitoring of additional production!
- Tune current fraction in dispatcher
- Tune current weight in broker
- Phase IV – implement new PanDA production shares for US sites, automate additional prod?
- Phase V – above pledge calculations in PD2P
- Phase VI – above pledge in JEDI (fka PDJD)

Summary

- ATLAS benefits greatly from above pledge resources available in US sites
  - Clearly, computing model estimates are too low
  - Pledges need to be increased for all sites
- US users benefit from being able to run anywhere – not just the US
- Slowly but steadily implementing direct benefit of above pledge resources for US users