Network awareness and network as a resource (and its integration with WMS)

Artem Petrosyan (University of Texas at Arlington)
BigPanDA Workshop, CERN, 21 October 2013

PanDA and Networking

- Goal for PanDA
  - Direct integration of networking with PanDA workflow – never attempted before for large scale automated WMS systems
- Why PanDA and networking
  - PanDA is a distributed computing workload management system
  - Data transfer/access is done asynchronously: by DQ2 in ATLAS, PhEDEx in CMS, pandamover/FAX for special cases…
    - Data transfer/access systems can provide first level of network optimizations – PanDA will use these enhancements as available
    - Will also discuss with FTS team at CERN IT in few weeks
  - PanDA relies on networking for workload data transfer/access
    - Higher level of network integration – directly in workflow management
    - Networking is assumed in PanDA – not integrated in workflow

Concept: network as a Resource

- PanDA as workload manager
  - PanDA automatically chooses job execution site
    - Multi-level decision tree – task brokerage, job brokerage, dispatcher
    - Also manages predictive future workflows – at task definition, PD2P
  - Site selection is based on processing and storage requirements
    - Can we use network information in this decision?
    - Can we go even further – network provisioning?
  - Further – network knowledge used for all phases of job cycle?
- Network as resource
  - Optimal site selection should take network capability into account
    - We do this already – but indirectly using job completion metrics
  - Network as a resource should be managed (i.e. provisioning)
    - We also do this crudely – mostly through timeouts, self throttling

Scope of Effort

- Three parallel efforts to integrate networking in PanDA
  - US ATLAS funded – primarily to improve integration with FAX
  - ASCR funded – BigPanDA project, taking PanDA beyond LHC
    - Next Generation Workload Management and Analysis System for Big Data, DOE funded (BNL, U Texas Arlington)
  - ANSE funded – NSF CC-NIE program

PanDA Use Cases

1) Use network information for cloud selection
2) Use network information for FAX brokerage
3) Use network information for job assignment
   - Improve flow of ‘activated’ jobs
   - Better accounting of ‘transferring’ jobs
4) Use network information for PD2P
5) Use network information for site selection
6) Provision circuits for PD2P transfers
7) Provision circuits for input transfers
8) Provision circuits for output transfers

FAX for User analysis

- Goal - reduce waiting time for user jobs (FAX job brokerage)
  - User analysis jobs go to sites with local input data
  - This can lead to long wait times occasionally (PD2P will make more copies eventually to reduce congestion)
  - While nearby sites with good network access may have idle CPU's
  - We could use network information to assign work to 'nearby' sites
- Use cost metric generated with HammerCloud tests initially – treat as 'typical cost' of data transfer between two sites
- Brokerage should use concept of 'nearby' sites
  - Calculate weight based on usual brokerage criteria (availability of CPU…) plus network transfer cost
  - Jobs to be sent to site with best weight – not necessarily the site with local data or with available CPU's

Cloud Selection Plan

- Optimize choice of T1-T2 pairings (cloud selection)
  - In ATLAS, production tasks are assigned to Tier 1's
  - Tier 2's are attached to a Tier 1 cloud for data processing
  - Any T2 may be attached to multiple T1's
  - Currently, operations team makes this assignment manually
  - This could/should be automated using network information
  - For example, each T2 could be assigned to a native cloud by operations team (e.g.. AGLT2 to US), and PanDA will assign to other clouds based on network performance metrics

First Steps

- Started the work on integrating network information into PanDA few months ago, following community discussions
- Step 1: Introduce network information into PanDA
  - Start with slowly varying (static) information from external probes
  - Populate various databases used by PanDA
- Step 2: Use network information for workload management
  - Start with simple use cases that lead to measurable improvements in work flow / user experience
- From PanDA perspective we assume
  - Network measurements are available
  - Network information is reliable
  - We can examine this assumption later – need to start somewhere

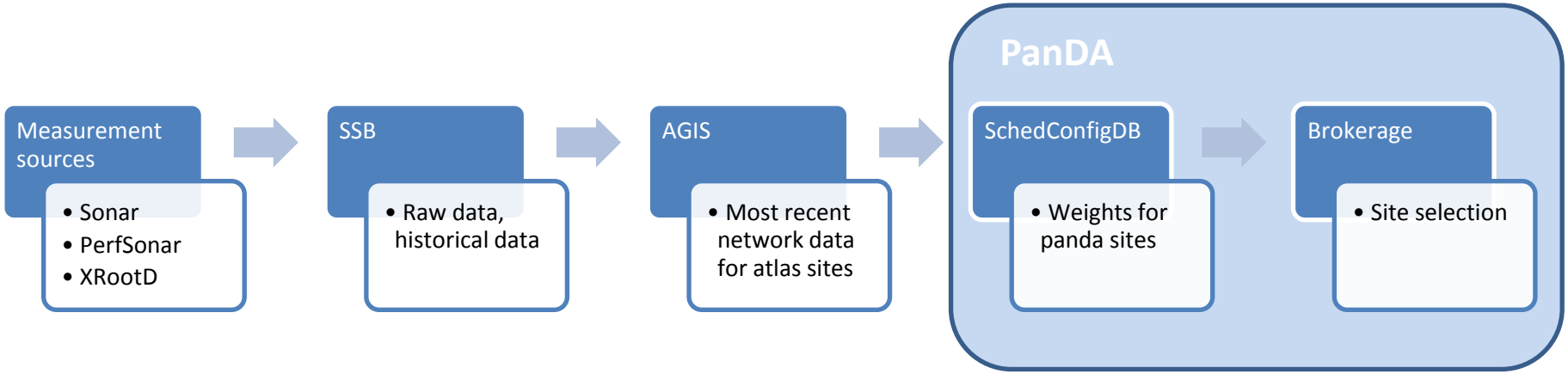
Sources of Network Information

- DDM Sonar measurements
  - ATLAS measures transfer rates for files between Tier 1 and Tier 2 sites (information used for site white/blacklisting)
  - Measurements available for small, medium, and large files
- PerfSonar measurements
  - All WLCG sites are being instrumented with PS boxes
  - US sites are already instrumented and fully monitored
- FAX measurements
  - Read-time for remote files are measured for pairs of sites
  - Standard PanDA test jobs (HammerCloud jobs) are used
- This is not an exclusive list – just a starting point

Data Repositories

- Native data repositories
  - Historical data stored from collectors
  - SSB – site status board for sonar and PS data (currently)
  - HC FAX data is kept independently and uploaded
- AGIS (ATLAS Grid Information System)
  - Most recent/processed data only – updated periodically
  - Pushed via JSON API
- SchedConfigDB
  - Internal Oracle DB used by PanDA for fast access
  - Data updated by extension of standard SchedConfig collector
- All work is currently in ‘dev’ branch

Dataflow



- Data is being transformed
  - Historical to most recent
  - Mb/sec to weights
  - Atlas sites to panda queues

Brokerage mechanism extension

- Site selection basing on network info is an extension of standard PanDA brokerage mechanism:
  - Select several sites basing on preset parameters – standard brokerage
  - Select additional N sites basing on network info - dynamic
- Network weights calculation formula:
  - Throughputs > 50Mb/sec considered “good” and equal 50
  - (Throughput/50)*0.5 – maximum weight should not exceed 0.5 so that we set priority to sites selected basing on configuration parameters
- Site selection module:
  - Returns N best destinations for source S with protocol P with weight higher than T:
    selectNBestNetworkSites(source=‘AGLT2’,protocol=‘xrd’,N=5,threshold=0.4)

Monitoring sources

- SSB

  http://dashb-atlas-ssb.cern.ch/dashboard/request.py/siteview?view=Sonar

  Update by different sources with different frequency
- AGIS throughput source-destination pairs (dev)

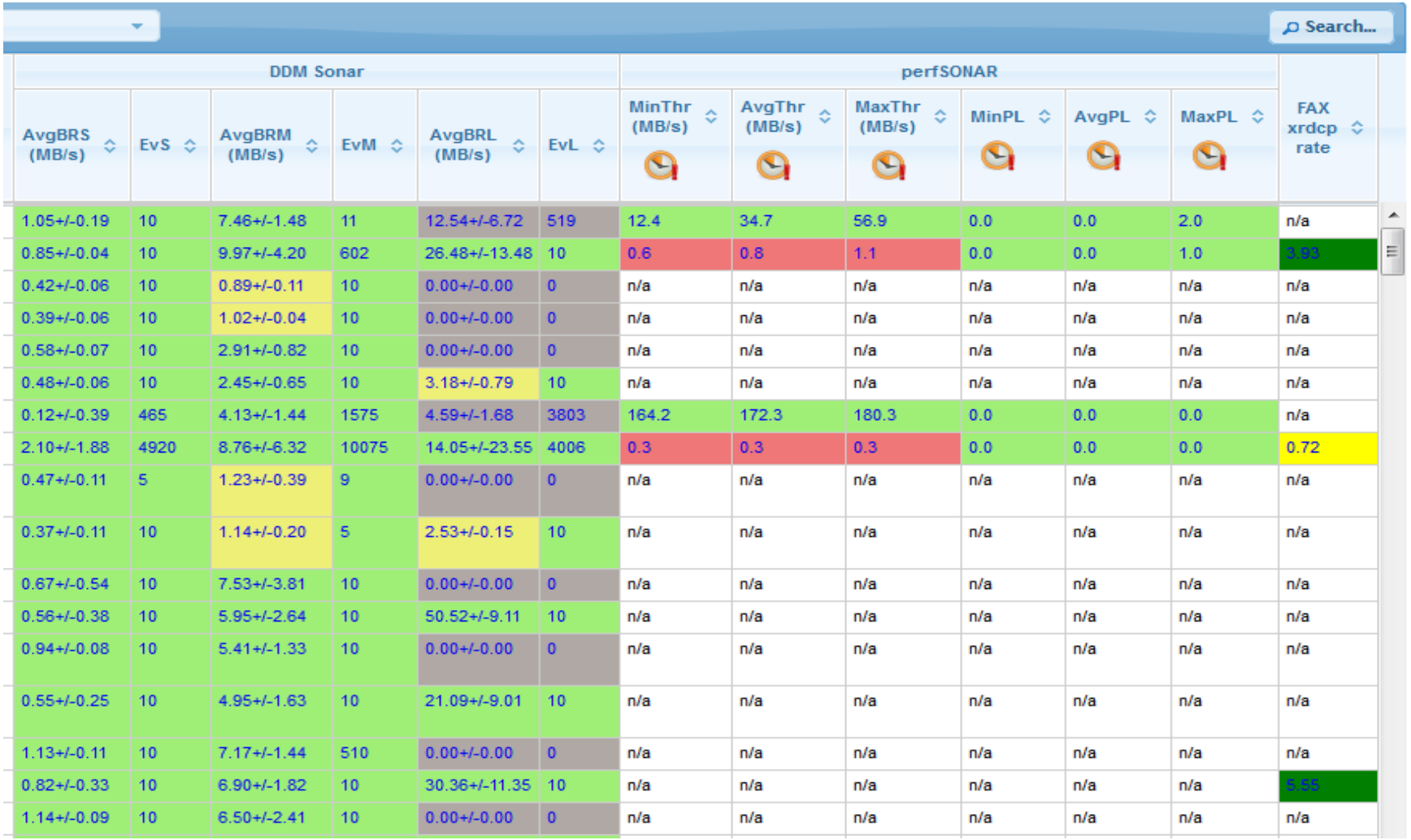
  http://atlas-agis-dev.cern.ch/agis/close_sites/atlassites_links/

  Updated every hour
- Intelligent Networking weights source-destination pairs (dev)

  http://voatlas142.cern.ch/network/sites_matrix/

  Updated every hour, synchronized with AGIS update

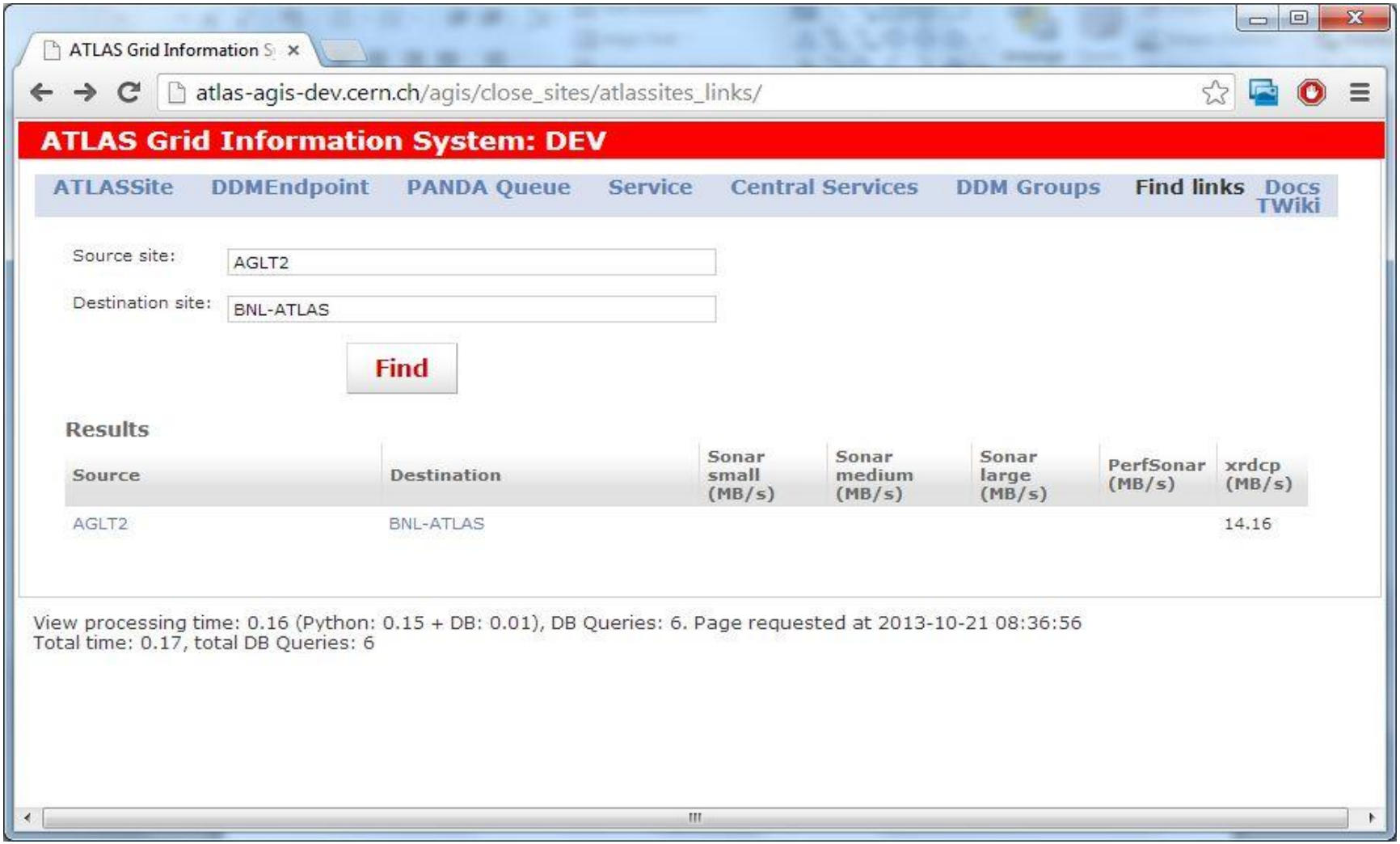
SSB



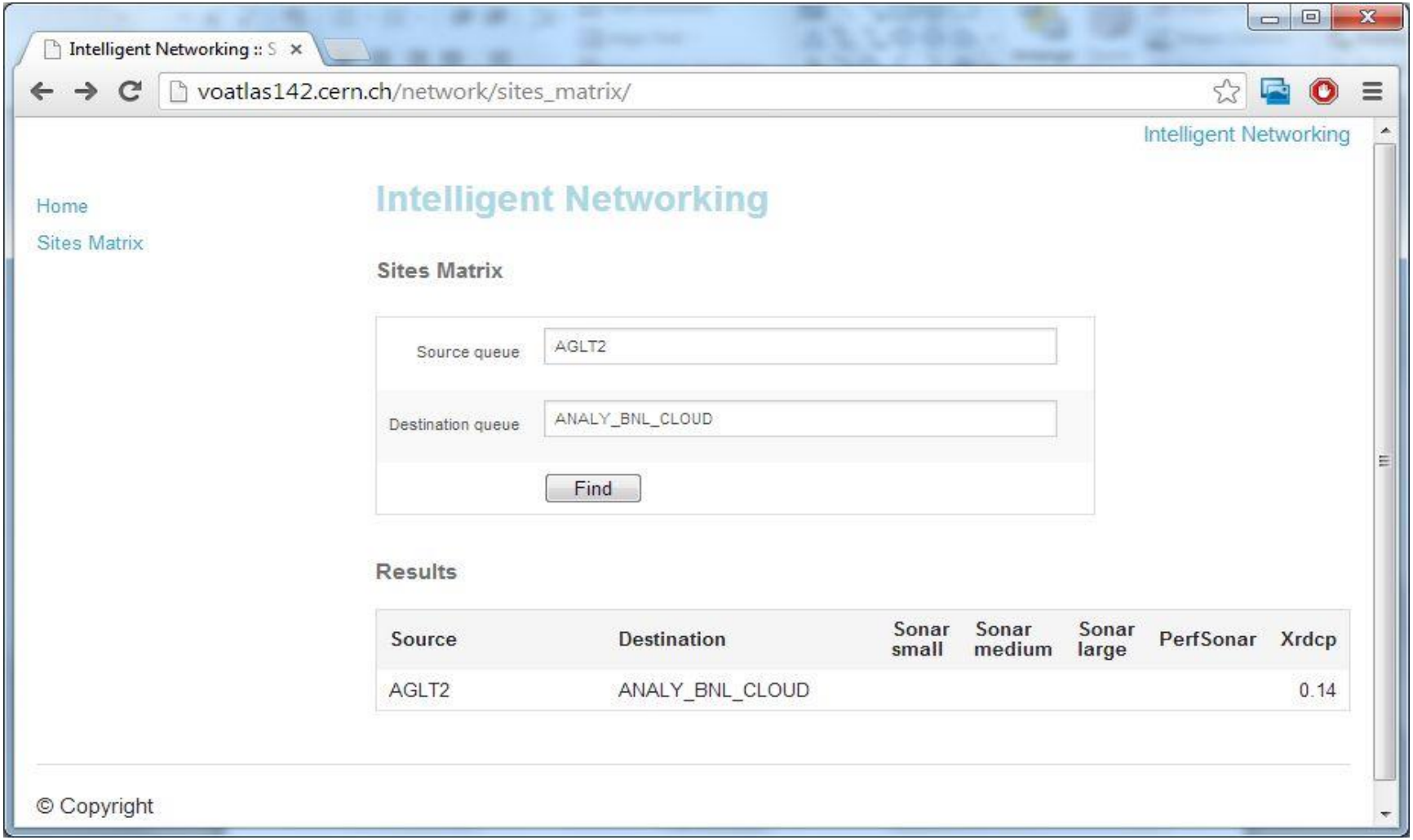
Search...

DDM Sonar						perfSONAR						
AvgBRS (MB/s)	EvS	AvgBRM (MB/s)	EvM	AvgBRL (MB/s)	EvL	MinThr (MB/s)	AvgThr (MB/s)	MaxThr (MB/s)	MinPL	AvgPL	MaxPL	FAX xrdcp rate
1.05+/-0.19	10	7.46+/-1.48	11	12.54+/-6.72	519	12.4	34.7	56.9	0.0	0.0	2.0	n/a
0.85+/-0.04	10	9.97+/-4.20	602	26.48+/-13.48	10	0.6	0.8	1.1	0.0	0.0	1.0	3.93
0.42+/-0.06	10	0.89+/-0.11	10	0.00+/-0.00	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
0.39+/-0.06	10	1.02+/-0.04	10	0.00+/-0.00	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
0.58+/-0.07	10	2.91+/-0.82	10	0.00+/-0.00	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
0.48+/-0.06	10	2.45+/-0.65	10	3.18+/-0.79	10	n/a	n/a	n/a	n/a	n/a	n/a	n/a
0.12+/-0.39	465	4.13+/-1.44	1575	4.59+/-1.68	3803	164.2	172.3	180.3	0.0	0.0	0.0	n/a
2.10+/-1.88	4920	8.76+/-6.32	10075	14.05+/-23.55	4006	0.3	0.3	0.3	0.0	0.0	0.0	0.72
0.47+/-0.11	5	1.23+/-0.39	9	0.00+/-0.00	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
0.37+/-0.11	10	1.14+/-0.20	5	2.53+/-0.15	10	n/a	n/a	n/a	n/a	n/a	n/a	n/a
0.67+/-0.54	10	7.53+/-3.81	10	0.00+/-0.00	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
0.56+/-0.38	10	5.95+/-2.64	10	50.52+/-9.11	10	n/a	n/a	n/a	n/a	n/a	n/a	n/a
0.94+/-0.08	10	5.41+/-1.33	10	0.00+/-0.00	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
0.55+/-0.25	10	4.95+/-1.63	10	21.09+/-9.01	10	n/a	n/a	n/a	n/a	n/a	n/a	n/a
1.13+/-0.11	10	7.17+/-1.44	510	0.00+/-0.00	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
0.82+/-0.33	10	6.90+/-1.82	10	30.36+/-11.35	10	n/a	n/a	n/a	n/a	n/a	n/a	5.55
1.14+/-0.09	10	6.50+/-2.41	10	0.00+/-0.00	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a

AGIS


ATLAS Grid Information S
atlas-agis-dev.cern.ch/agis/close_sites/atlassites_links/
ATLAS Grid Information System: DEV
ATLASSite
DDMEndpoint
PANDA Queue
Service
Central Services
DDM Groups
Find links
Docs
TWiki
Source site:
AGLT2
Destination site:
BNL-ATLAS
Find
Results
Source
Destination
Sonar small (MB/s)
Sonar medium (MB/s)
Sonar large (MB/s)
PerfSonar (MB/s)
xrdcp (MB/s)
AGLT2
BNL-ATLAS
14.16
View processing time: 0.16 (Python: 0.15 + DB: 0.01), DB Queries: 6. Page requested at 2013-10-21 08:36:56
Total time: 0.17, total DB Queries: 6

Intelligent Networking



Short Term Plan

- In the next few months
  - Network information in PanDA – move from development to production branch
  - Implement FAX brokerage algorithm
  - Implement cloud selection algorithm
  - Evaluate algorithm after few months
  - Extend monitoring

Medium Term Plan

- Improve source of information
  - Reliability of network information
  - Dynamic network information
  - Internal measurements

Long Term Plans

- Look at list of topics discussed at various meetings over the past year