Network awareness in PanDA

Artem Petrosyan, JINR

Big Data processing and analysis challenges in mega-science experiments
January 20, 2015, JINR, Dubna, Russia

Why PanDA should care about networking? 1 of 2

- Networking is important for data management in PanDA
  - Distributed workload management systems need to access data both for input and output for processing
  - Data transfers/access is done in multiple steps in PanDA: pilot data movers, direct access, DQ2 for transfers in ATLAS, PhEDEx in CMS, pandamover/FAX, PD2P…
- Future data transfer systems may be optimised for network performance — PanDA will automatically use them
- But network information can be also used directly in workflow management in PanDA at a higher level — first step to try
- We should optimise PanDA workflow for data transfer/access using network information

Why PanDA should care about networking? 2 of 2

- Network performance is important for workflow decisions
  - PanDA automatically chooses job execution site
  - It is multi-level decision tree — task brokerage, job brokerage, dispatcher, policy driven or predictive (PD2P)
  - Site selection can benefit from network information
    - Currently decisions are based on processing and storage requirements
    - We should try to use network information in these decisions
    - Can we go even further — network provisioning?
- Main Goal — network as resource
  - Optimal WMS design should take network capability into account
  - Network as resource should be managed (i.e. provisioned)

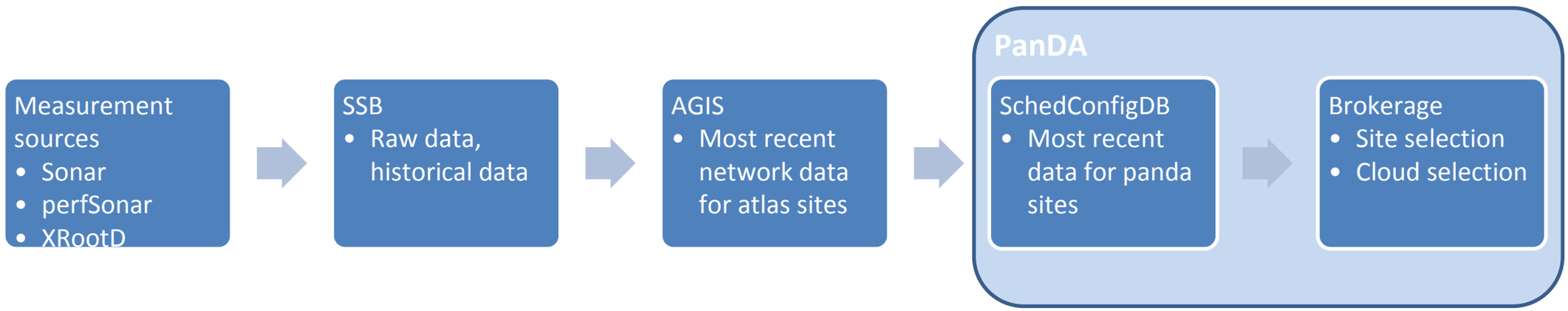
Steps

- Collect network information
- Storage and access
- Using network information
- Using dynamic circuits

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
- FAX measurements
  - Read time for remote files are measured for pairs of sites
  - Standard PanDA test jobs (HammerCloud jobs) are used

Dataflow

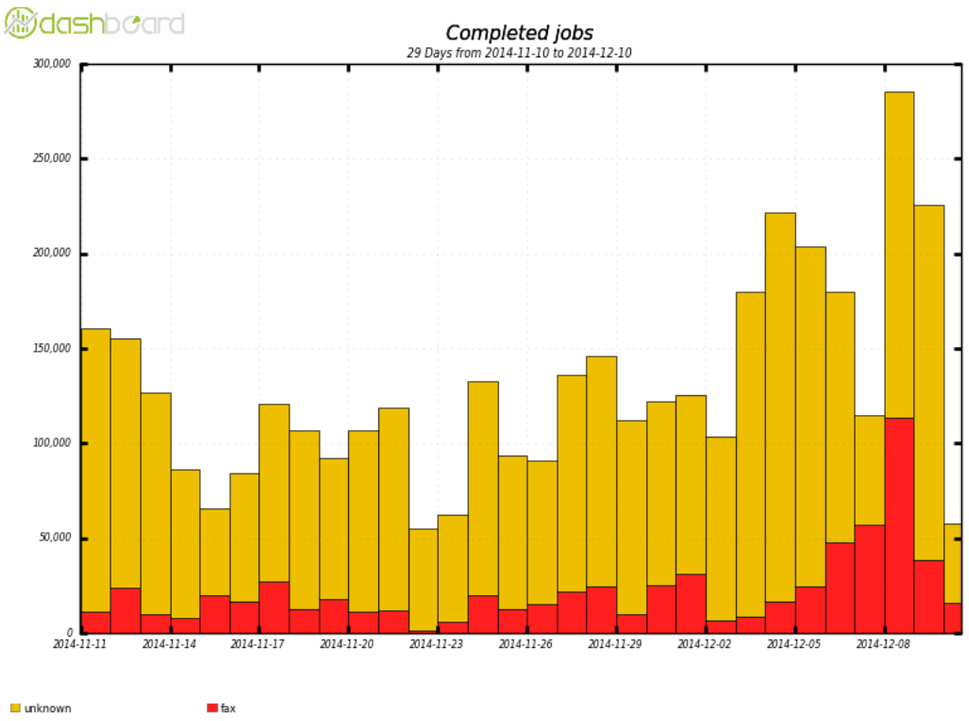


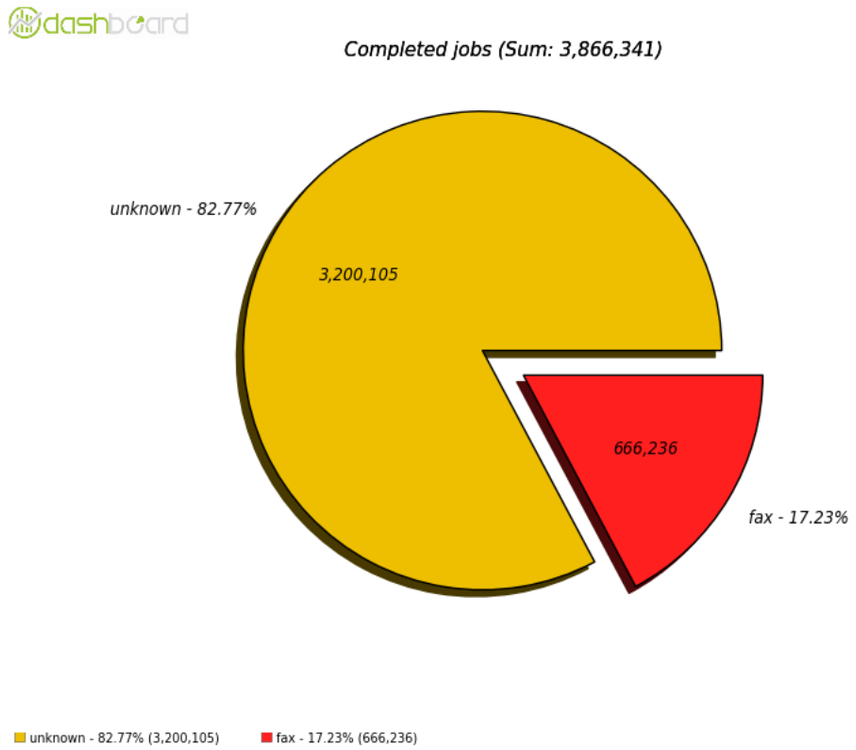
- Data is being transformed
  - Historical to most recent
  - ATLAS sites to PanDA queues

Faster User Analysis through FAX

- First use case for network integration with PanDA
- PanDA brokerage will use concept of nearby sites
  - Calculate weight based on usual brokerage criteria (availability of CPU, release, pilot rate…)
  - Add network transfer cost to brokerage weight
  - Jobs will be sent to the site with best weight — not necessary the site with local data
  - If nearby site has less wait time, access the data through FAX

FAX remote reading stats





- Actual jobs statistics for period of November-December
- https://twiki.cern.ch/twiki/bin/view/AtlasComputing/MonitoringFax#More_Monitoring_links

Conclusions for Case 1

- Network data collection working well
- PanDA brokerage working well
- Achieved goal of reducing waiting time
- Well balanced local vs remote access
- Is being tuned

Cloud Selection

- Second use case for network integration with PanDA
- Optimise choice of T1-T2 pairing (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, and PanDA will assign to other clouds based on network performance metrics

DDM Sonar Data

Legend: Sonar Small|Sonar Medium|Sonar Large

Green, bold: the best T2D site for T1 for each file size

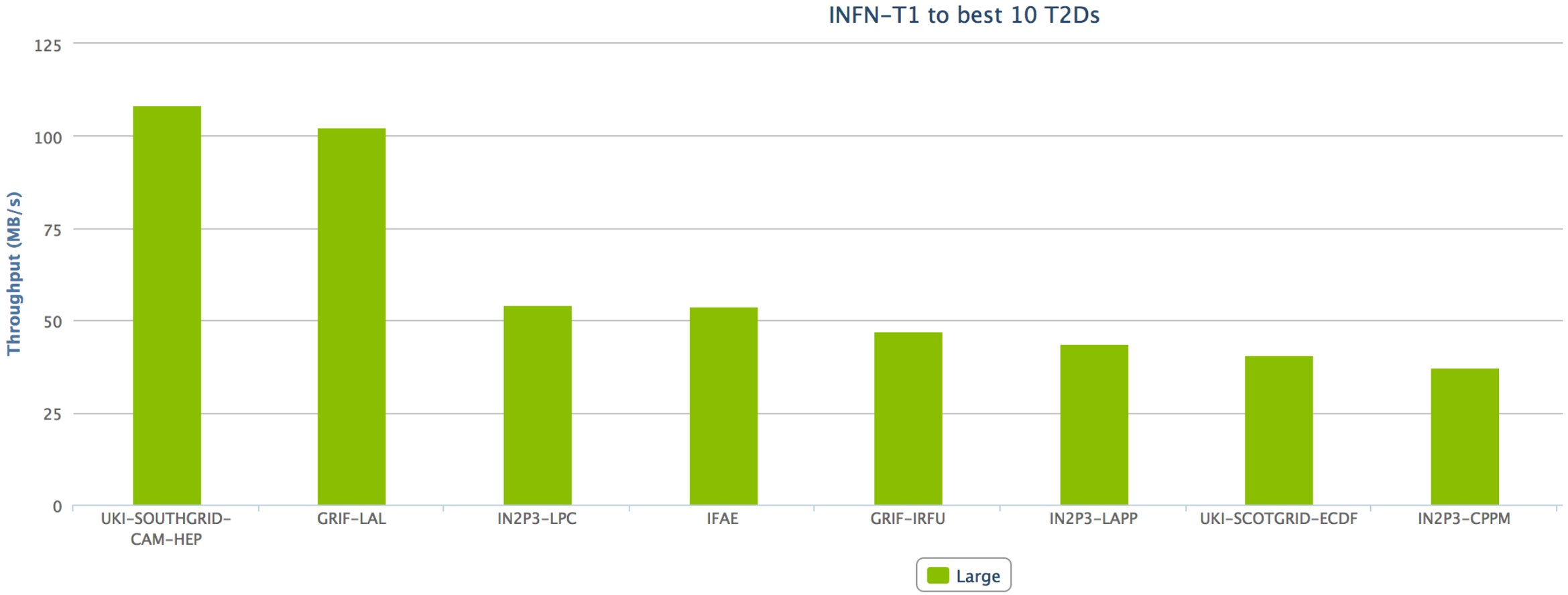
	ARC	BNL	FZK-LCG2	IN2P3-CC	INFN-T1	NDGF-T1
BEIJING-LCG2	0.0|0.0|19.77	0.0|0.0|16.17		same cloud	0.0|1.02|1.06	0.0|0.0|19.77
BostonU		same cloud	0.0|1.05|0.66		0.0|2.09|0.0	
CA-MCGILL-CLUMEQ-T2	0.85|0.0|73.75	0.0|0.0|29.54	1.71|13.04|29.6	0.0|8.13|18.34	0.46|9.84|15.94	0.85|0.0|73.75
CA-SCINET-T2	0.0|0.0|62.91	2.89|20.71|20.83	3.98|17.82|28.05	2.77|24.1|30.93	0.0|0.0|27.25	0.0|0.0|62.91
CA-VICTORIA-WESTGRID-T2	0.0|13.9|33.57	2.37|18.69|25.66	0.0|15.62|23.41	0.55|17.46|16.8	0.0|23.27|26.86	0.0|13.9|33.57
CSCS-LCG2	0.0|0.0|89.36		same cloud			0.0|0.0|89.36
DESY-HH	**6.59**|15.96|50.41	3.7|9.76|23.33	same cloud	3.69|25.82|30.59	5.43|34.61|33.16	**6.59**|15.96|50.41
DESY-ZN	0.0|7.05|11.04	0.46|26.91|39.8	same cloud	0.0|0.0|3.85	6.65|0.0|3.76	0.0|7.05|11.04
GRIF-IRFU	0.0|0.0|57.49	0.0|2.41|4.05	0.11|23.27|30.3	same cloud	**7.75**|0.0|47.09	0.0|0.0|57.49
GRIF-LAL	0.0|60.78|**92.17**	1.22|2.48|9.02	0.0|15.43|52.52	same cloud	6.4|0.0|102.46	0.0|60.78|**92.17**
GRIF-LPNHE	0.0|2.48|25.68	0.88|2.02|3.78	0.0|0.0|13.1	same cloud	0.0|3.47|6.91	0.0|2.48|25.68
GoeGrid	1.38|5.64|11.17	0.47|2.24|9.24	same cloud	0.0|9.82|12.47	0.48|5.83|8.46	1.38|5.64|11.17
GreatLakesT2	2.75|20.43|30.82	same cloud	2.03|15.26|23.71	0.0|7.97|23.59	1.8|17.0|15.64	2.75|20.43|30.82
HarvardU		same cloud				
IFAE	0.0|37.78|39.86	0.0|23.94|13.7	3.91|36.43|56.52	0.0|12.49|30.19	4.91|**54.54**|53.92	0.0|37.78|39.86
IFIC-LCG2	0.0|5.3|5.17	0.0|4.69|6.73	0.0|5.38|8.18	0.0|0.0|15.19	0.0|0.0|11.55	0.0|5.3|5.17

- http://aipanda021.cern.ch/networking/t1tot2d_matrix/

Tier 1 View

Best 10 T2Ds for INFN-T1, large files

	UKI-SOUTHGRID-CAM-HEP	GRIF-LAL	IN2P3-LPC	IFAE	GRIF-IRFU	IN2P3-LAPP	UKI-SCOTGRID-ECDF	IN2P3-CPPM	LRZ-LMU	MidwestT2
INFN-T1	108.28	102.46	54.3	53.92	47.09	43.86	40.51	37.4	34.15	33.87

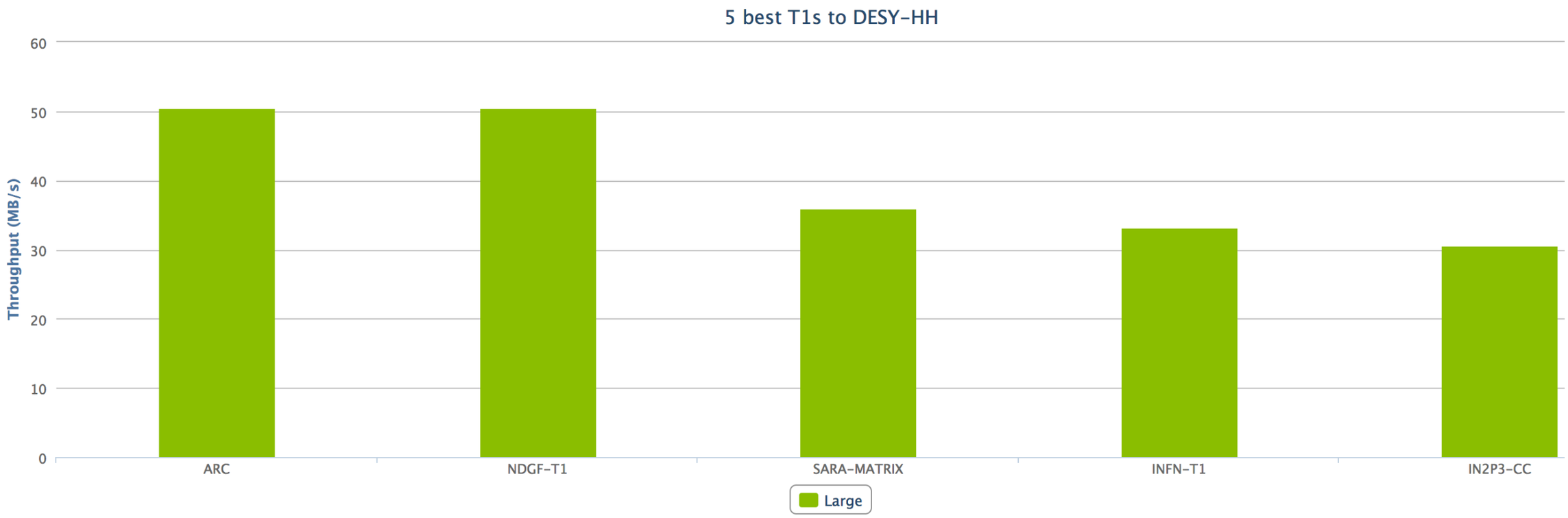


- http://aipanda021.cern.ch/networking/t2dfort1/INFN-T1/

Tier 2 View

Best 5 T1s for DESY-HH, large files, throughput >= 10MB/s

	ARC	NDGF-T1	SARA-MATRIX	INFN-T1	IN2P3-CC
DESY-HH	50.41	50.41	36.04	33.16	30.59



- http://aipanda021.cern.ch/networking/t1fort2d/DESY-HH/

Improving Site Association

Multicloud statistics for queues on DESY-HH

	Current	Suggested	History of suggested
ANALY_DESY-HH	None	ND,NL,IT,FR,ES	
ANALY_DESY-HH_TEST	NL	ND,NL,IT,FR,ES	
DESY-HH-all-prod-CEs	ES,FR,IT,UK	ND,NL,IT,FR,ES	2014-06-24: ND,NL,IT,FR,ES 2014-06-20: ,ND,NL,IT,FR,ES 2014-06-17: ,ND,NL,IT,FR,ES 2014-06-17: ,ND,NL,IT,FR,ES 2014-06-17: ,ND,NL,IT,FR,ES
DESY-HH_TEST	CERN,NL,ES,IT	ND,NL,IT,FR,ES	

- Values of multicloud calculated automatically basing on actual network links between T2 site and T1 sites from another clouds

Conclusion for Case 2

- Working well in real time
- Currently in testing stage
- Multicloud values calculated but updated only for several US sites
- Update of the other sites still in hands of ADC experts

Summary

- We collect, store and use network data, all modules deployed as services on ATLAS infrastructure and work in automatic mode
- First 2 use cases for network integration with PanDA implemented and working well
  - FAX overflow since spring
  - Cloud selection since fall
- Reliability of services, involved into data collecting, delivering and storing is becoming a highly important point
- We need more monitoring options for newly implemented features
- Next step: software defined network integration, i.e. network provisioning
  - Study is ongoing
  - PheDEx circuit booking and DaTRi integration project has already started