PanDA Configurator and Network Aware Brokerage

Fernando Barreiro Megino, Kaushik De,
Tadashi Maeno

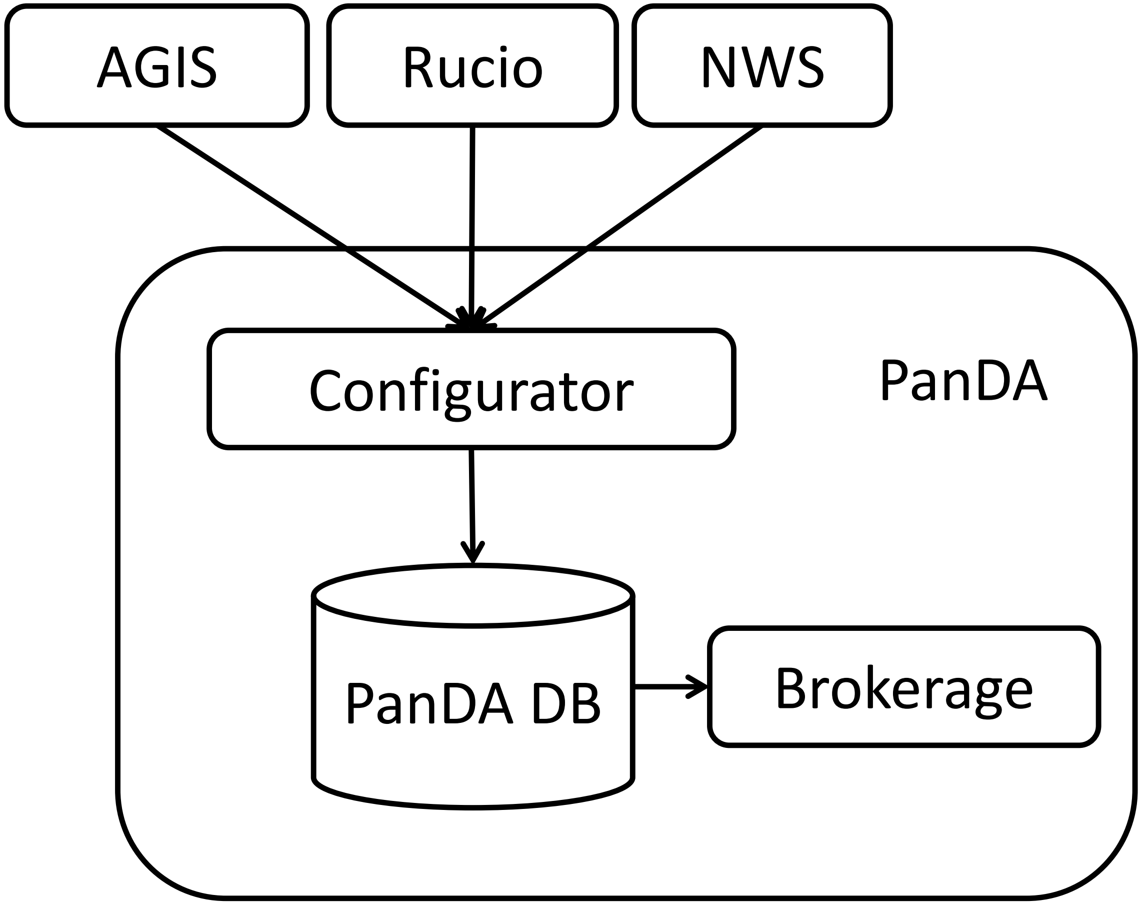
14 March 2015, US ATLAS Distributed Facilities Meeting, Clemson University

UNIVERSITY OF
TEXAS
ARLINGTON

PanDA

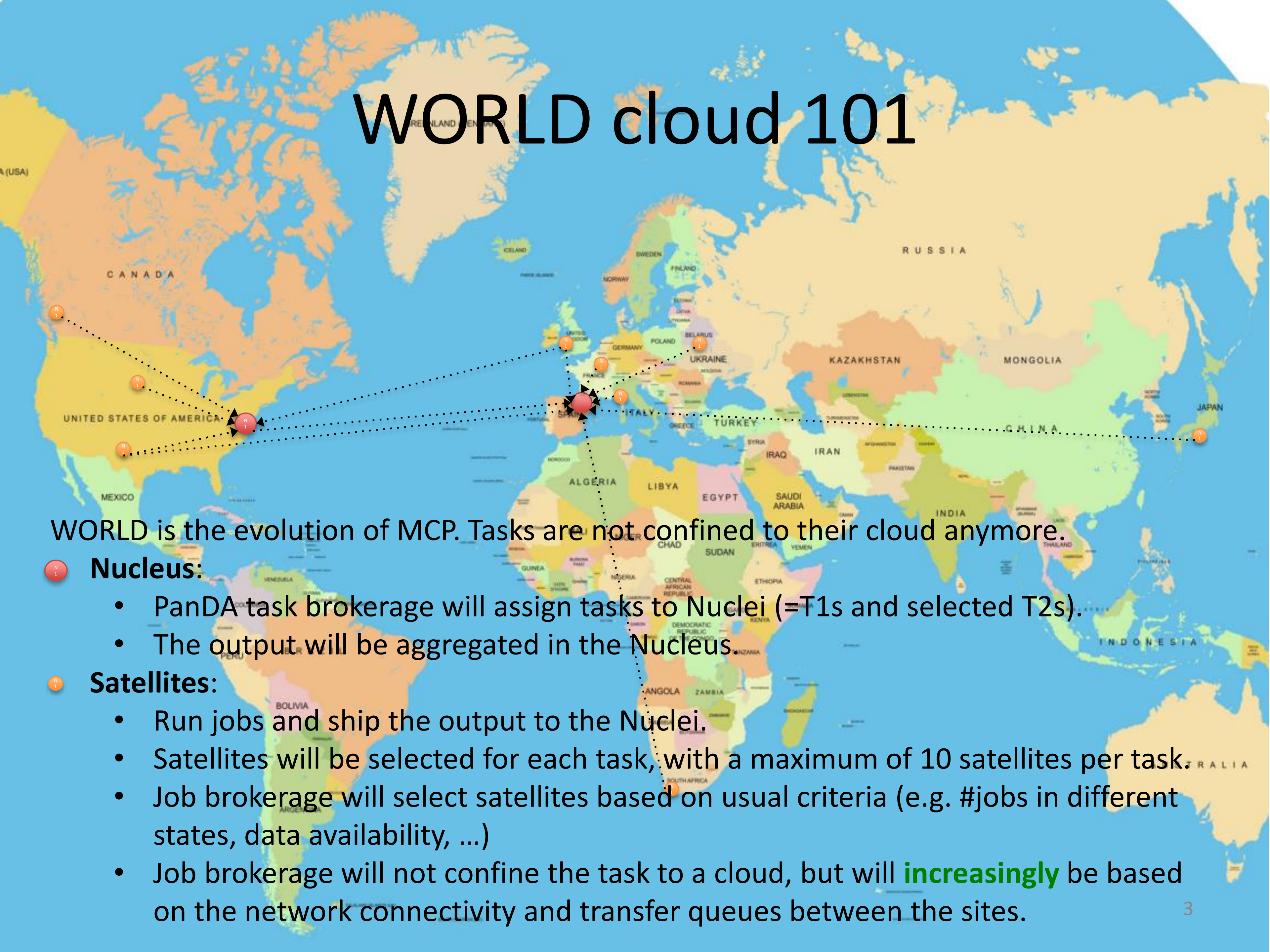
BROOKHAVEN
NATIONAL LABORATORY

Configurator: overview



- PanDA agent running every 30 minutes collecting information useful for brokerage, in particular regarding **WORLD** cloud migration
- Adding progressively new sources as we see the need

WORLD cloud 101



WORLD is the evolution of MCP. Tasks are not confined to their cloud anymore.

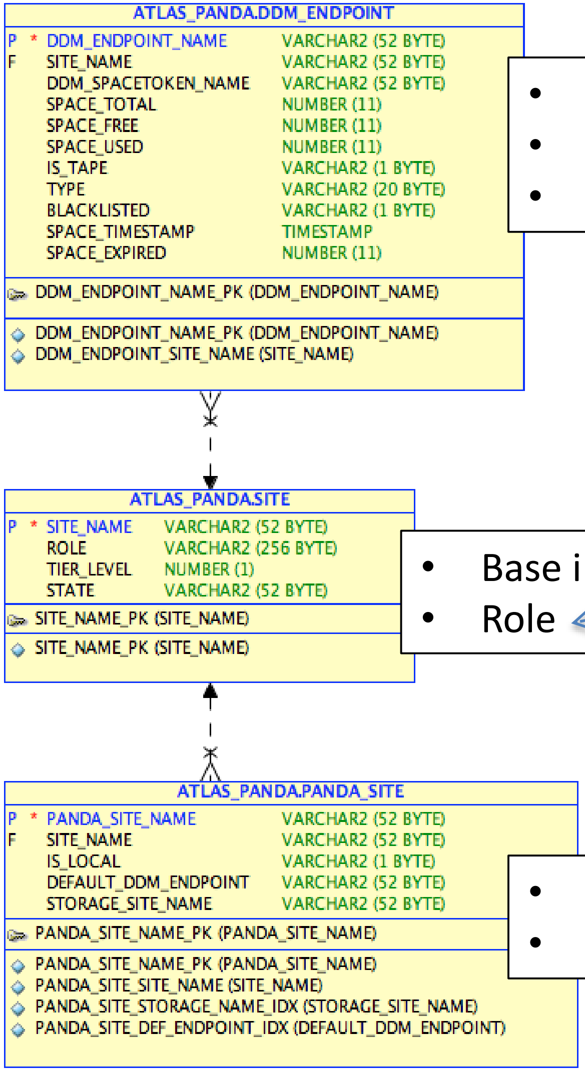
● **Nucleus:**

- PanDA task brokerage will assign tasks to Nuclei (=T1s and selected T2s).
- The output will be aggregated in the Nucleus.

● **Satellites**:

- Run jobs and ship the output to the Nuclei.
- Satellites will be selected for each task, with a maximum of 10 satellites per task.
- Job brokerage will select satellites based on usual criteria (e.g. #jobs in different states, data availability, …)
- Job brokerage will not confine the task to a cloud, but will **increasingly** be based on the network connectivity and transfer queues between the sites.

Configurator: topology data (reminder)



ATLAS_PANDA.DDM_ENDPOINT

P	*	DDM_ENDPOINT_NAME	VARCHAR2 (52 BYTE)
F		SITE_NAME	VARCHAR2 (52 BYTE)
		DDM_SPACETOKEN_NAME	VARCHAR2 (52 BYTE)
		SPACE_TOTAL	NUMBER (11)
		SPACE_FREE	NUMBER (11)
		SPACE_USED	NUMBER (11)
		IS_TAPE	VARCHAR2 (1 BYTE)
		TYPE	VARCHAR2 (20 BYTE)
		BLACKLISTED	VARCHAR2 (1 BYTE)
		SPACE_TIMESTAMP	TIMESTAMP
		SPACE_EXPIRED	NUMBER (11)

DDM_ENDPOINT_NAME_PK (DDM_ENDPOINT_NAME)

DDM_ENDPOINT_NAME_PK (DDM_ENDPOINT_NAME)
DDM_ENDPOINT_SITE_NAME (SITE_NAME)

- Base info
- Space usage
- Downtime info

ATLAS_PANDA.SITE

P	*	SITE_NAME	VARCHAR2 (52 BYTE)
		ROLE	VARCHAR2 (256 BYTE)
		TIER_LEVEL	NUMBER (1)
		STATE	VARCHAR2 (52 BYTE)

SITE_NAME_PK (SITE_NAME)

SITE_NAME_PK (SITE_NAME)

- Base info
- Role

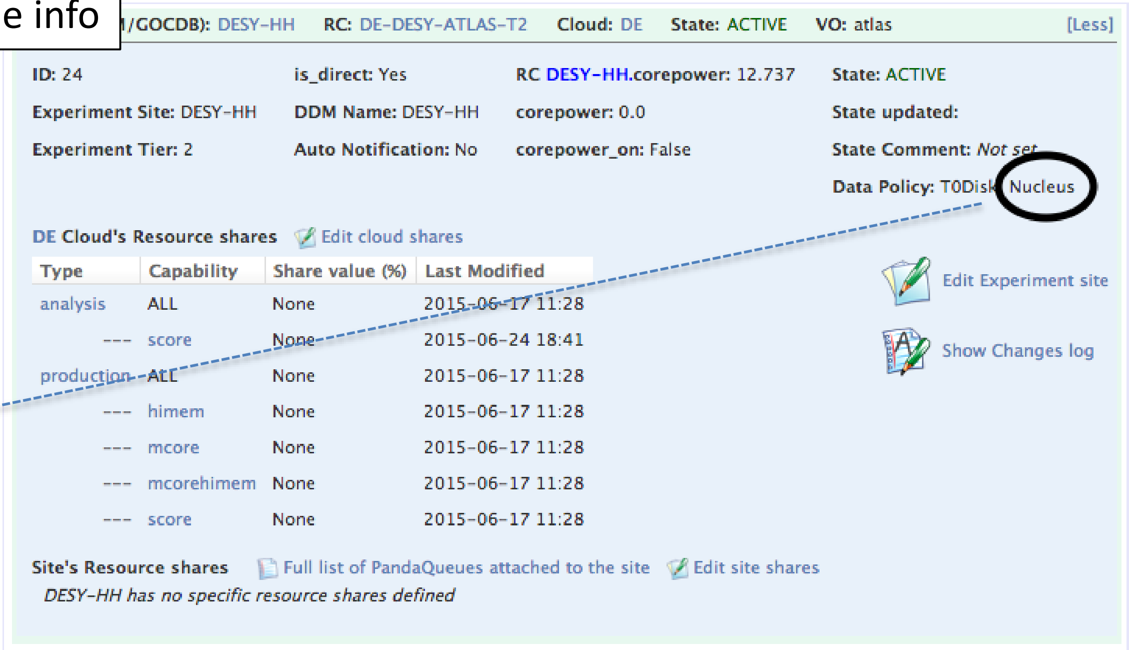
ATLAS_PANDA.PANDA_SITE

P	*	PANDA_SITE_NAME	VARCHAR2 (52 BYTE)
F		SITE_NAME	VARCHAR2 (52 BYTE)
		IS_LOCAL	VARCHAR2 (1 BYTE)
		DEFAULT_DDM_ENDPOINT	VARCHAR2 (52 BYTE)
		STORAGE_SITE_NAME	VARCHAR2 (52 BYTE)

PANDA_SITE_NAME_PK (PANDA_SITE_NAME)

PANDA_SITE_NAME_PK (PANDA_SITE_NAME)
PANDA_SITE_SITE_NAME (SITE_NAME)
PANDA_SITE_STORAGE_NAME_IDX (STORAGE_SITE_NAME)
PANDA_SITE_DEF_ENDPOINT_IDX (DEFAULT_DDM_ENDPOINT)

- Base info
- Locality



/GOCDB): DESY-HH RC: DE-DESY-ATLAS-T2 Cloud: DE State: ACTIVE VO: atlas [Less]

ID: 24
is_direct: Yes
RC DESY-HH.corepower: 12.737
State: ACTIVE
Experiment Site: DESY-HH
DDM Name: DESY-HH
corepower: 0.0
State updated:
Experiment Tier: 2
Auto Notification: No
corepower_on: False
State Comment: *Not set*
Data Policy: T0Disk Nucleus

DE Cloud's Resource shares Edit cloud shares

Type	Capability	Share value (%)	Last Modified
analysis	ALL	None	2015-06-17 11:28
	--- score	None	2015-06-24 18:41
production	ALL	None	2015-06-17 11:28
	--- himem	None	2015-06-17 11:28
	--- mcore	None	2015-06-17 11:28
	--- mcorehimem	None	2015-06-17 11:28
	--- score	None	2015-06-17 11:28

Edit Experiment site
Show Changes log

Site's Resource shares Full list of PandaQueues attached to the site Edit site shares

DESY-HH has no specific resource shares defined

Configurator: static network data

- Configurator agent downloads and processes network information every 30 min from AGIS and NWS. Data is cached in a key-value table in PanDA DB
  - Table structure avoids adding/removing columns every time a new metric appears/disappears
- AGIS closeness matrix: static closeness between each source-destination pair:
  - Value between *1* (good) and *9* (bad) based on the hierarchic cloud model. Examples:
    - `T1 → T1:` *1*
    - `T2 → T1 in same cloud:` *2*
    - …
    - `T2 → T2 in different clouds:` *7*
    - `T3 → T3 in different clouds:` *9*
  - Special values:
    - -1 to blacklist a channel
    - 0 to define a combined site (in progress)

Configurator: dynamic network data

- The Analytics platform contains a lot of raw network information (FTS, FAX, PerfSonar). We are working with the Rucio team to get aggregates per source-destination pair combined with Rucio queue data:
  - #files transferred in last 1 and 6 hours (by activity)
  - #files queued (by activity)

  Available in 1st NWS version

  - Throughput according to FTS, based on 1 week data
  - Throughput according to FAX
  - PerfSonar metrics (latency, packet loss, throughput)

  Available since 2nd NWS version (work in progress)

WORLD cloud: satellite selection in JEDI production job brokerage

1. Filter out candidates with blacklisted AGIS closeness (closeness = -1)
2. Calculate network weight for remaining candidates, combining static and dynamic info

$$NW\_Weight_{static} = 1 + \frac{MAX\_CLOSENESS - closeness}{MAX\_CLOSENESS - MIN\_CLOSENESS}$$

$$NW\_Weight_{dynamic} = 1 + \frac{queued\_bestsource / done6h\_bestsource}{queued\_channel / done6h\_channel}$$

$$\Rightarrow \quad NW\_Weight = 0.7 \times NW\_Weight_{static} + 0.3 \times NW\_Weight_{dynamic}$$

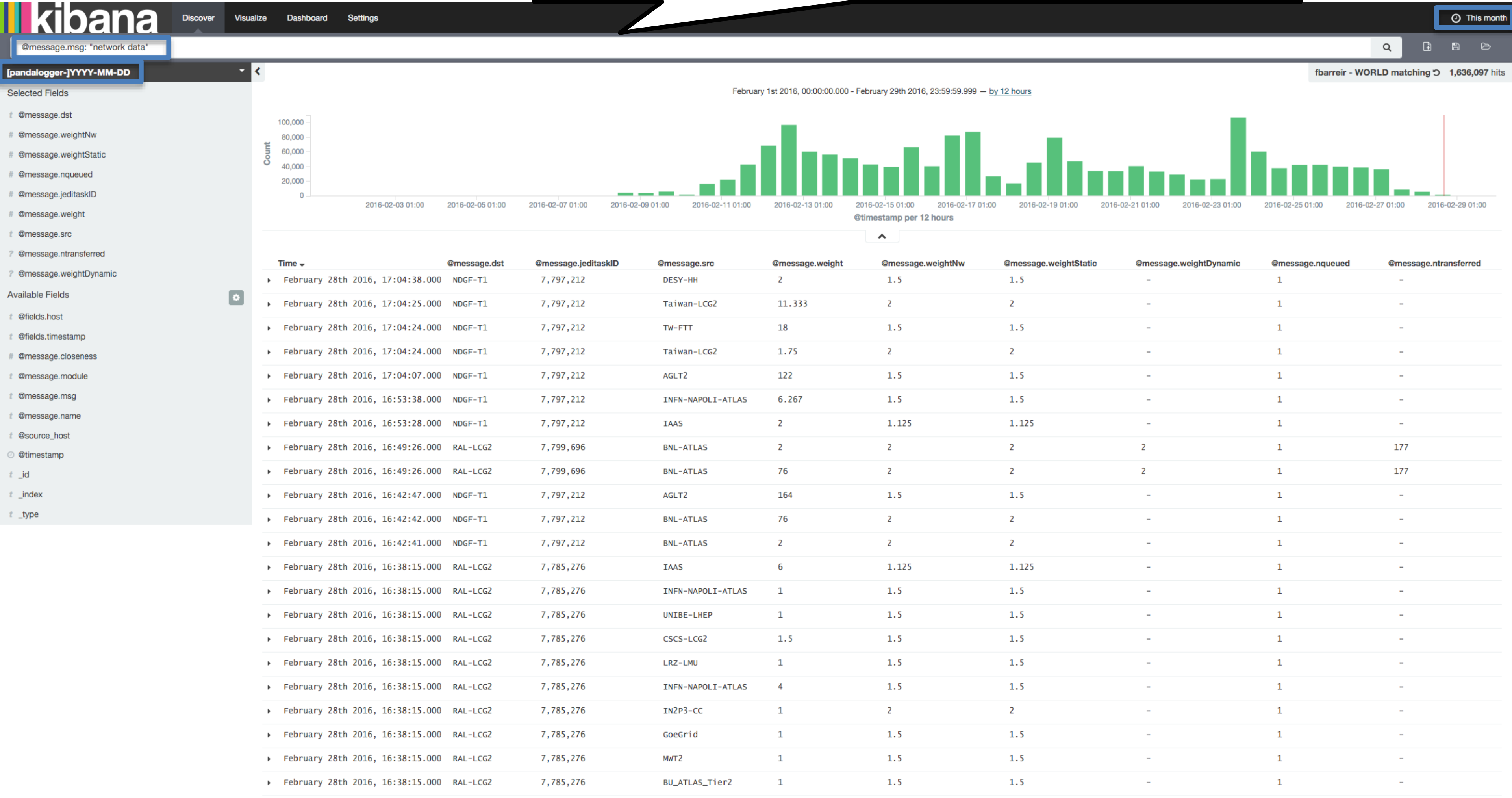
(exceptions apply)

3. Multiply traditional weight (based on data availability, #jobs in different stages, etc.) by network weight

Currently we are running in passive mode and sending the network brokerage decisions to Analytics platform, so we can tune the network model and algorithm

Monitoring NW brokerage: ES messages

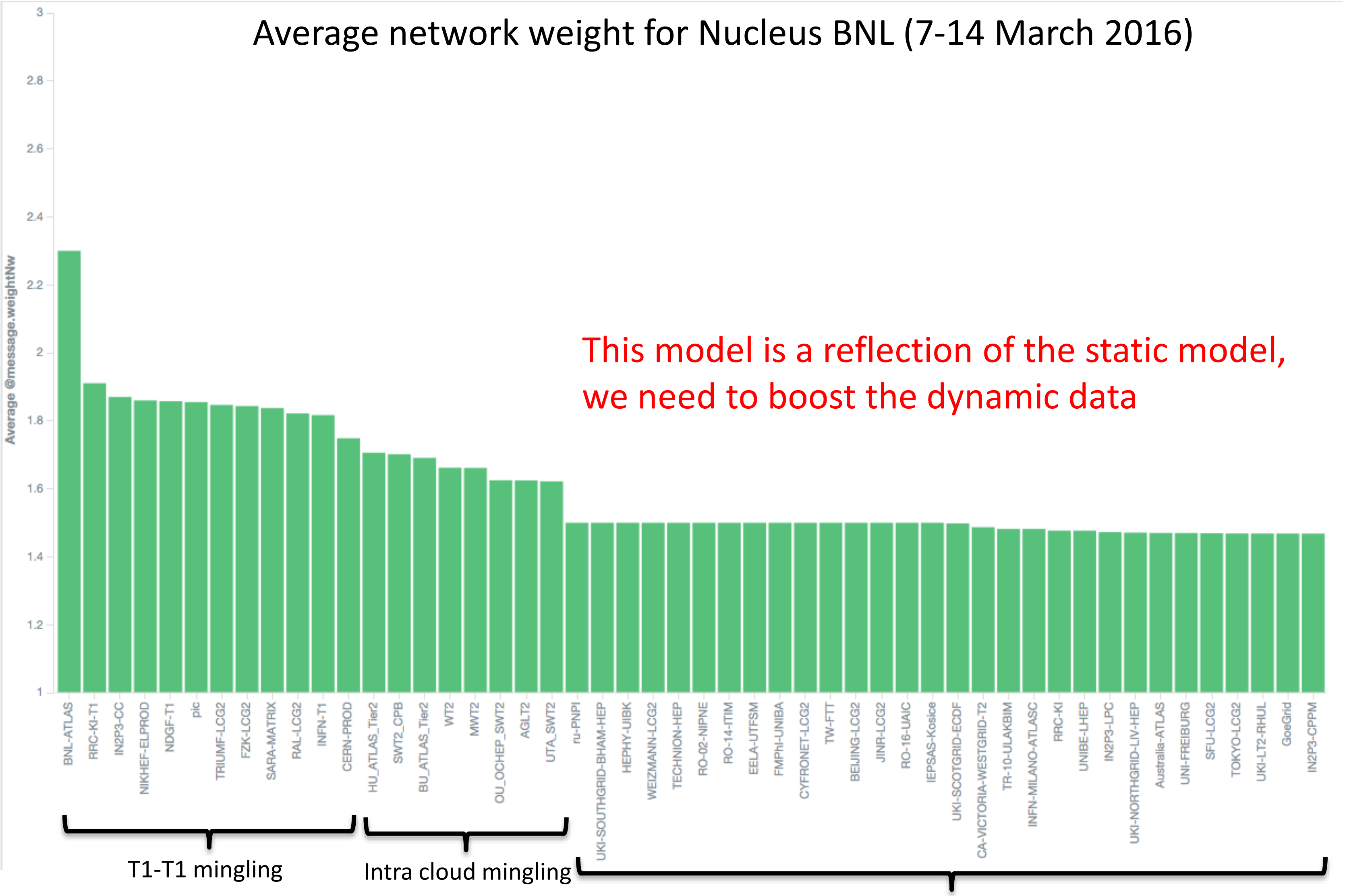
You can explore the data yourself: https://aianalytics01.cern.ch/



Time	@message.dst	@message.jeditaskID	@message.src	@message.weight	@message.weightNw	@message.weightStatic	@message.weightDynamic	@message.nqueued	@message.ntransferred
February 28th 2016, 17:04:38.000	NDGF-T1	7,797,212	DESY-HH	2	1.5	1.5	-	1	-
February 28th 2016, 17:04:25.000	NDGF-T1	7,797,212	Taiwan-LCG2	11.333	2	2	-	1	-
February 28th 2016, 17:04:24.000	NDGF-T1	7,797,212	TW-FTT	18	1.5	1.5	-	1	-
February 28th 2016, 17:04:24.000	NDGF-T1	7,797,212	Taiwan-LCG2	1.75	2	2	-	1	-
February 28th 2016, 17:04:07.000	NDGF-T1	7,797,212	AGLT2	122	1.5	1.5	-	1	-
February 28th 2016, 16:53:38.000	NDGF-T1	7,797,212	INFN-NAPOLI-ATLAS	6.267	1.5	1.5	-	1	-
February 28th 2016, 16:53:28.000	NDGF-T1	7,797,212	IAAS	2	1.125	1.125	-	1	-
February 28th 2016, 16:49:26.000	RAL-LCG2	7,799,696	BNL-ATLAS	2	2	2	2	1	177
February 28th 2016, 16:49:26.000	RAL-LCG2	7,799,696	BNL-ATLAS	76	2	2	2	1	177
February 28th 2016, 16:42:47.000	NDGF-T1	7,797,212	AGLT2	164	1.5	1.5	-	1	-
February 28th 2016, 16:42:42.000	NDGF-T1	7,797,212	BNL-ATLAS	76	2	2	-	1	-
February 28th 2016, 16:42:41.000	NDGF-T1	7,797,212	BNL-ATLAS	2	2	2	-	1	-
February 28th 2016, 16:38:15.000	RAL-LCG2	7,785,276	IAAS	6	1.125	1.125	-	1	-
February 28th 2016, 16:38:15.000	RAL-LCG2	7,785,276	INFN-NAPOLI-ATLAS	1	1.5	1.5	-	1	-
February 28th 2016, 16:38:15.000	RAL-LCG2	7,785,276	UNIBE-LHEP	1	1.5	1.5	-	1	-
February 28th 2016, 16:38:15.000	RAL-LCG2	7,785,276	CSCS-LCG2	1.5	1.5	1.5	-	1	-
February 28th 2016, 16:38:15.000	RAL-LCG2	7,785,276	LRZ-LMU	1	1.5	1.5	-	1	-
February 28th 2016, 16:38:15.000	RAL-LCG2	7,785,276	INFN-NAPOLI-ATLAS	4	1.5	1.5	-	1	-
February 28th 2016, 16:38:15.000	RAL-LCG2	7,785,276	IN2P3-CC	1	2	2	-	1	-
February 28th 2016, 16:38:15.000	RAL-LCG2	7,785,276	GoeGrid	1	1.5	1.5	-	1	-
February 28th 2016, 16:38:15.000	RAL-LCG2	7,785,276	MWT2	1	1.5	1.5	-	1	-
February 28th 2016, 16:38:15.000	RAL-LCG2	7,785,276	BU_ATLAS_Tier2	1	1.5	1.5	-	1	-

There is no network information for most links over the last 6h, so static closeness will prevail. We should also include metrics that cover longer time periods

Example: Favorite satellites for nucleus BNL



Average network weight for Nucleus BNL (7-14 March 2016)

This model is a reflection of the static model, we need to boost the dynamic data

T1-T1 mingling

Intra cloud mingling

Inter cloud

Observations (1)

- This work is fairly new – we started after the Sitges TIM in December
  - We have improved considerably the data transmission model and are starting to use network data for a very broad case in PanDA
  - **We have powerful building blocks, but we need to get them operational and tune the algorithms**
  - We need a good data analysis of all the available information and recommendations (WIP by Mario et al.)

- AGIS closeness is too much of a reflection of the MONARC model
  - IMHO the data should reflect the reality, not a theoretical, obsolete model
    - Simple, semi-static classification?

- Aggregation of data from NWS is work in progress

Observations (2)

- Verify, activate, improve algorithm for nuclei-satellite matching
  - Start using the second version of aggregated data, containing more info than nqueued and ntransferred:
    - FTS Mbps over last week to have dynamic data over longer period
    - PerfSonar data
  - Boost the dynamic data
  - Analyze if the network weight should be stronger
- Extend to other Rucio and FAX use cases

Other possible network brokerage use cases

- Network weight for input file transfers (AKA Rodney Walker’s case):
  - Input data is in site A and B, but sites are busy
  - Sites C and D are free, but don’t have input data

  → Consider network for the brokerage
- FAX network weight for Event Service jobs
- **BUT:** Both cases require that PanDA and the respective DM system (Rucio, FAX) follow a similar source selection logic
  - Otherwise PanDA might be taking useless decisions
- Review Overflow jobs?