Distributing Postgres with Postgres-XL for Very Large Astronomical Databases

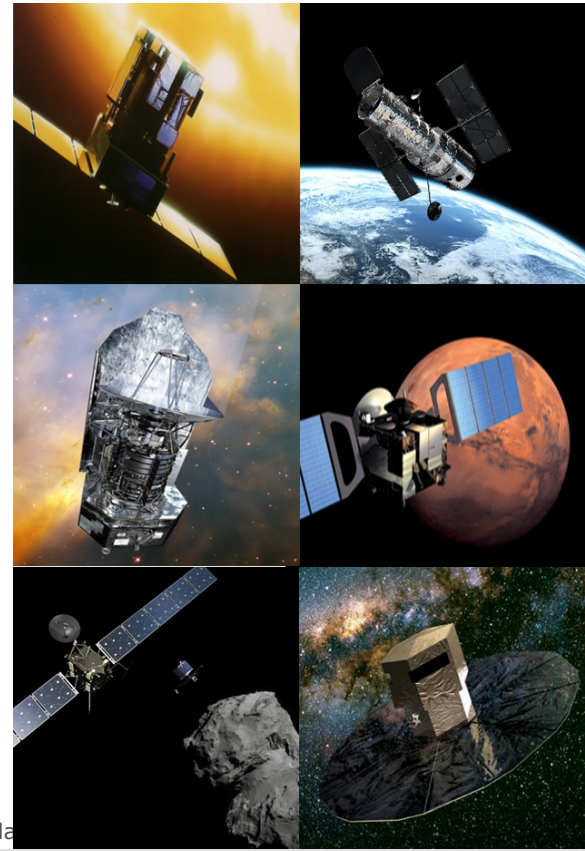
Pilar de Teodoro
European Space Astronomy Centre, Madrid, Spain

PGCONF.US 2018, 04/19/2018

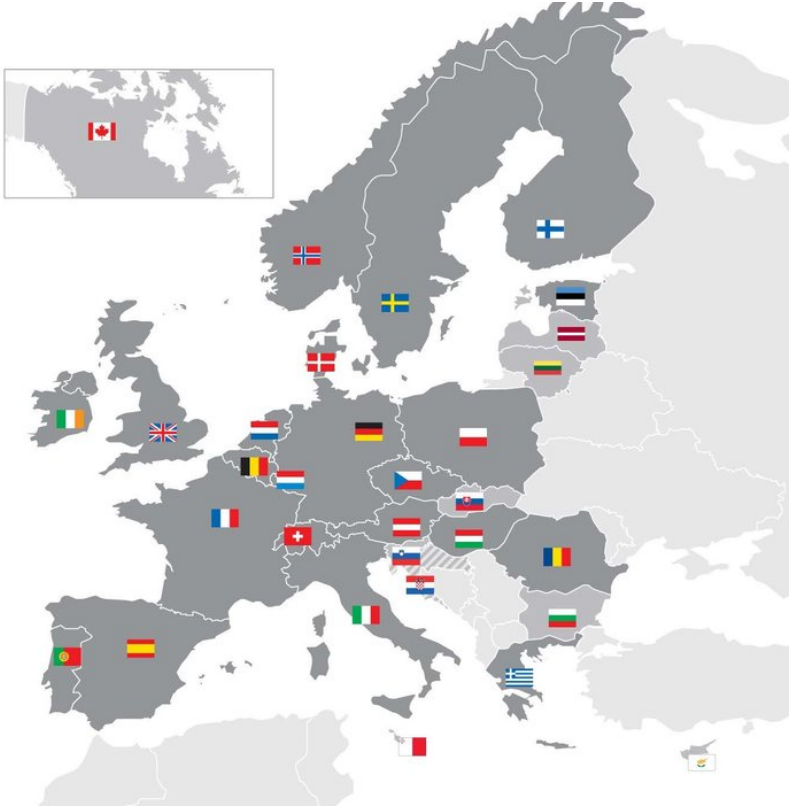
@pteodoro

Outlines

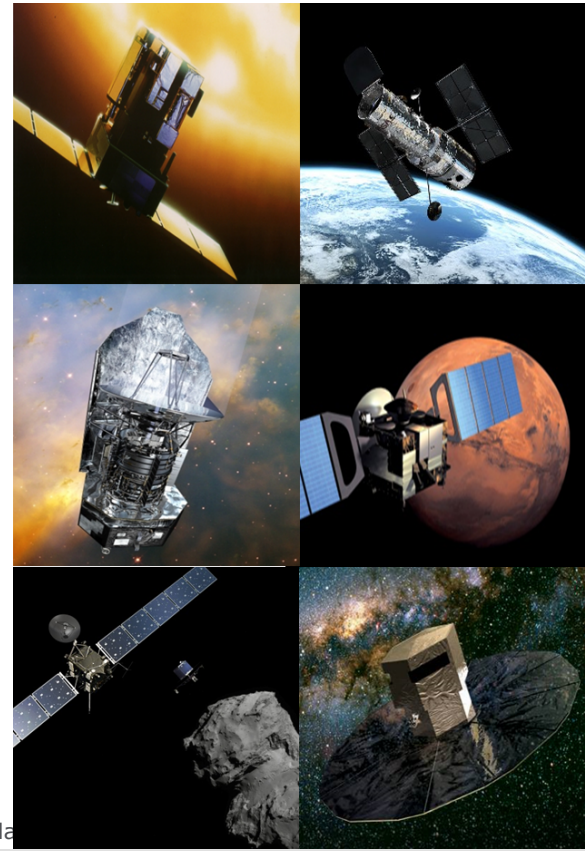
- What is ESA and ESDC
- Why postgres-XL
- Our architecture
- Tests performed
- Issues found
- Lessons learnt
- Next steps



European Space Agency



ESAC Science Data Centre
The Digital Library of the Universe



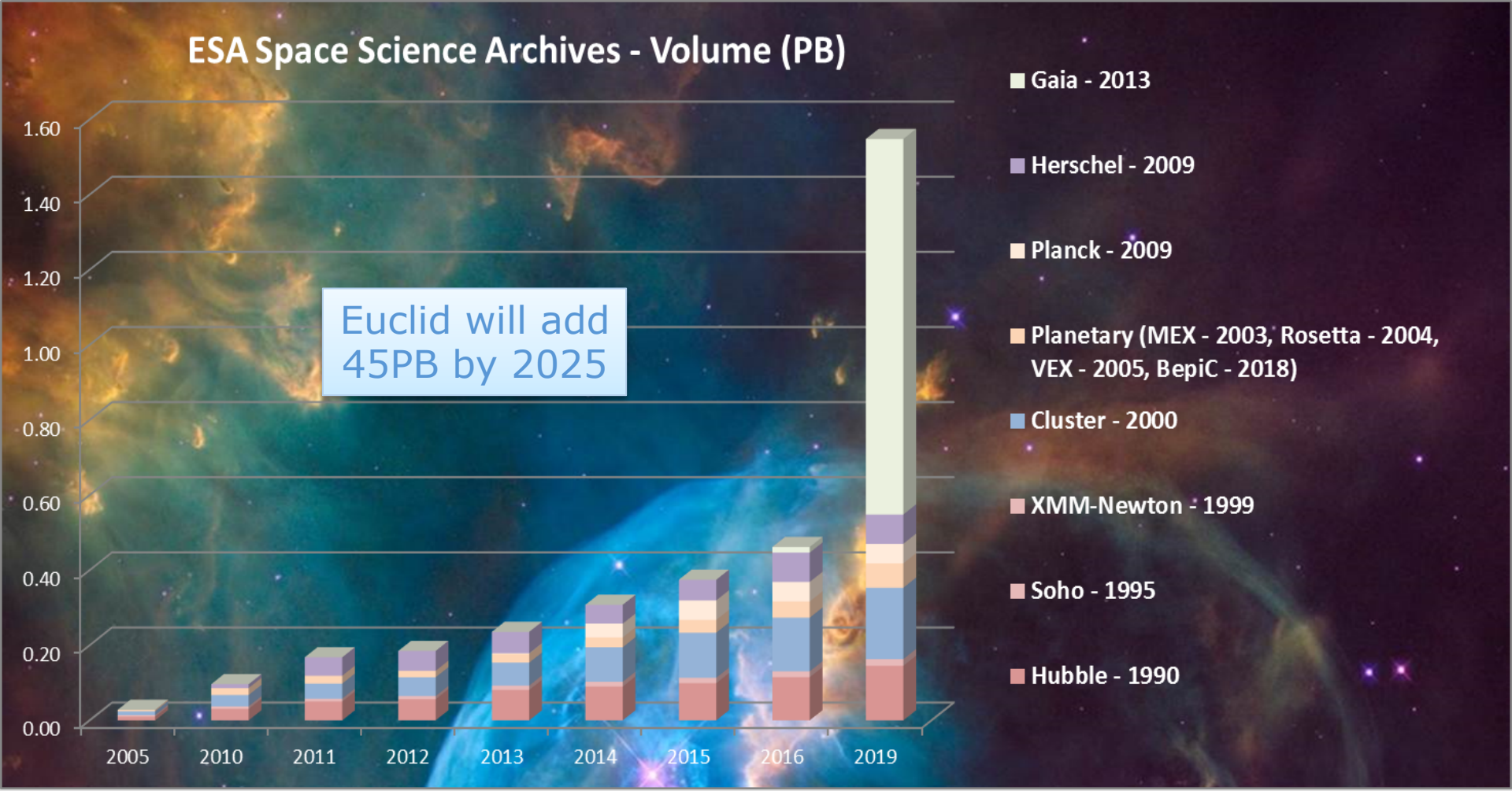
At ESA's European Space Astronomy Centre near Madrid

Science Archives from ~20 space missions:

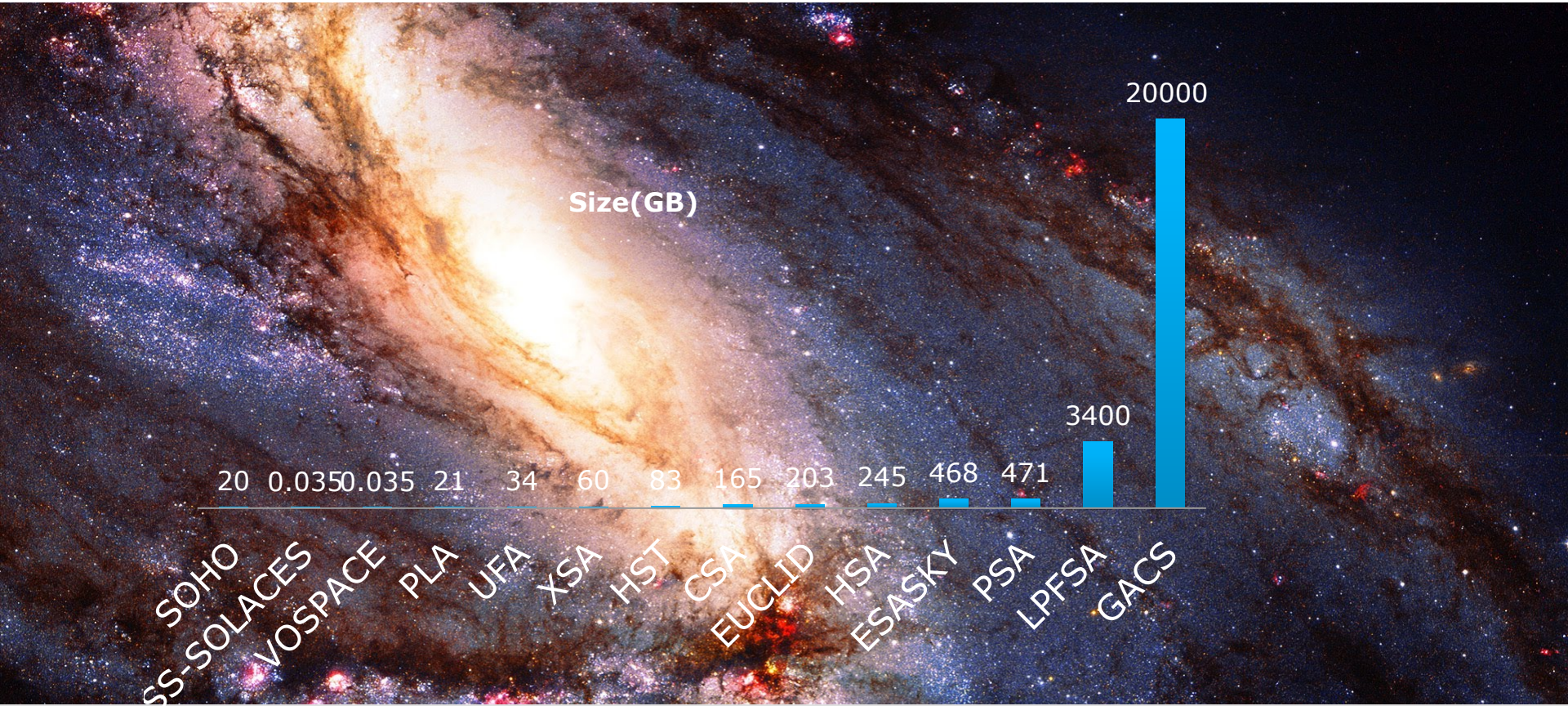
- Astronomy, Planetary, Heliophysics
- From all phases (development, operations, post-ops, legacy)
- http://archives.esac.esa.int/

Different Users:

- Scientific Community (public access)
- Instrument teams and observers (controlled access)
- Science Operations Team (privileged access)


ESA Space Science Archives - Volume (PB)
1.60
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
2005
2010
2011
2012
2013
2014
2015
2016
2019
Euclid will add 45PB by 2025
Gaia - 2013
Herschel - 2009
Planck - 2009
Planetary (MEX - 2003, Rosetta - 2004, VEX - 2005, BepiC - 2018)
Cluster - 2000
XMM-Newton - 1999
Soho - 1995
Hubble - 1990

Databases Size at ESDC (October 2017)



→ ESA'S FLEET ACROSS THE SPECTRUM

lisa pathfinder
Testing the technology for gravitational wave detection (2015-2017)

iso
Chemical analysis of celestial objects (1995-1998)

hipparcos
The first astrometry satellite (1989-1993)

gaia
Surveying a billion stars of exoplanets (2013-)

iue
Analysing ultraviolet light from stars (1978-1996)

cheops
Sizing and first characterisation

xmm-newton
Seeing deeply into the hot and violent Universe (1999-)

herschel
Unveiling the cool and dusty Universe (2009-2013)

Databases at ESDC

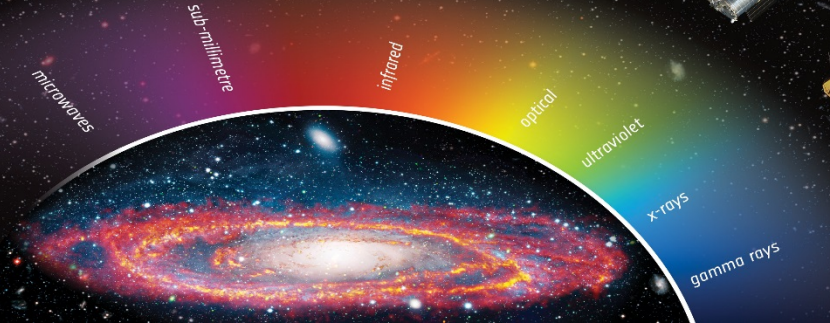
of the visible Universe (1990-)

integral
Seeking out the extremes of the Universe (2002-)

planck
Looking back at the dawn of time (2009-2013)

exosat
X-ray survey of high-energy phenomena (1983-1986)

cos-b
Surveying the high energy Galaxy (1975-1982)



Database systems at ESDC

Most archives: **PostgreSQL**: 9.0 to 10.0

- Open source
- Big community
- Spherical queries (pg_sphere, q3c, pg_healpix, postgis)

ISO: **Sybase**

Euclid DPS: **Oracle** -> **PostgreSQL?**

ORACLE®

ESASky, PSA (some functionalities): **ElasticSearch**

Import from Files, MySQL…

Plan for Scale-out Tests

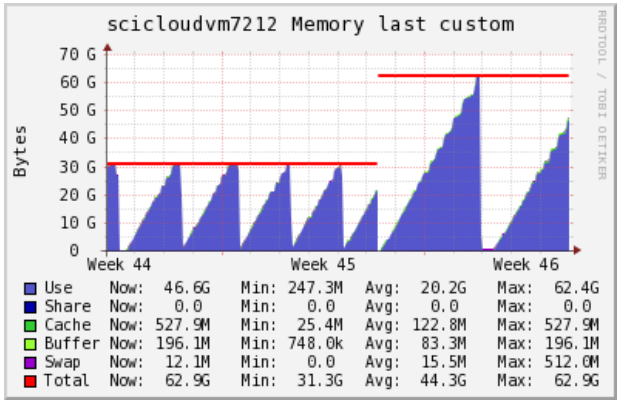
Objective: Big Datasets queried in efficient times (Requirements). Ingestion performance.

Pre-Requisites: Resources, data to ingest (Catalogues, Relational tables)

Comparison: ingestion time, retrieving time, administration, problems arisen.

Criteria:

- Performance
- High availability
- Administration
- Support



PostgreSQL Scale-out Solutions

Postgres-XL by 2nd Quadrant

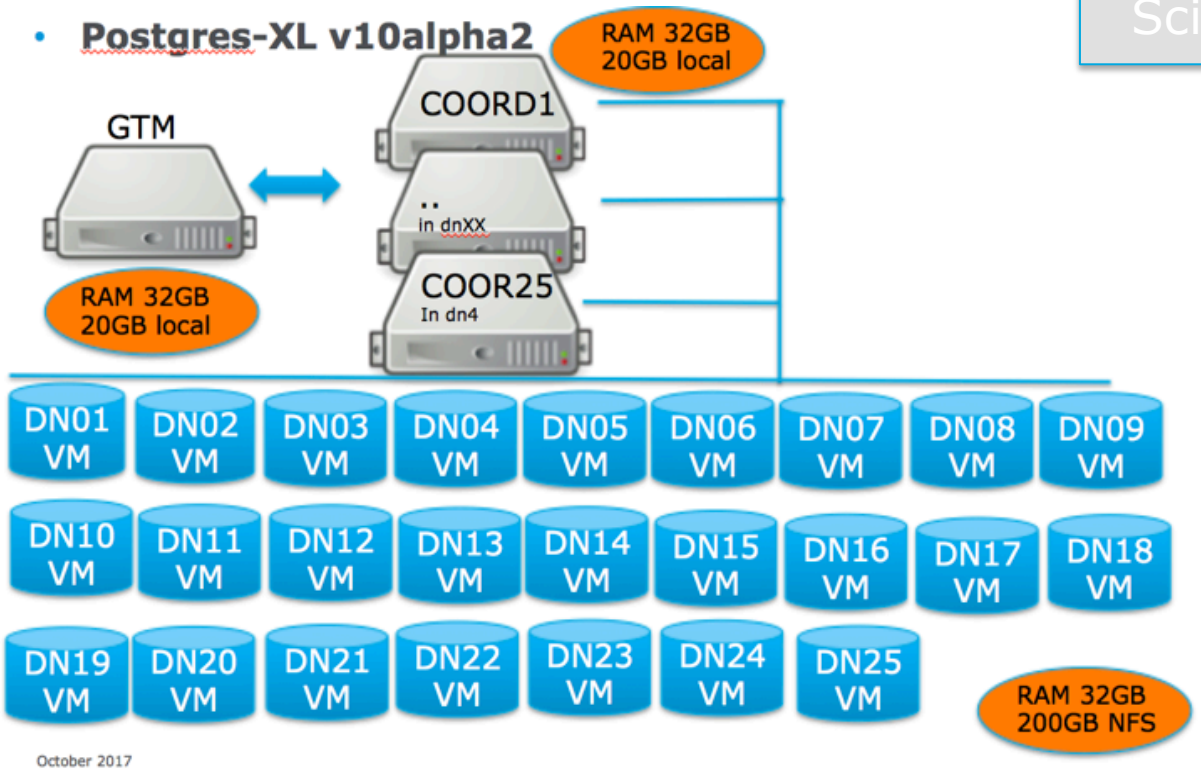
- Postgres-10 alpha2 for Euclid

Postgres-XL by 2nd Quadrant

- Postgres-9.5.2 for Gaia

Postgres-XL Architecture for Testing

Scicloud-27 VMs

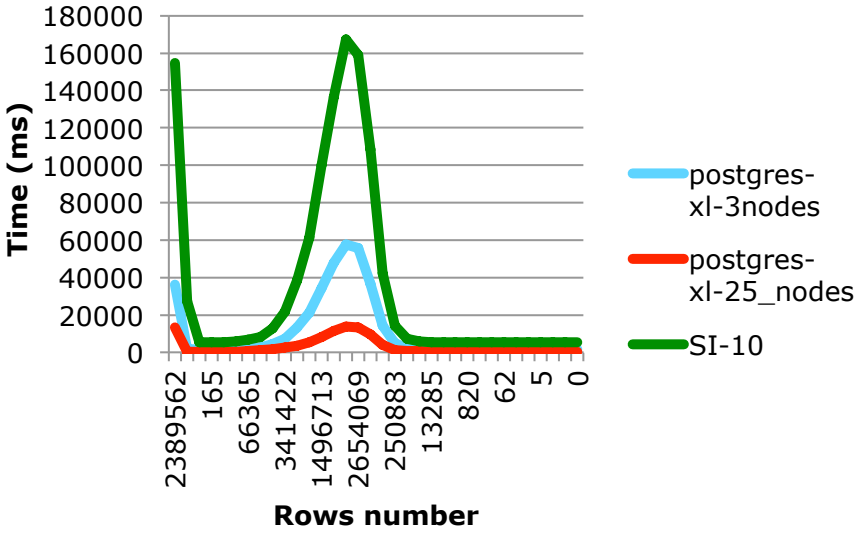


Simple Query Tests

On Kids catalogue:

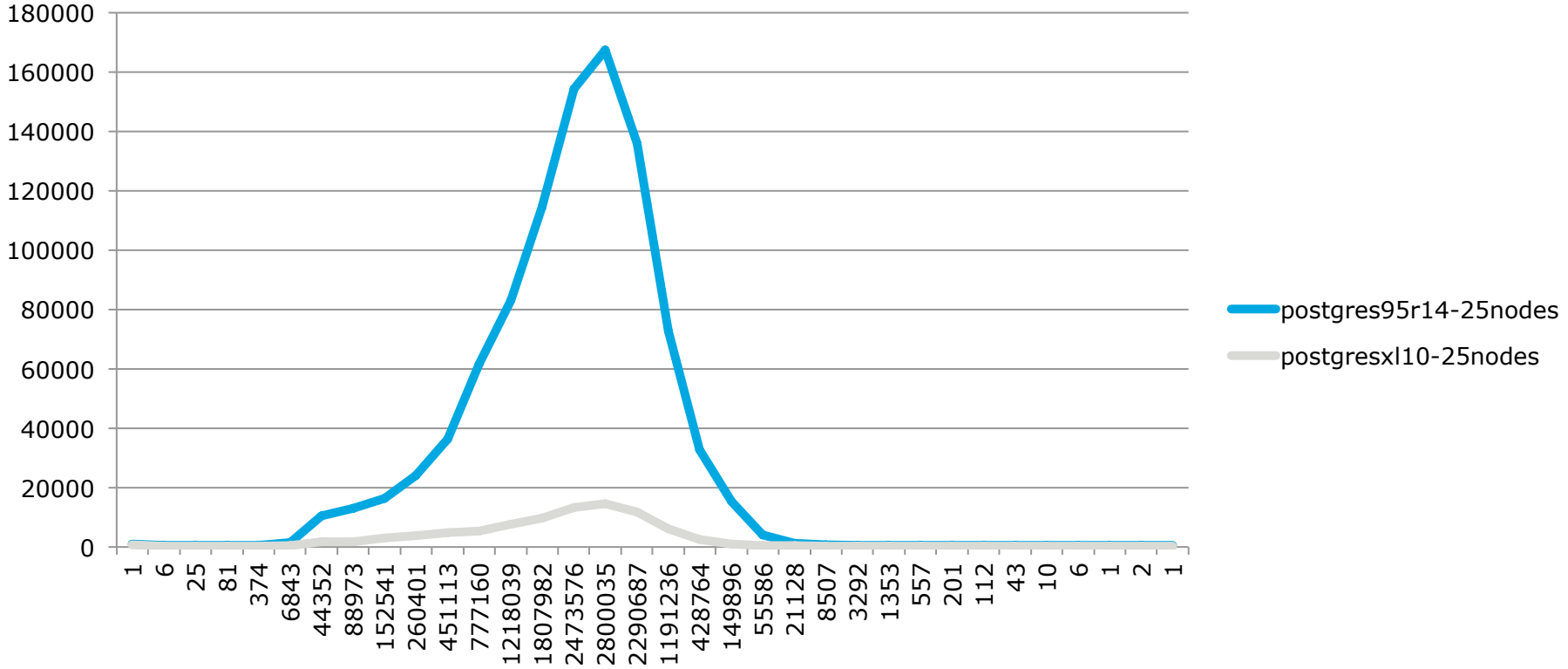
- 16M rows
- 12GB total
- Comparison for Single instance(SI) on postgres v10 vs postgres-xl running on 3 nodes and postgres-xl running on 25 nodes. Time grows with the number of rows.
- Filter: mag_auto_i_trunc column so values from 10..43 with distribution by hash:

Filter	Rows	Filter	Rows
10	238956	27	730583
11	1	28	250883
12	4	29	90818
13	16	30	34346
14	161	31	13285
15	2193	32	5320
16	6636	33	2149
17	11656	34	820
18	19918	35	361
19	34142	36	139
20	60052	37	62
21	97729	38	27
22	149671	39	4
23	214430	40	5
24	272247	41	3
25	265406	42	1
26	176139	43	0

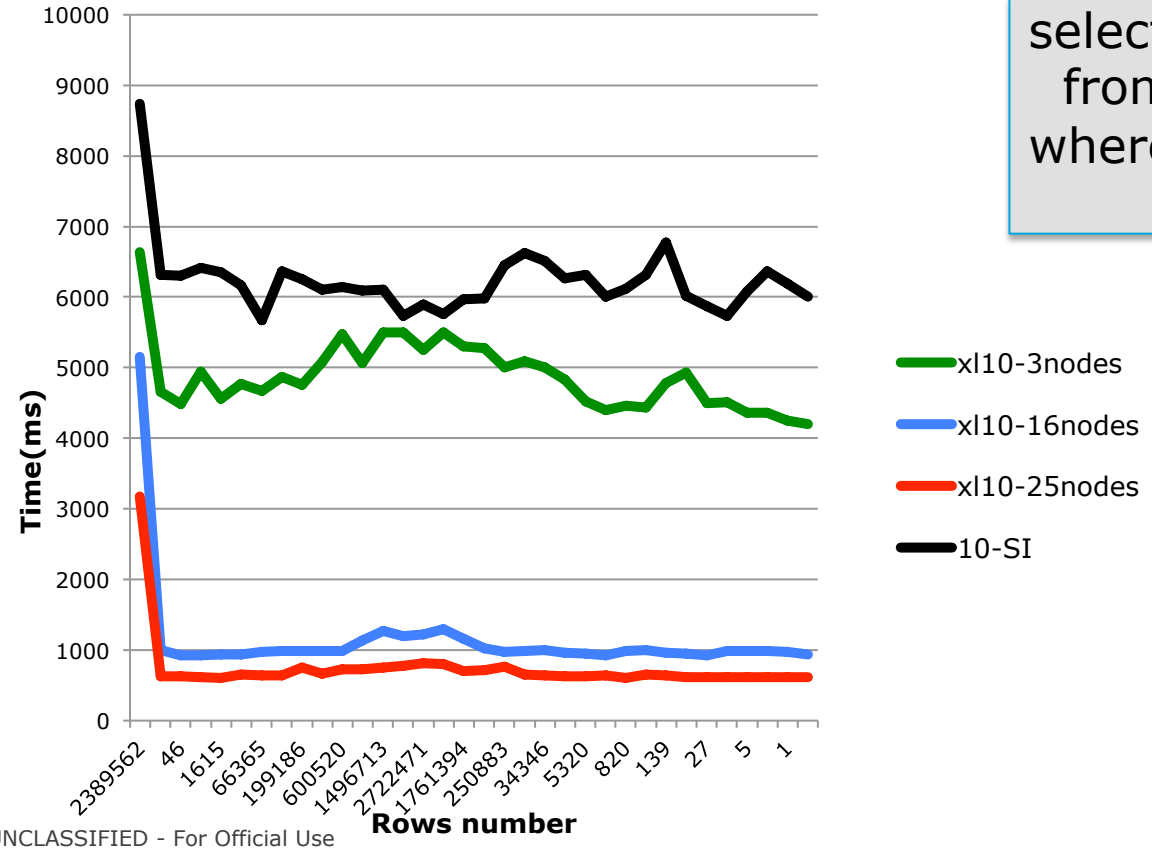



```
Select *
  from kids_mb_catalogue
Where mag_auto_i_trunc=XX
```

Postgres-xl 9.5 vs 10



Aggregate queries in postgres-xl scalability



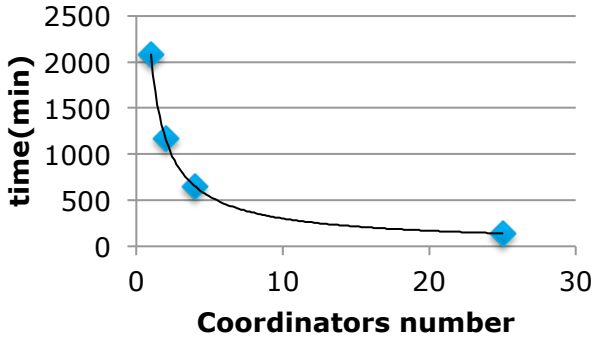

```
select count(*)
  from kids_mb_catalogue
where mag_auto_i_trunc=XX
```

- No index used.
- First count takes longer: parsing and caching.
- Postgres single instance v. 10 is the slowest
- Xl solution scales well but 25 is not much much better than 16
- Netapp issues in the VMs.

Ingestion scalable tests in Postgres-XL

Ingestion:SPV02 catalogue: 1,2TB;~10M rows/file;260 files
Ingesting in parallel in X#coordinators:

Coordinators number	Time(min)	Number of files ingested per node	Total time(min)
1	8	260	2080
2	9	130	1170
4	10	65	650
25	14	10	140



Ingesting in parallel with 12-15 coordinator should be enough according to plot.

PGBENCH for GAIA catalogue

pgbench is a simple program for running benchmark tests on PostgreSQL. It runs the same sequence of SQL commands over and over, possibly in multiple concurrent database sessions, and then calculates the average transaction rate (transactions per second)

2 queries:

Table name: dr2.epoch_photometry

Table size: 5965 MB

Rows (count): 3054247

1) get the epoch_photometry for a random source_id

```
 begin
select * from dr2.epoch_photometry where source_id=
(select source_id from dr2.epoch_photometry order by random() limit 1);
end;
```

2) get the epoch_photometry for a fixed source_id

```
select * from dr2.epoch_photometry where source_id= 5233697651993379456;
```

PGBENCH for GAIA CATALOGUE

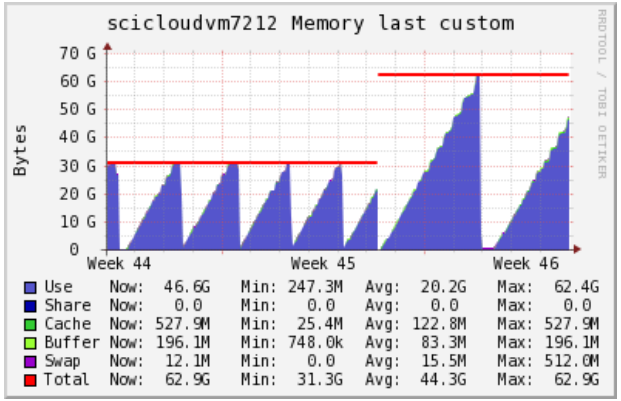
Test#	Connections	Threads	Query	Transactions per second	Latency	Transactions	Transactions done
1	1	1	1	40	0,024	10	10
5	10	10	1	351	0,028	10	100
9	50	2	1	1	32	10	298/500
11	50	50	2	560	0,89	10	500
12	50	50	1	1	30	10	324/500
22	300	10	1	172	1,7	10	3000
26	500	10	2	23	21	10	1020/5000

pgbench -p 8300 -h gacsrelcluster01coord1 -f query1.sql -c 1 -j 1 -t 1 gacs

In conclusion, we expect that running a **pool of 100/150 simultaneous connections** will run **stable** in PG-XL

Issues Found (I)

1- GTM memory leak:



Solution: Add GTM proxies in every datanode and coordinator. Only 1GB of memory would be needed for GTM.

Issues Found (II)

2-GTM crash and PGXL not working after that:

ERROR: Could not obtain a transaction ID from GTM. The GTM might have failed or lost connectivity

ERROR: Snapshot too old - RecentGlobalXmin (372323) has already advanced past the snapshot xmin (10000)

Solution: a GTM crash might cause this, using GTM standby and failing over to it when something bad happens to the GTM. We needed to increase the value of **next_xid, global_xmin** to a larger value than the one the logs.

3- Reshuffling a large table (1,1TB):

euclid=# alter table spv02 delete node(easdpsdn25,easdpsdn24,easdpsdn23,easdpsdn22,easdpsdn21,easdpsdn20,easdpsdn19,easdpsdn18,easdpsdn17,easdpsdn16,easdpsdn15,easdpsdn14,easdpsdn13,easdpsdn12);

ERROR: write failed

Issues Found (III)

4-When an “init all” is issued on an already configured cluster, the datanodes data folders are removed, it depends on the shell. It is confirmed for:

Bash in RHEL6 with PGXL 10alpha2, based on PG 10beta3

5-Migrating Relational Model to PG-XL can be very complicated.

No direct way to do it.

6-Foreign Data wrappers are not supported in this version

postgres=# create extension postgres_fdw;

ERROR: Postgres-XL does not support FOREIGN DATA WRAPPER yet

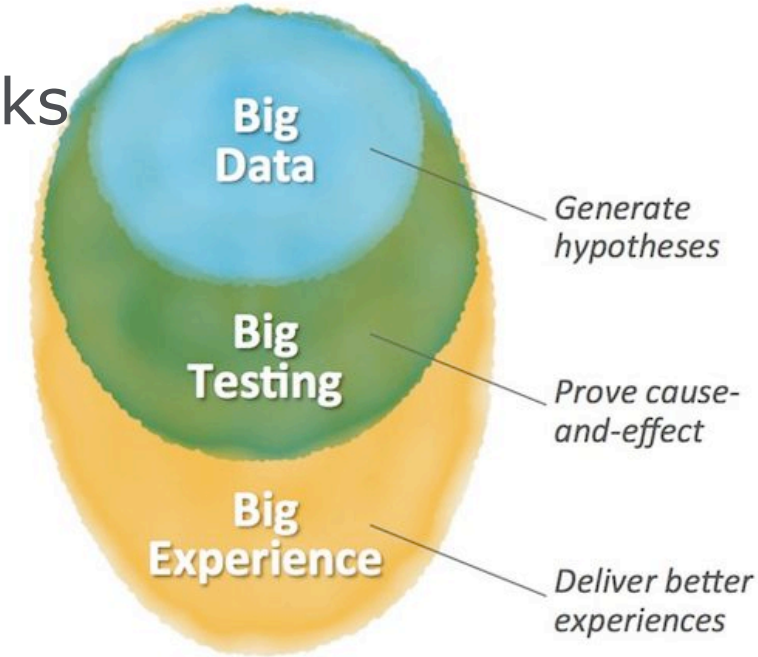
DETAIL: The feature is not currently supported

PGXL distribution lists

- postgres-xl-announce
- postgres-xl-bugs
- postgres-xl-developers
- postgres-xl-general

Next Steps

- Query profiling: queries with geometrical filters(q3c, healpix, pg_sphere)
- Identify performance bottlenecks
- Improve monitoring
- Backup and restore
- Testing…testing..testing

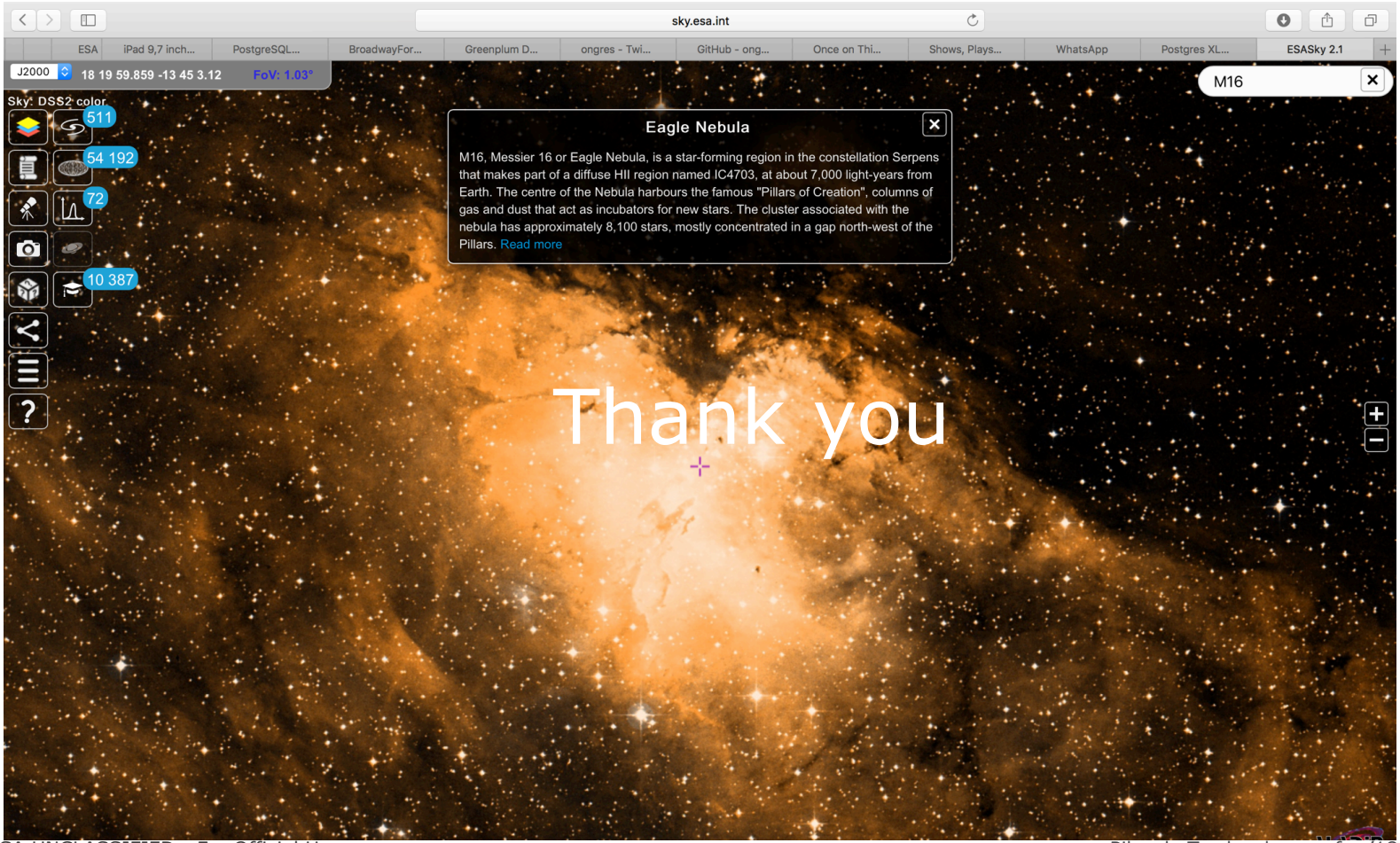


Lessons Learnt and what is missing(I)

- Yes, Postgres-XL is scaling but:
  - still some issues about stability: GTM memory leak
  - backup and recovery can be very dangerous
  - does not work with foreign data wrappers
  - GTM does not rotate logs
  - not direct monitoring tool
  - Configuration to be done with puppet, ansible…
  - Test plan (based on requirements)
  - Talk with the product owners from the beginning..willing to help

Lessons Learnt and what is missing(II)

- Support via distribution lists
- Support license?
- Identify all necessary teams to get maximum benefits(hardware, network, stakeholders)
- Continuation of the PG-XL project?
- Great for Write once, Read many
- Alternative technologies?


sky.esa.int
ESA
iPad 9,7 inch...
PostgreSQL...
BroadwayFor...
Greenplum D...
ongres - Twi...
GitHub - ong...
Once on Thi...
Shows, Plays...
WhatsApp
Postgres XL...
ESASky 2.1
J2000
18 19 59.859 -13 45 3.12
FoV: 1.03°
Sky: DSS2 color
511
54 192
72
10 387
M16
Eagle Nebula
M16, Messier 16 or Eagle Nebula, is a star-forming region in the constellation Serpens that makes part of a diffuse HII region named IC4703, at about 7,000 light-years from Earth. The centre of the Nebula harbours the famous "Pillars of Creation", columns of gas and dust that act as incubators for new stars. The cluster associated with the nebula has approximately 8,100 stars, mostly concentrated in a gap north-west of the Pillars. Read more
Thank you