Technological Challenges in the GAIA Archive

Juan Gonzalez – jgonzale at sciops.esa.int
Jesus Salgado – jsalgado at sciops.esa.int
ESA Science Archives Team

IVOA Interop 2013, Heidelberg
May 2013

Presentation Overview

Gaia Archive Core Systems (GACS)

- Work packages and Subsystems involved
- Architectural draft, technologies to be applied

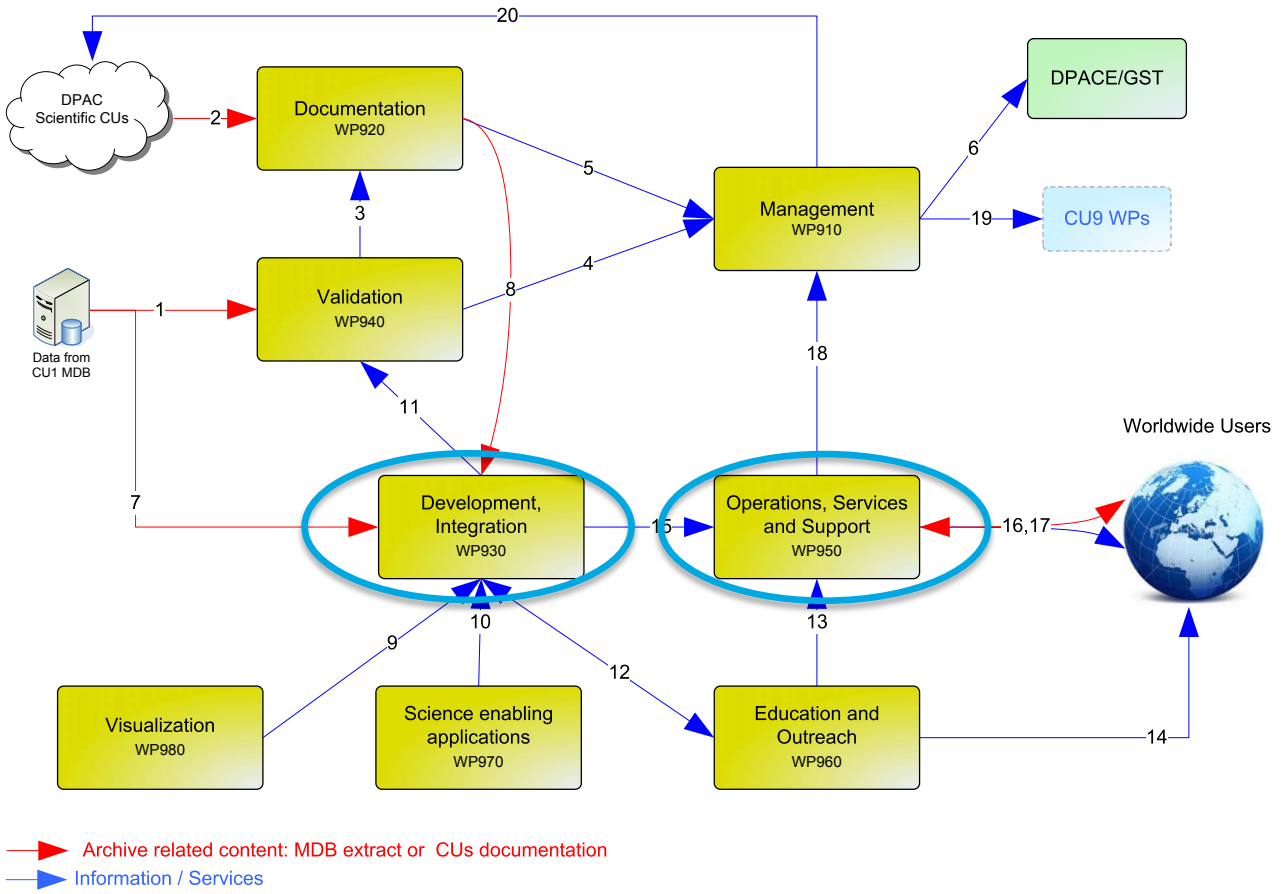
The Interrogator

- TAP+
- Catalogue Databases

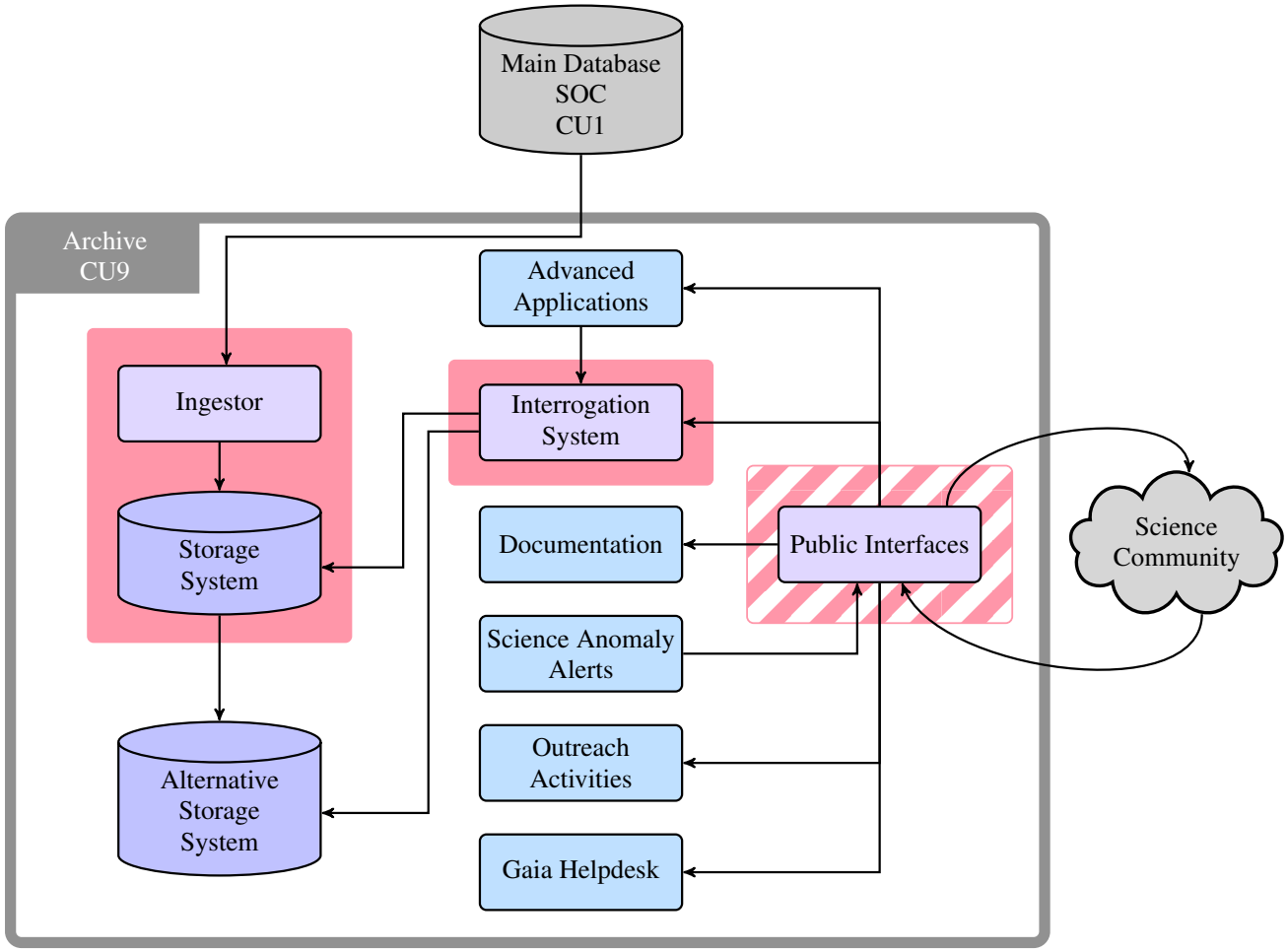
Data products storage

- VOSpace, other technologies
- Simple storage vs Reprocessing infrastructures

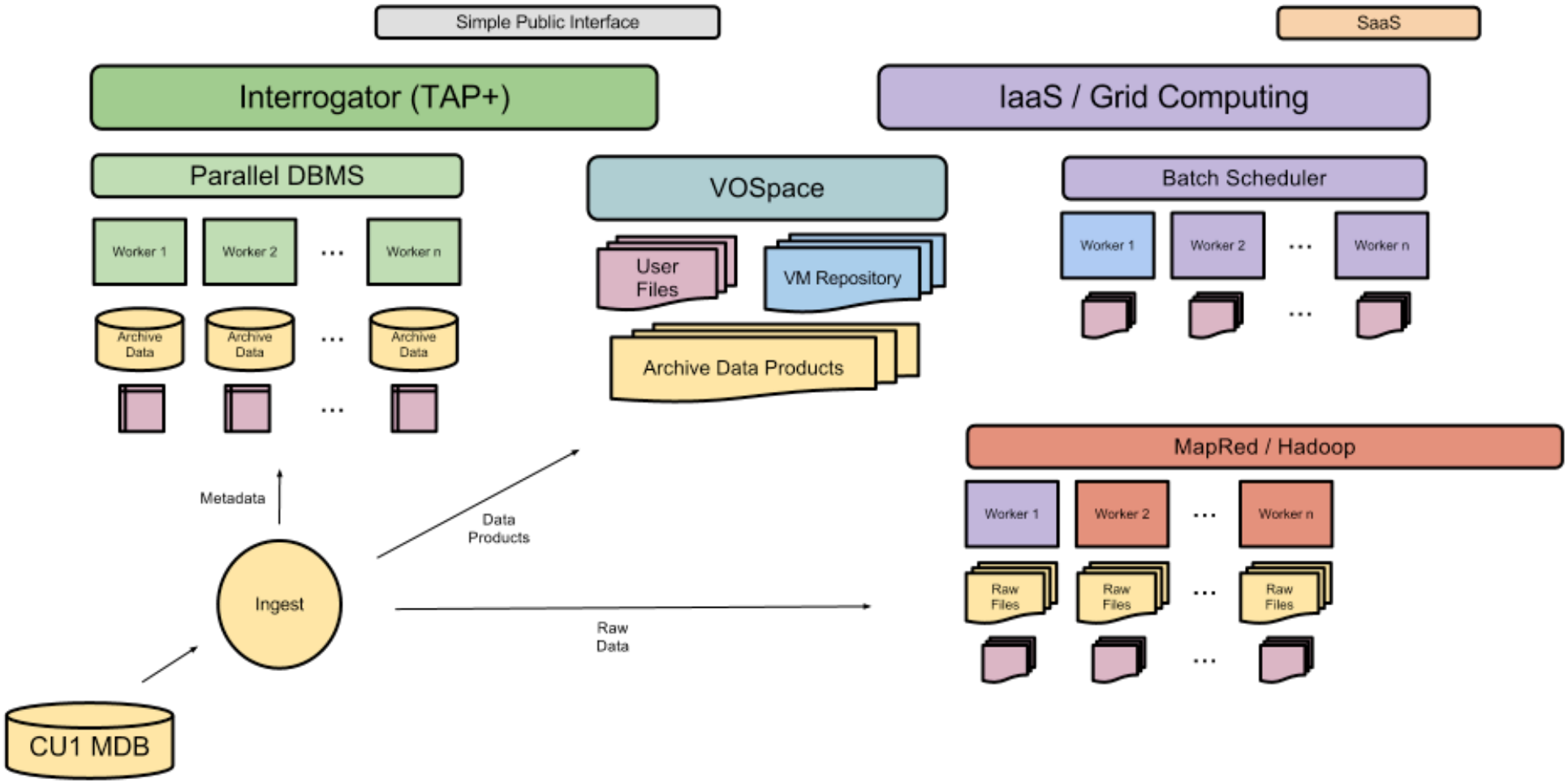
Gaia Archive Core Systems (GACS) (I)



Gaia Archive Core Systems (GACS) (II)



Gaia Archive Core Systems (GACS) (II)



Presentation Overview

Gaia Archive Core Systems (GACS)

- Work packages and Subsystems involved
- Architectural draft, technologies to be applied

The Interrogator

- TAP+
- Catalogue Databases

Data products storage

- VOSpace, other technologies
- Simple storage vs Reprocessing infrastructures

The Interrogator: TAP+

- TAP+: extensions needed

  - **Pagination**:
    - Needed for requesting data, specially from user interfaces
  - **Linking to Products:**
    - Several products per observation, ObsTap not setting solutions for it. Datalink?
  - **Hierarchical / Multidimensional / Object Oriented Output:**
    - How to represent data that's multidimensional in nature in tables. VO-DML?

The Interrogator: Catalogue Database

- Current trends on archiving large datasets

  - **No SQL**
    - Relaxing ACID rules will bring higher performance
  - **Shared-Nothing architectures:**
    - Splitting your dataset among machines will increase locality of the data, and computing power per volume of data.

The Interrogator: Catalogue Database

- No-SQL promess a lot of performance, but, they do so by reducing SQL premises
  - **No ACID**:
    - Might not be a big deal for scientific usage
  - **No Full SQL-92:**
    - Are you sure your project does not need full SQL? ADQL does. TAP does.
- Shared-nothing architectures promess as well more performance, but:
  - **Data is partitioned:**
    - Partitioning reduces by itself the general purpose orientation of one database

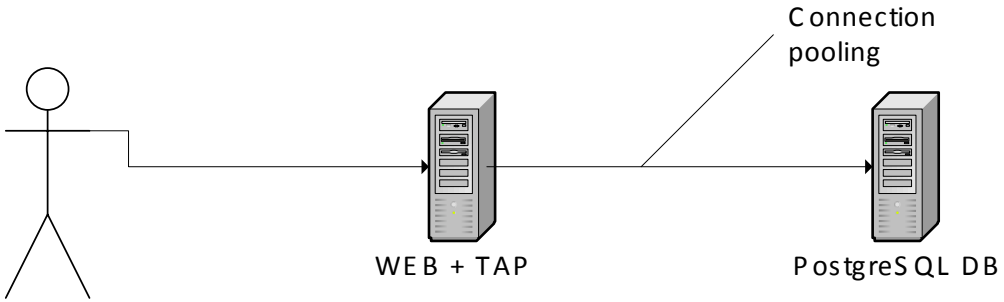
The Interrogator: Catalogue Database

- "Traditional" Monolithic relational databases have many adavantages in terms of cost
  - **Way less administration costs**:
    - Running and maintaining a shared-nothing cluster requires specialized people, tools and procedures
  - **Great Open Source free software: no very expensive software licenses**
    - Parallel databases with good administration procedures implemented are often very expensive.
  - **Cheaper hardware:**
    - Vertically scaling your machines is way less expensive (up to a certain point)

The Interrogator: Catalogue Database

So the only disadvantage of monolithic DBs is performance for large amounts of data? How large is my data, then?

The Interrogator: Catalogue Database



GUMS 10 Catalogue

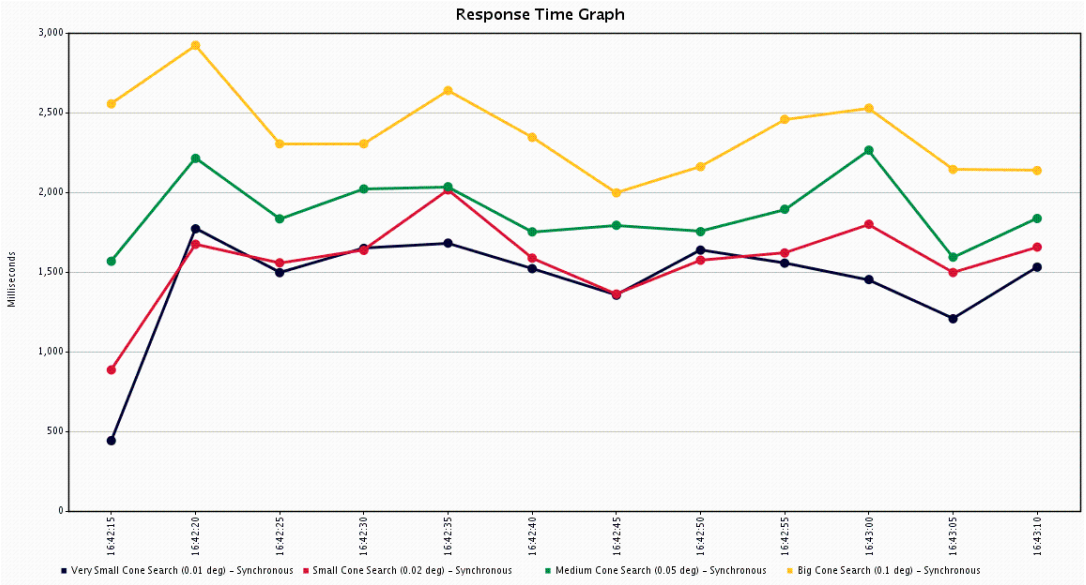
~1.5 Billion (10E9) sources
DB space ~ 1TB

PostgreSQL DB specs
- 128 GB RAM
- 16 cores
- Disk
    - o 2 x 700 GB local SAS 10K rpm HDs
    - o OS-defined RAID 0,
    - o ~140MB/s peak. sequential access

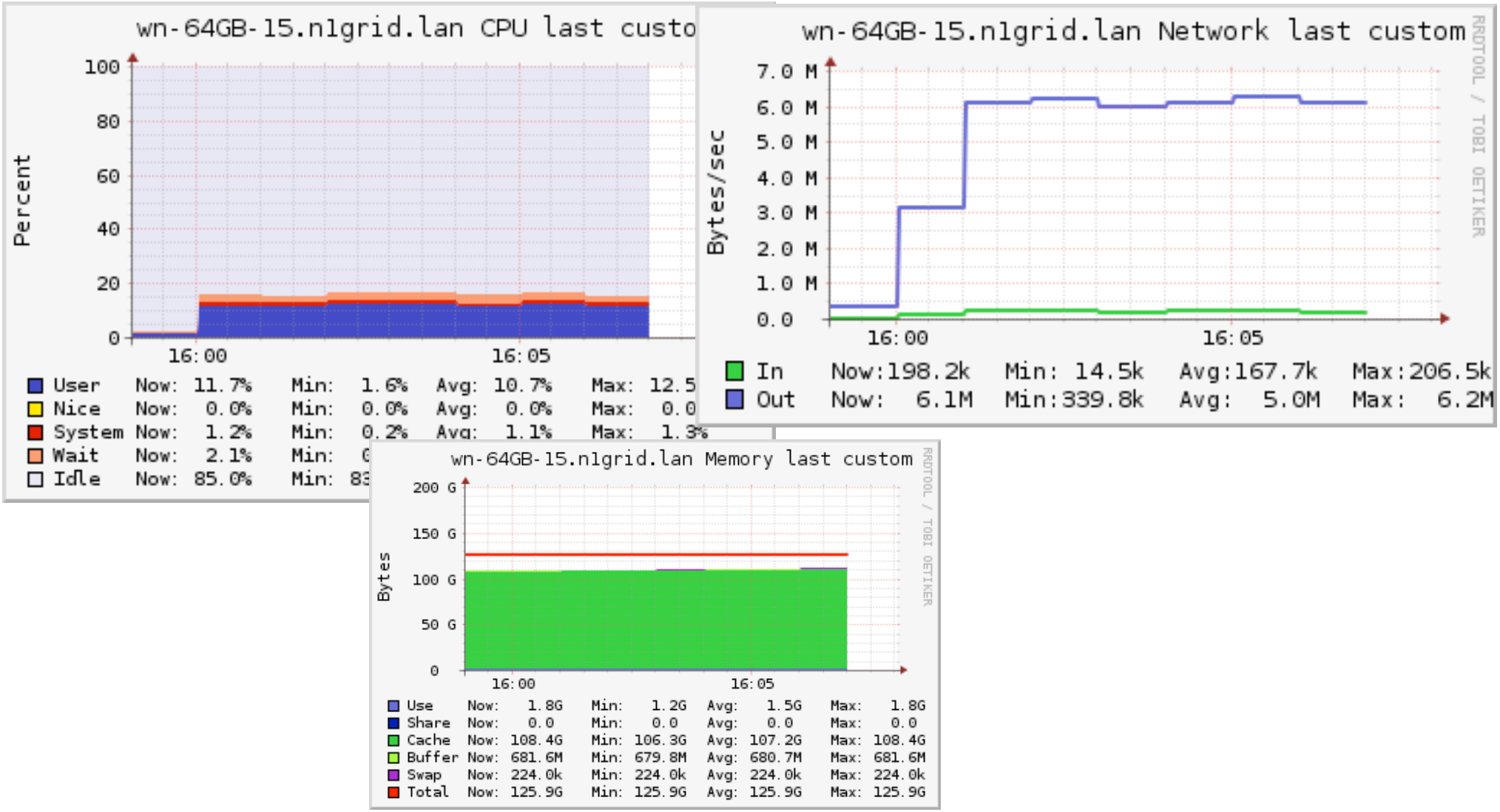
The Interrogator: Catalogue Database

- **100 execution threads**:

User Pattern	avg response time	Error rate (%)	Effective Rate	Bandwidth (KB/s)	Response size (bytes)
Random CS (0.01 deg) - Sync	1380	0	14.57614116	263.2784715	18495.78375
Random CS (0.02 deg) - Sync	1528	0	14.2077783	662.0614915	47716.8881
Random CS (0.05 deg) - Sync	1848	0	13.79264281	2854.584875	211931.4588
Random CS (0.1 deg) - Sync	2340	0	13.38874273	7066.149202	540434.373
TOTAL	1762	0	55.72940941	10786.13252	198189.7855

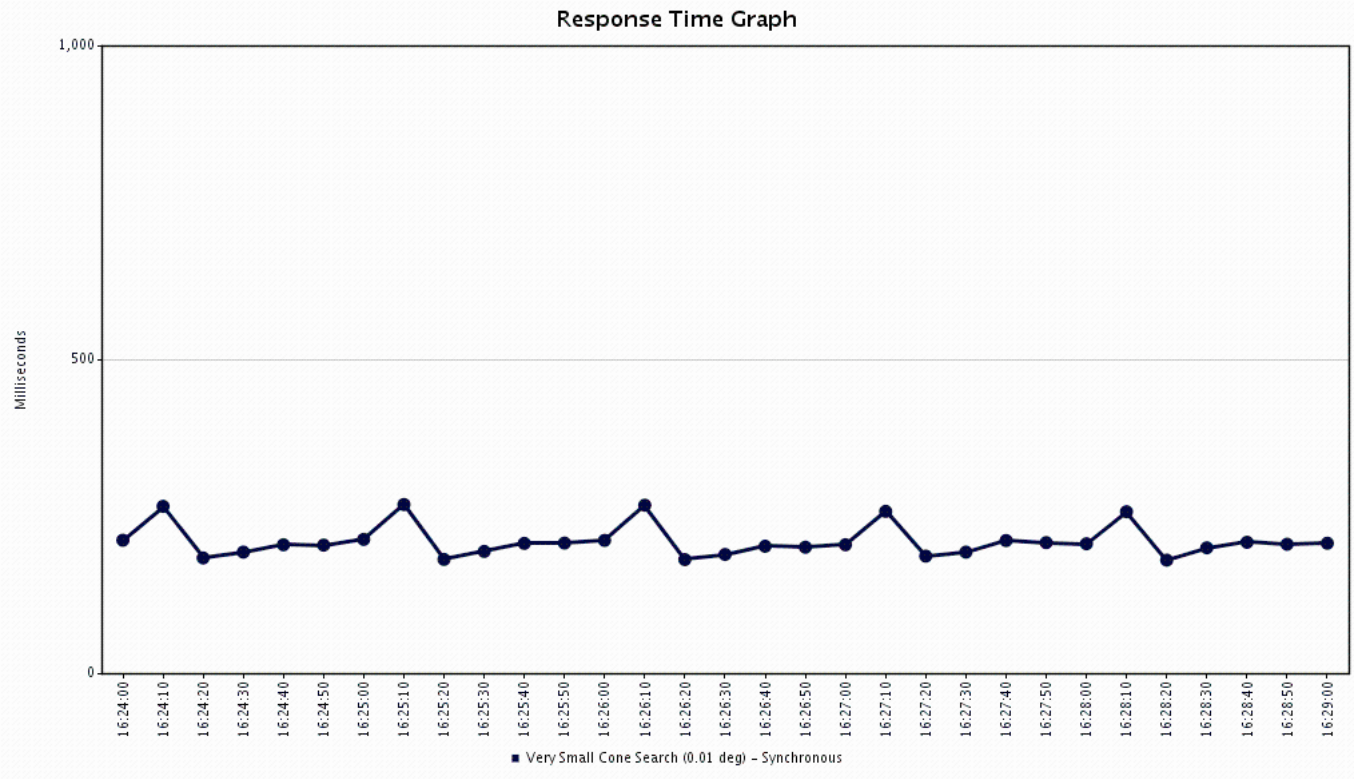


The Interrogator: Catalogue Database

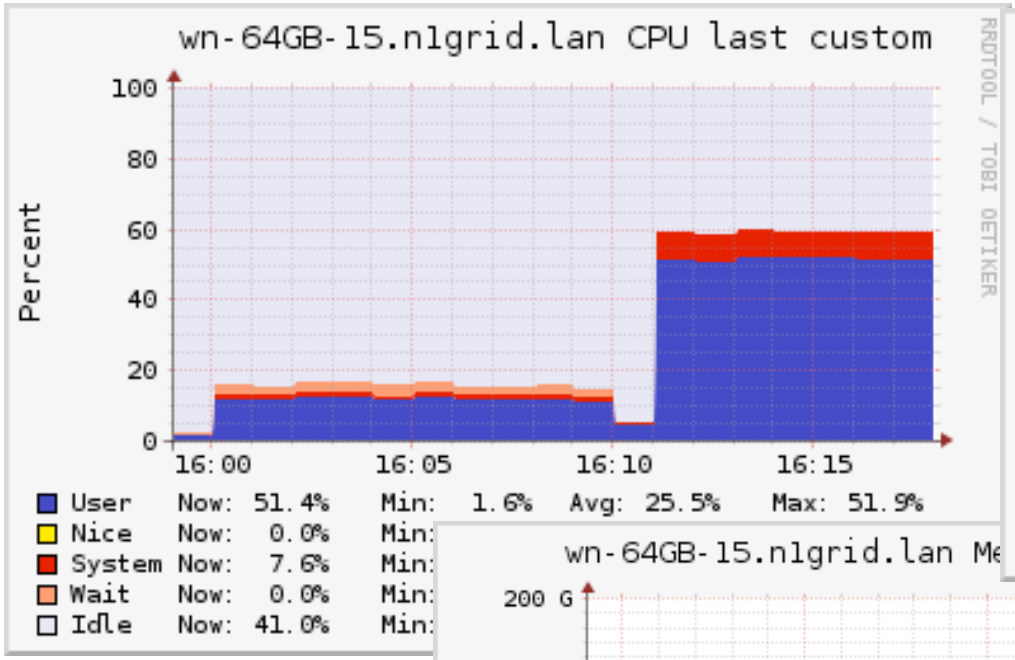


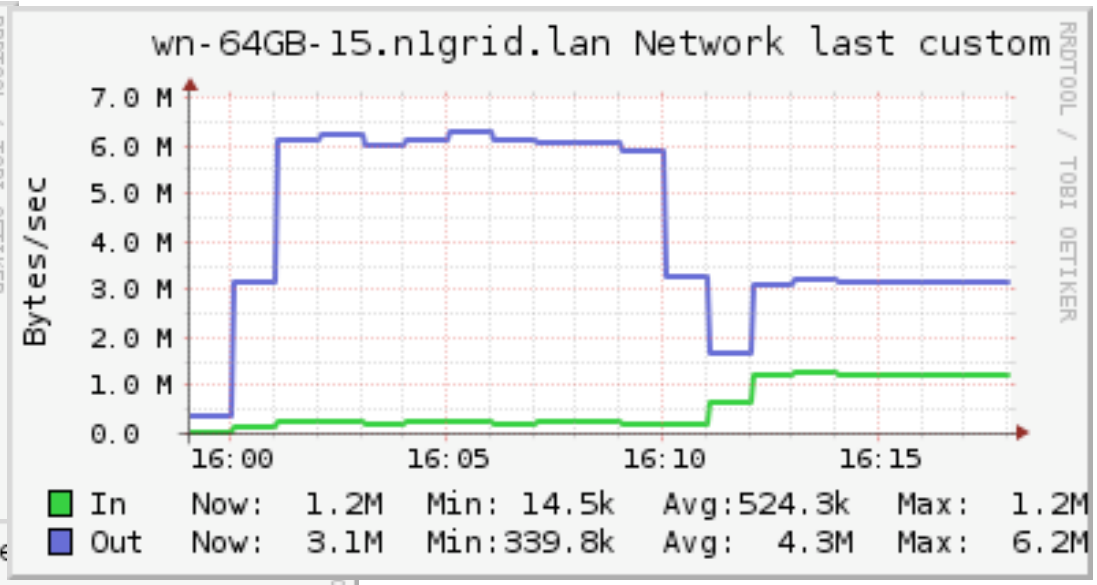
The Interrogator: Catalogue Database

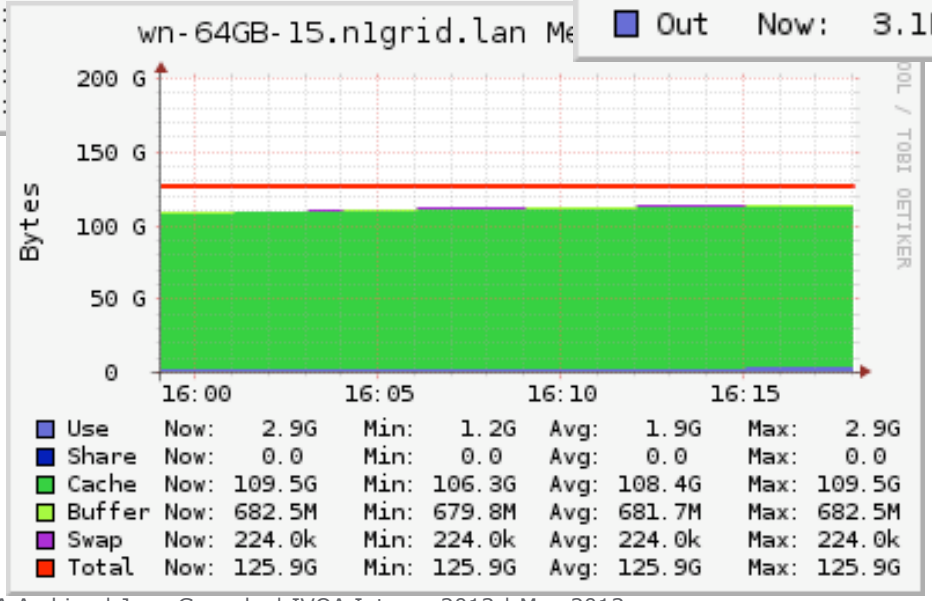
User Pattern	avg response time(ms)	Error rate (%)	Effective Rate(s)	Bandwidth (KB/s)	Response size (bytes)
Random CS (0.01 deg) - Sync	211	0	413.5	7539.87	18671.1



The Interrogator: Catalogue Database

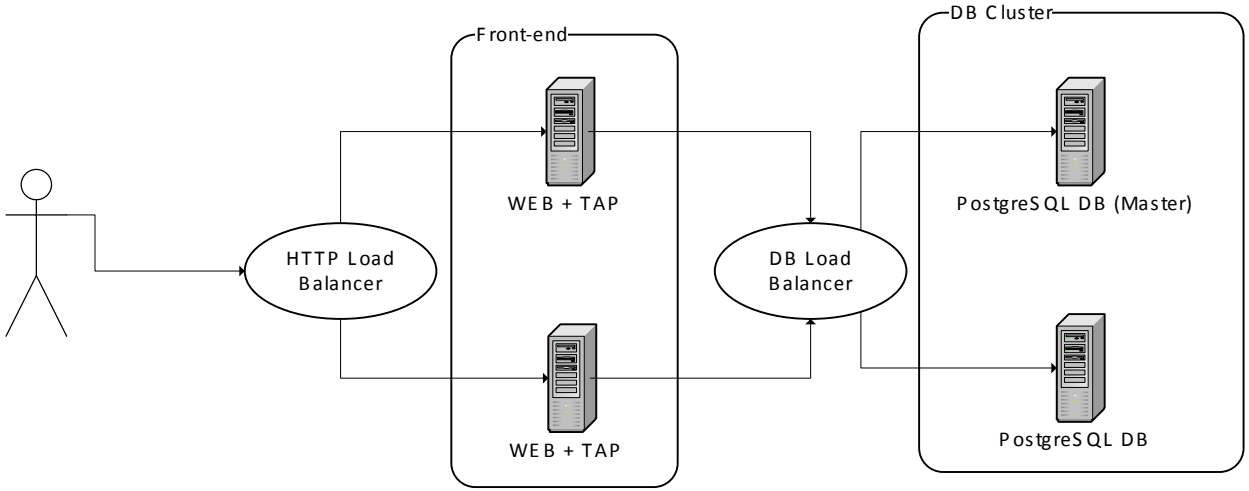






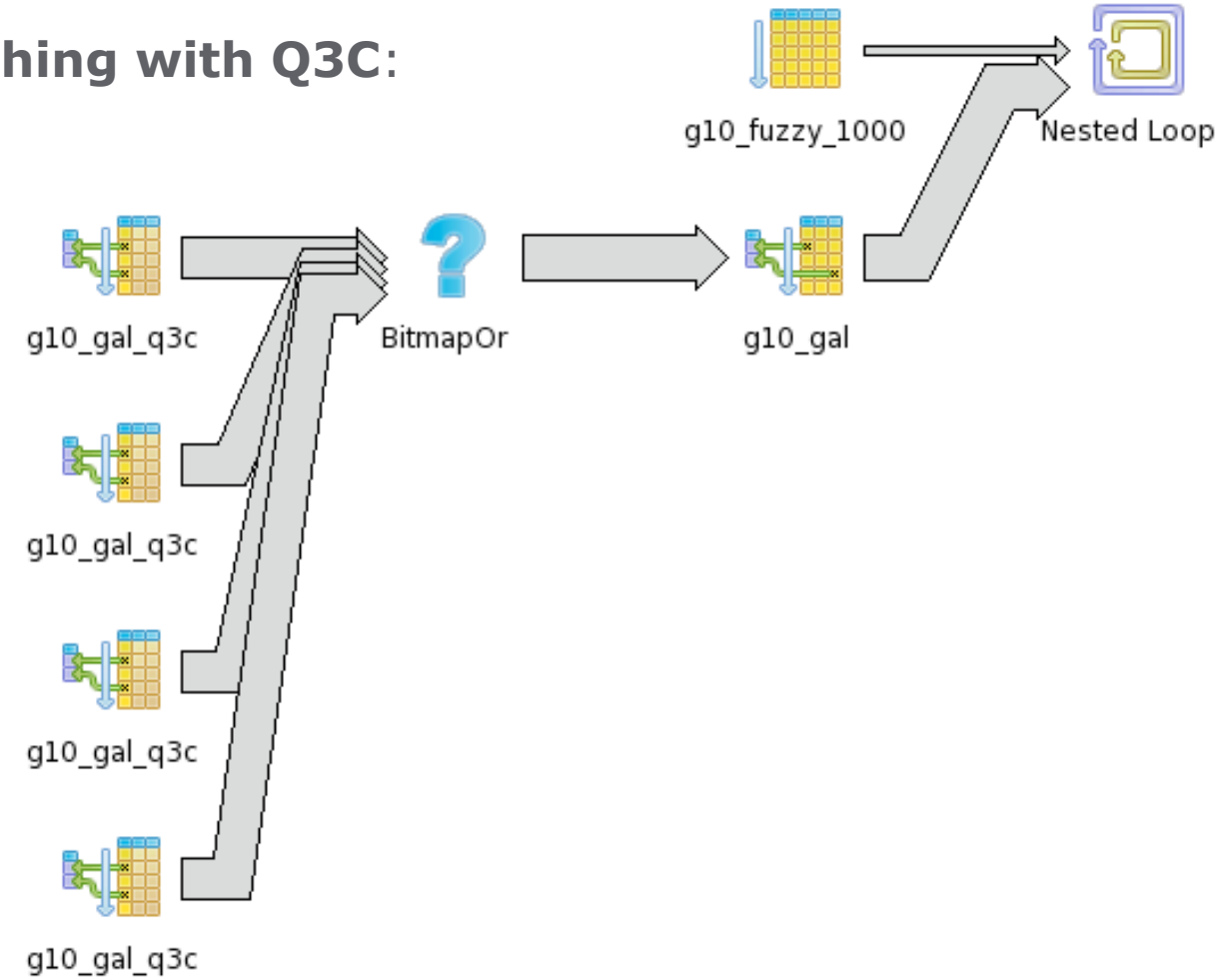
The Interrogator: Catalogue Database

- **Scaling up in number of users:**



The Interrogator: Catalogue Database

- **Crossmatching with Q3C**:



The Interrogator: Catalogue Database

CrossMatch sorted by proximity

Query:

```
SELECT q3c_dist(t.alpha, t.delta, m.alpha, m.delta) AS dist, *
FROM g10_fuzzy_1000 AS t, g10_mw AS m
WHERE q3c_join(t.alpha, t.delta, m.alpha, m.delta, 0.00027)
ORDER BY dist
```

	Interactive Query	Full output	Rows
Fresh execution	78 ms	217 ms	1,857
In memory	75 ms	217 ms	

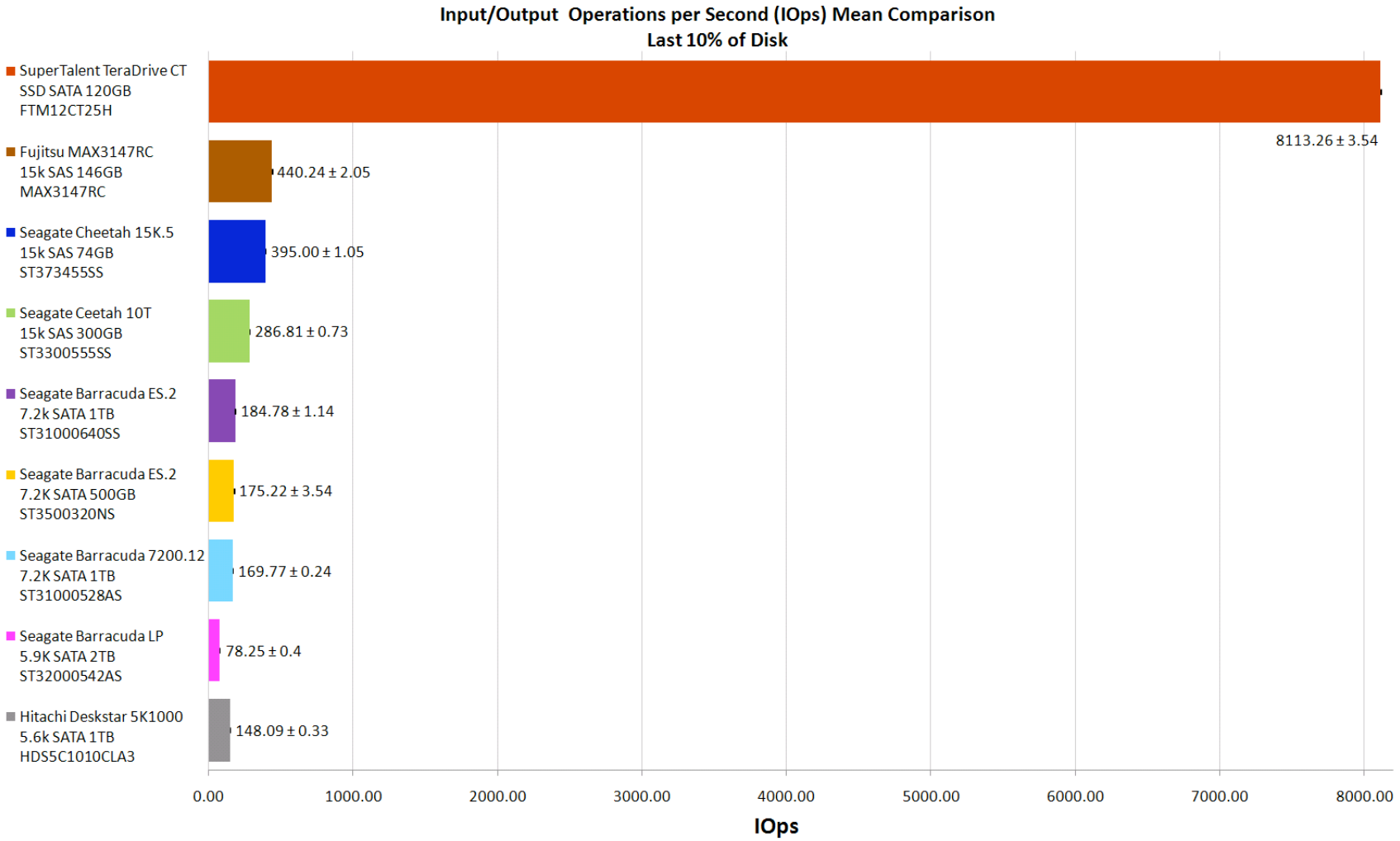
The Interrogator: Catalogue Database

- Random reads kill performance in systems with HDDs

  - **Performance for our test system**:
    - Sequential reading ~140 MB/s
    - Random reads <10 MB/s

  - **How to rise IOPS:**
    - Larger Disk Cache (High memory systems, 2TB is COTS now)
    - Storage with larger IOPS

The Interrogator: Catalogue Database



The Interrogator: Catalogue Database

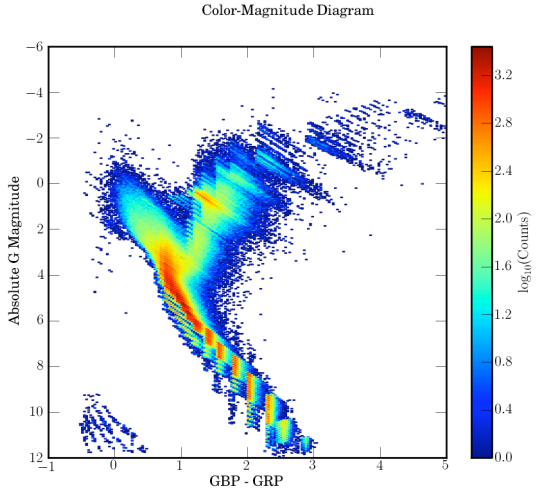
480K * 4KB = 1.9 GB/s !

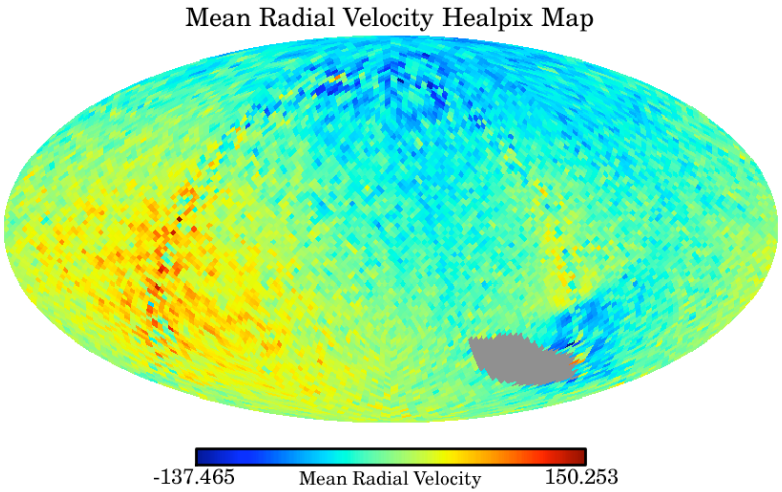
ioDrive2 Duo Capacity	2.4TB MLC*	1.2TB SLC*
Read Bandwidth (1 MB)	3.0 GB/s	3.0 GB/s
Write Bandwidth (1 MB)	2.5 GB/s	2.5 GB/s
Ran. Read IOPS (512B)	540,000	700,000
Ran. Write IOPS (512B)	1,100,000	1,100,000
Ran. Read IOPS (4K)	480,000	580,000
Ran. Write IOPS (4K)	490,000	535,000
Read Access Latency	68µs	47µs
Write Access Latency	15µs	15µs
Bus Interface	PCI-Express 2.0 x8 electrical x8 physical	
Weight	Less than 11 ounces	
Form Factor	Full-height, half-length	
Warranty	5 years or maximum endurance used	

The Interrogator: Catalogue Database

- There is however a lot of room for large shared-nothing clusters, even in cases where data would "fit" in a single machine
  - **Large shared-nothing clusters are great if you may identify a certain set of usage scenarios beforehand**:
    - Tasks which don't involve complex joins among data that is not likely going to be stored on the same machine (Greenplum, Teradata, Vertica, etc.)
    - Fixed functionalities for which might be even possible to develop specific partitionings and DB software (crossmatches)
    - Tasks which might be mapped to Map-Reduce jobs (Hadoop)
  - **Generation of Histograms or density plots through Hadoop clusters is a great example**

The Interrogator: Catalogue Database





- Execution in Amazon Clusters, data in S3 storage and analysis run on Elastic Map Reduce
  - **Map/Reduce functions have to be written by the astronomer (support framework developed @ ESAC)**
  - **Now rethinking approach to Hive**

Presentation Overview

Gaia Archive Core Systems (GACS)

- Work packages and Subsystems involved
- Architectural draft, technologies to be applied

The Interrogator

- TAP+
- Catalogue Databases

Data products storage

- VOSpace, other technologies
- Simple storage vs Reprocessing infrastructures

Data products storage

- Estimations on reduced data volume up to **1 PB**

  - **Expected delivery date ~2021**:
    - That will leave some room for technology improvement ☺

  - **Currently, high cost differences from just storage to effective batch analysis on raw data**
    - Moving 1PB of data from NAS storage to your processing Grid at runtime is not an efficient option.
    - Map/Reduce infrastructures?

Exposing GAIA Spectra in the VO

Any questions?

Feedback:

jgonzale at sciops.esa.int
sat_gaia at sciops.esa.int

http://www.sciops.esa.int/index.php?project=SAT&page=index