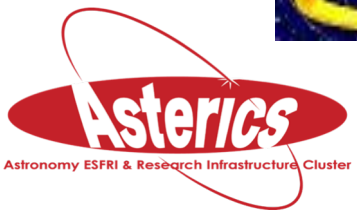
Feedback about last experiments: HiPS and clouds

André Schaaff, Thomas Boch, Pierre Fernique

Centre de Données astronomiques de Strasbourg

IVOA, Victoria, 28/05-01/06/2018, GWS Session 2

CENTRE DE DONNÉES
ASTRONOMIQUES DE STRASBOURG



H2020-Astronomy ESFRI and Research Infrastructure Cluster (Grant Agreement number: 653477).

Foreword

- During the previous Interop in Victoria in 2010, i had a presentation about VizieR in the clouds…
  - Around 8000 euros per year, it was expensive compared to an in-house server
- The experiment is now around HiPS (larger use case) and especially with AWS
  - Hosting HiPS
  - Generation of HiPS
- Acknowledgment: AWS Cloud Credits for Research

Remarks

- Not exhaustive, on going tests (credits end 31 July)
- The aim is mainly to have a better idea about the cost and the performances
- We do not plan to switch from our own servers to the Cloud
- But it could have an interest for a specific usage like the HiPS generation (explanation is following)

My opinion before the tests

- Hosting
  - Flexible with more space on demand
  - No on site hardware to manage
  - World deployment
  - Availability
  - Not sensible to simultaneous access (release of a new HiPS for example)
  - But not really enthusiastic concerning the ”pay as you use” concept…

My opinion before the tests (2)

- Generation
  - The generation of a HiPS is punctual and
    - depends on the initial material (FITS files)
    - depends on the number of orders and the final size
  - Needs sometimes to replay the process a few times
    - If it takes 24 hours per processing you have to wait before the next checking and correction
    - If we can reduce this time (for example) to 2 hours it will possible to finalize perhaps in one day
    - On demand resizing of the computing resource could have an interest, but for which cost ?

Hosting HiPS in the Cloud

- On AWS, S3 and CloudFront
- S3 is cheaper but the data is located in one datacenter, S3 is not recommended for data streaming, large images, etc.
- CloudFront benefits from the World presence of Amazon and the location of the user is taken into account, in this case you don't really know how it works but it should be most efficient on the user side

Use case

- Remark: a small HiPS to facilitate all the manipulations (upload duration on S3, etc.)

- Hosting
  - HiPS AKARI-N60
  - Order 5
  - 5 GB
  - ~300KB / tile

- Generation
  - AKARI-N60 FITS files
  - 13 GB

Hosting cost (/ month)

- S3
  - 1 TB 24.46$
  - 24.87$ with 1,000,000 ”get”
  - 28.65$ with 10,000,000 ”get”
  - 66.45$ with 100,000,000 ”get”
  - Data transfer to CloudFront is free

- HiPS examples
  - Most accessed HiPS: 300 GB with 10,000,000 “get” => 11,27 $
  - AKARI-N60 with 1,000,000 “get” => < 1$

Hosting cost (2)

- CloudFront (with the best coverage)
  - Out transfer (“get”)
    - 2GB / day => 1$ / month (tile ~ 100KB)
    - 100GB / day => 260$ / month (tile ~ 500KB)
    - 100GB / day => 323$ / month (tile ~ 50KB)
- Remark: …

In the real life (/ month)

- A HiPS server with 200 TB
  - 5012.35$ for S3
  - (1 PB => 25586$ = ~310,000$ / year)
- CloudFront => 262$ for 100GB/day and an average of 300KB / tile
- => 5275$ * 12 => ~ 63,300$ / year

Performances

- AKARI-N60 on S3, accessed via CloudFront
  - S3 data center is Paris
  - CloudFront best coverage
- From the same location(from Victoria) with the same network and with Aladin, the average loading was ~20% better with AWS
- Has to be deeply tested on a long period with AWS options like transfer acceleration

Generating HiPS in the Cloud

- We have used EC2
- Reference time at CDS: 21 minutes
- 2 configurations
  - c5d.2xlarge (8 vcpu, 16GB RAM, 0,14$ / hour)
    - ~24 minutes
  - r4.8xlarge (32 vcpu, 244GB RAM, 0.8$ / hour)
    - ~12 minutes
- Remark: without optimization

Conclusion (1) about HiPS hosting

- Putting the data on AWS is easy but “pay as you use” means “pay for your users” and you don’t really know how much data they will download, how much “get” you will have at the end of the Month
- Access and hosting of 200TB is (too) expensive
- It could have an interest for a specific usage like
  - hosting of the most popular(s) HiPS
  - hosting some HiPS in other regions

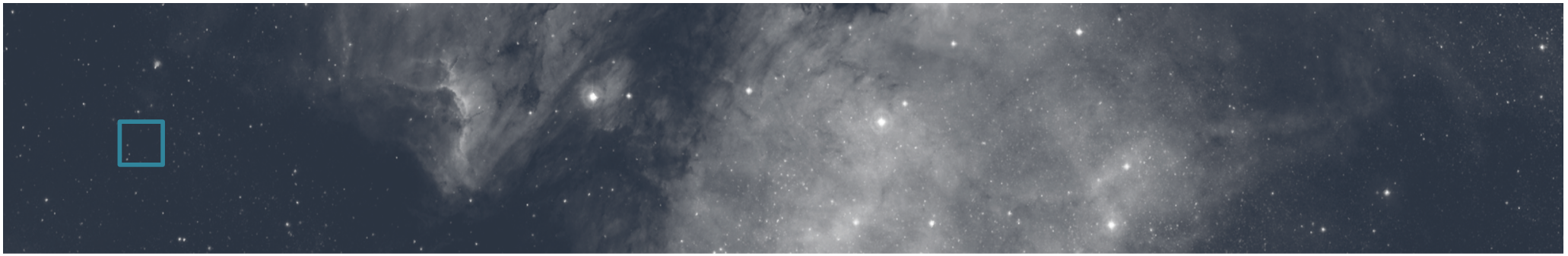
Conclusion (2) about HiPS generation

- Easy to size your hardware on demand
- It is easy (you are the user) to evaluate how much you will pay
- First results very good without optimization

- Next step: generation of a “big” HiPS and optimization of the (cloud) hardware
- For the use case AKARI-N60, a “good” performance (to reduce the whole time (replaying) for other future HiPS) should be under 5 minutes
- Global cost, including the (out) data transfer, to be evaluated

Few words about Spark

- Work done with François-Xavier Pineau and Corentin Sanchez
- Since Santiago Interop the (Spark) X-Match Algorithm has been improved
- Next step is to test it with Cassandra
- Report next time



H2020-Astronomy ESFRI and Research Infrastructure Cluster (Grant Agreement number: 653477).