Implementing VO services in the commercial cloud

André Schaaff, Thomas Boch, Mihaela Buga, Pierre Fernique, Grégory Mantelet, Anaïs Oberto, François-Xavier Pineau

Centre de Données astronomiques de Strasbourg

IVOA, College Park, 8-10/11/2018, Operations IG session

CENTRE DE DONNÉES ASTRONOMIQUES DE STRASBOURG

IVOA

France

Foreword

- During the last Interop in Victoria i gave a short feedback about quick experiments around HiPS with AWS to host / generate HiPS
- We made deeper tests over the Summer to have a better and realistic idea about the cost and the performances
- The aim was not to switch to the cloud but to make a comparison with our own hardware / architecture and to answer questions like: are we better ?, cheaper ?, ...

(Remark: prices are at the time of the experiments, you can probably find similar offers and prices with other cloud providers)

- Acknowledgment: AWS Cloud Credits for Research

HiPS generation in the Cloud

- The generation of a HiPS depends on the initial material (FITS files)
- It is not always done in a single process => multiple tries with several hours / days at each time
- What did we expect ?
  - Benefit from AWS EC2 resizing to reduce the time between 2 tries
  - Benefit from data storage resizing to handle large datasets
  - A reasonable rate
- Our use case : VISTA band K (3.5TB to upload, HiPS => 2TB)

HiPS generation in the Cloud (2)

- HiPS generation is threaded on a same server, no cluster parallelization
  - AWS EC2 is cheap for small instances but for HiPS generation we need large instances

$0.448 per On Demand Linux m5.2xlarge Instance Hour	369.524 Hrs	$165.55
$0.896 per On Demand Linux m5.4xlarge Instance Hour	147.739 Hrs	$132.37
$1.818 per On Demand Linux c5.9xlarge Instance Hour	8.054 Hrs	$14.64
$1.937 per Dedicated Linux c5.9xlarge Instance Hour	1.266 Hrs	$2.45
$16.806 per On Demand Linux x1.32xlarge Instance Hour	2.120 Hrs	$35.63
$3.636 per On Demand Linux c5.18xlarge Instance Hour	27.239 Hrs	$99.04
$5.376 per On Demand Linux m5.24xlarge Instance Hour	55.035 Hrs	$295.87

- Even with the larger instances we had not a better duration than with our own server (we tested with all the proposed storage possibilities, with dedicated instances, …)
- We have not experimented with dedicated physical instances as it is too expensive

HiPS generation in the Cloud (3)

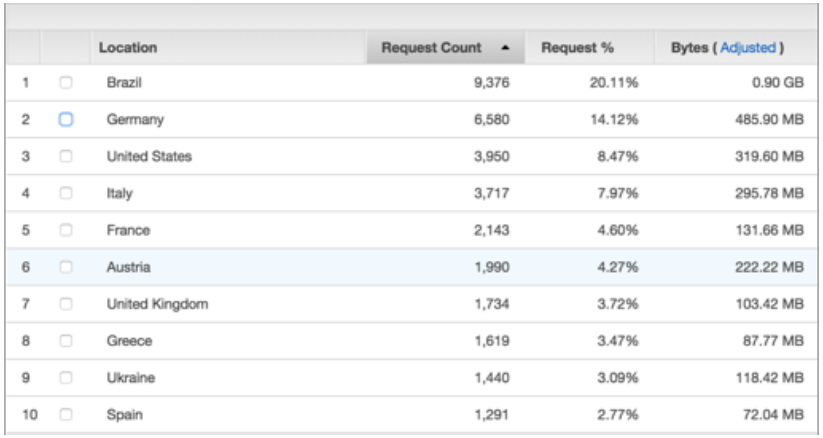
- We had not always the same duration for the same HiPS => cloud is a sharing of disks, cpu, … => at the end not the same cost !

- The upload of the initial data is not free, depends on the storage / instance type

- The download of the result is expensive (around 70$/TB)
  - 2TB for our use case => 140 $ to download the HiPS
  - But PanSTARRs = 30 TB / band => 2100$ (it is possible to transfer it directly on S3 / Cloud Front)

Hosting HiPS in the Cloud

- AWS S3
  - The data is located in one data center, in a bucket
  - Cheap (upload is free)
- CloudFront
  - You don't really know how it works but it is more efficient, the data (or a part ?) is replicated
  - You can choose a worldwide coverage or regions
  - You have access statistics by country, etc.
  - An AWS S3 overlay with a thin price difference



		Location	Request Count ▲	Request %	Bytes (Adjusted)
1	☐	Brazil	9,376	20.11%	0.90 GB
2	☐	Germany	6,580	14.12%	485.90 MB
3	☐	United States	3,950	8.47%	319.60 MB
4	☐	Italy	3,717	7.97%	295.78 MB
5	☐	France	2,143	4.60%	131.66 MB
6	☐	Austria	1,990	4.27%	222.22 MB
7	☐	United Kingdom	1,734	3.72%	103.42 MB
8	☐	Greece	1,619	3.47%	87.77 MB
9	☐	Ukraine	1,440	3.09%	118.42 MB
10	☐	Spain	1,291	2.77%	72.04 MB

HiPS server on AWS S3 / Cloud Front

- Example of statistics (3 consecutive days) in July 2018, from the list of all the HiPS server

- The cloudfront HiPS server hosted only the DSSColor (270MB) which represents 50% of other server download

N.B.: Statistics provided by Aladin desktop

```
2918.0 MB alasky.u-strasbg.fr
2825.5 MB alaskybis.u-strasbg.fr
1716.0 MB skies.esac.esa.int
 409.0 MB axel.u-strasbg.fr
 357.8 MB d1pim2jdtjozp.cloudfront.net
...
```

```
1638.5 MB alasky.u-strasbg.fr
1065.9 MB alaskybis.u-strasbg.fr
 389.5 MB d1pim2jdtjozp.cloudfront.net
...
```

```
2053.1 MB alasky.u-strasbg.fr
 793.6 MB alaskybis.u-strasbg.fr
 315.6 MB d1pim2jdtjozp.cloudfront.net
...
```

2053.1 MB => ~ 1000 MB for the DSSColor

315.6 MB for the DSSColor

X-Match in the Cloud

- We have also experimented the X-Match on AWS EC2
- We had the same kind of results than for the HiPS generation, not far from the CDS X-Match server but never better even with the best instances

- => We think that we had a bottleneck at the disk level (EBS volumes) both for the X-Match and the HiPS generation

- (we had also tested quickly AWS RDS / Aurora in the frame of Simbad)

Our opinion after these tests

- Hosting on AWS S3 / Cloud Front is really flexible, no consideration like the remaining space on your disks…
- No hardware to manage on site, everything can be done from your laptop
- No mirroring needed, World deployment through AWS Cloud Front
- Availability is high
- Efficient AWS support even at the basic price level
- Amazing when you release a new HiPS, you will not be afraid by the user **rush** !

But…

- You pay for the hosting => reasonable if you have not hundreds of TB
- Upload is free ?
  - ~ True if you upload your data on AWS S3
  - Not true if you upload data on an AWS EC2 instance associated disk as the instance is running and you pay for that
- Download is not free !
- You pay on demand for the traffic => you must be aware of this point (you can set alerts)

Conclusion

- If you want to propose a large set of (large) HiPS surveys, it is probably too expansive (many factors depending on your lab (existing manpower, own hardware, etc.))
- If you want to propose a HiPS server with a few TB of surveys and if you have no manpower to manage the hardware, AWS Cloud Front is a good solution, it can be setup in a few hours and it will be easy to manage online
- (compared to our server) Too expansive for HiPS generation with no real added value if you generate regularly HiPS
- A usage in the frame of the X-Match, even if we had good performances, is for us risky/costly as large volumes (depending on the sources and the parameters) are susceptible to be uploaded at many times in a month

Miscellaneous cost examples

- AWS S3 (/ month)
  - 1 TB 24.46$ (not linear)
  - 24.87$ with 1,000,000 ”get”
  - 28.65$ with 10,000,000 ”get”
  - 66.45$ with 100,000,000 ”get”
  - Data transfer to AWS CloudFront is free
- HiPS examples (/ month)
  - Most accessed HiPS: 300 GB with 10,000,000 “get” => 11,27 $
  - AKARI-N60 with 1,000,000 “get” => < 1$

Miscellaneous cost examples (2)

- A HiPS server with 200 TB
  - 5012.35$ / month for AWS S3
  - AWS CloudFront layer depends on the traffic => 262$ / month for 100GB / day and an average of 300KB / tile
- Total cost per year => 5275$ * 12 => ~ 63,300$

- (1 PB on AWS S3 => 25586$ = ~310,000$ / year)

Miscellaneous cost examples (3)

- AWS CloudFront - Out transfer (“get”)
  - 2GB / day => 1$ / month (tile ~ 100KB)
  - 100GB / day => 260$ / month (tile ~ 500KB)
  - 100GB / day => 323$ / month (tile ~ 50KB)