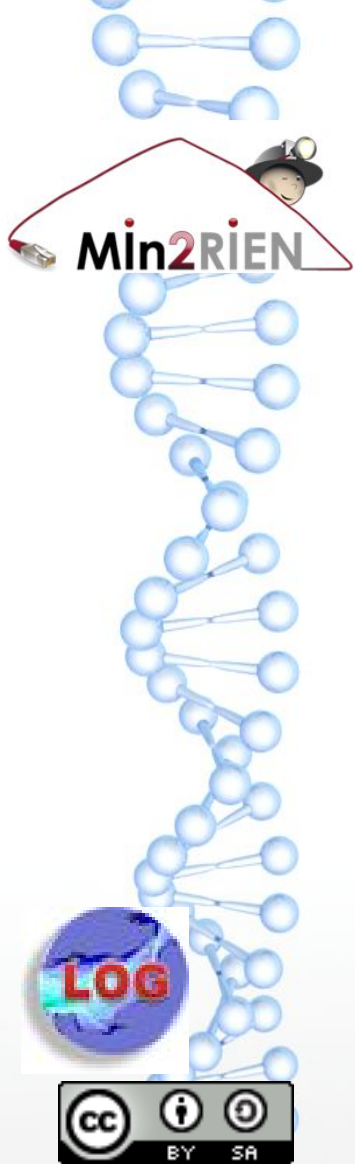




Mise en place d'un cluster

par David Dessailly & Régis Sion

UMR8187- LOG

Laboratoire d'Océanologie et de Géosciences
MREN – Université du Littoral Côte d'Opale.



Sommaire

- **Le LOG au sein du réseau de l'ULCO.**
- **Origine du besoin et les moyens**
- **Appel d'offre et partenaires.**
- **Choix, achat, installation, configuration et tests**
- **Utilisation par l'équipe des océanologues**
- **Conclusion**



Université multi-sites

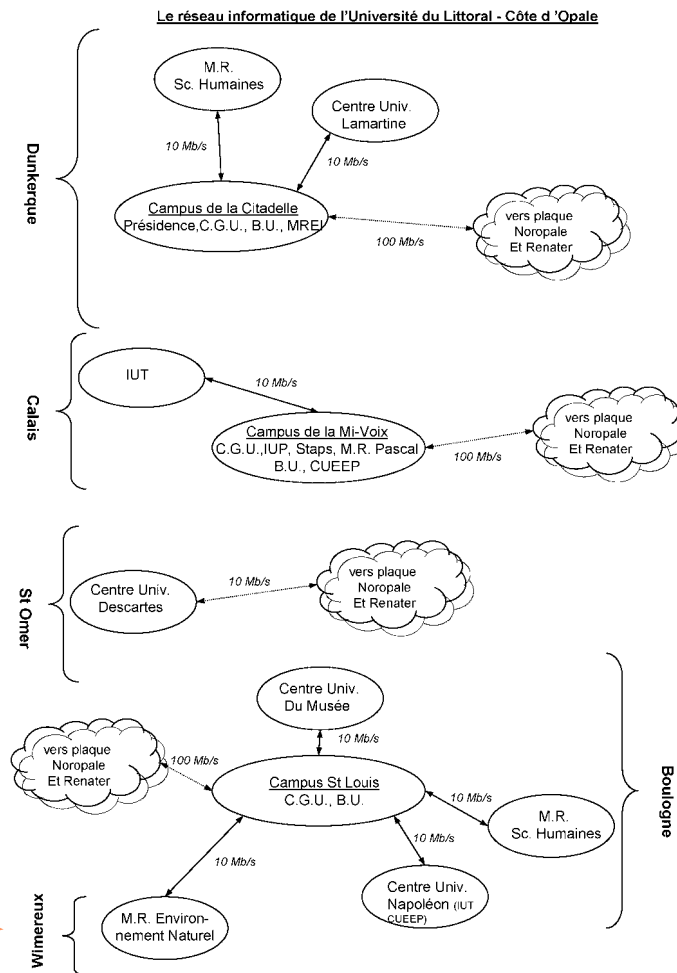
Dunkerques

Calais

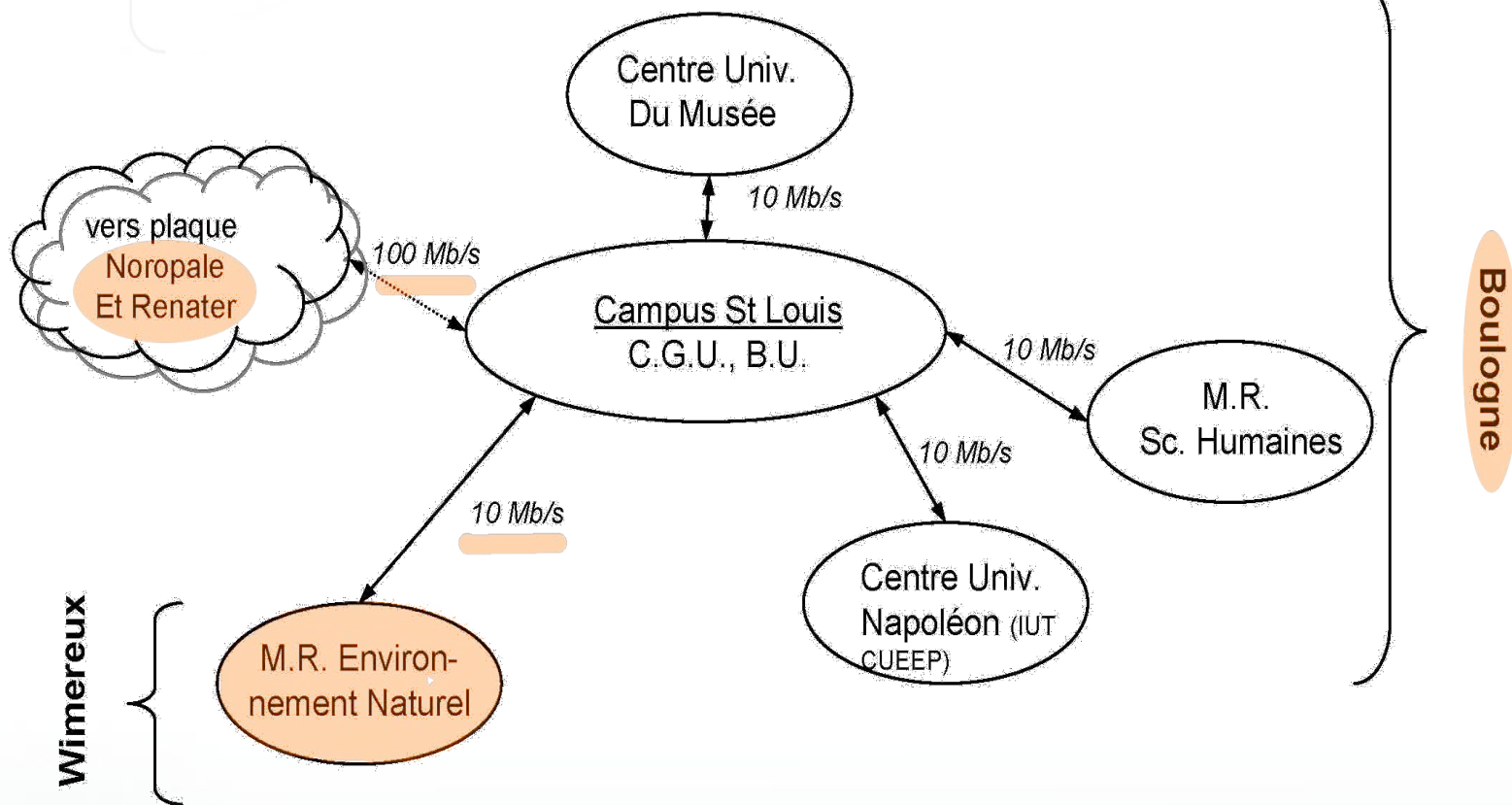
St Omer

**Boulogne/Mer
Wimereux**

Le Réseau De l'ULCO



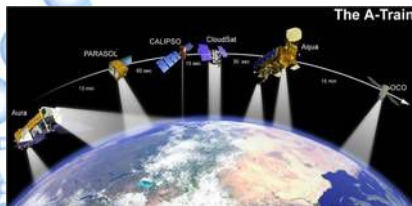
Le réseau local



ULCO - Division Informatique – C. Pierri – février 2012

Le besoin

Satellites



Modélisation

$$\frac{1}{c} \frac{\partial}{\partial t} I_\nu + \hat{\Omega} \cdot \nabla I_\nu + (k_{\nu,s} + k_{\nu,a}) I_\nu = j_\nu + \frac{1}{4\pi c} k_{\nu,s} \int_{\Omega} I_\nu d\Omega$$

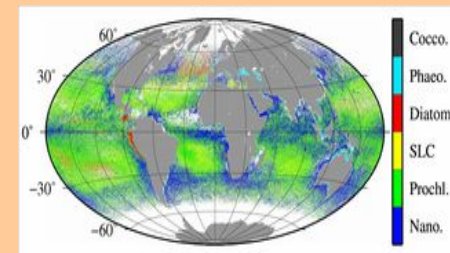
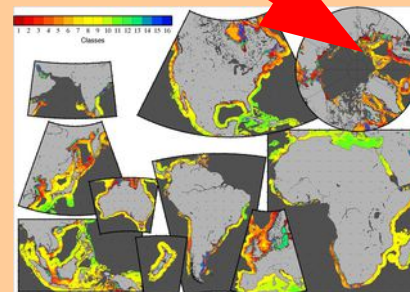
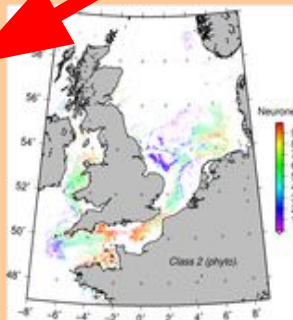
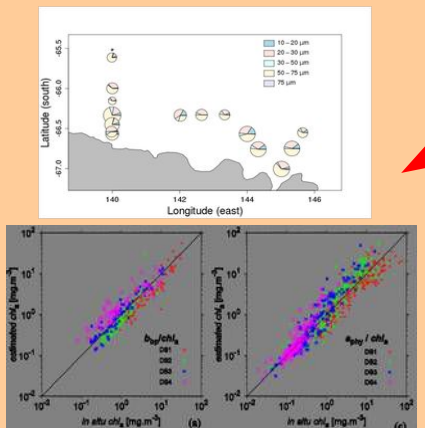
Programmation & SGBD
(C,Python,FORTRAN, PostgrSQL)

Terrain



calculateur

Résultats



Les moyens précédents

- Les postes de travail des chercheurs !

- Station de calcul (achat 2009 sur crédits PNTS/INSU)

Gros PC (HP Z800) configuré en serveur de calcul
2 x Xeon E5520 = 8 cpu - 16 Go ram

- Stockage

Baie SAS (raid6) 16x1To (sata) = 14 To (brut)



Les moyens \$\$\$

Financements

- ULCO

BQR* (ESCAL) : **E**quipement de **S**tockage et de **CA**lcul du **L**og : (34,2K€)

ANR GlobCost : (3,3K€)

- Projets CNES (CNRS)

TOSCA* (PhytoCot) : (4,5K€)

TOSCA (ICAC) : (2K€)

Total budget : 44 K€ HT

* Budget Qualité Recherche

** Terre, Océan, Surfaces continentales, Atmosphère.



La demande

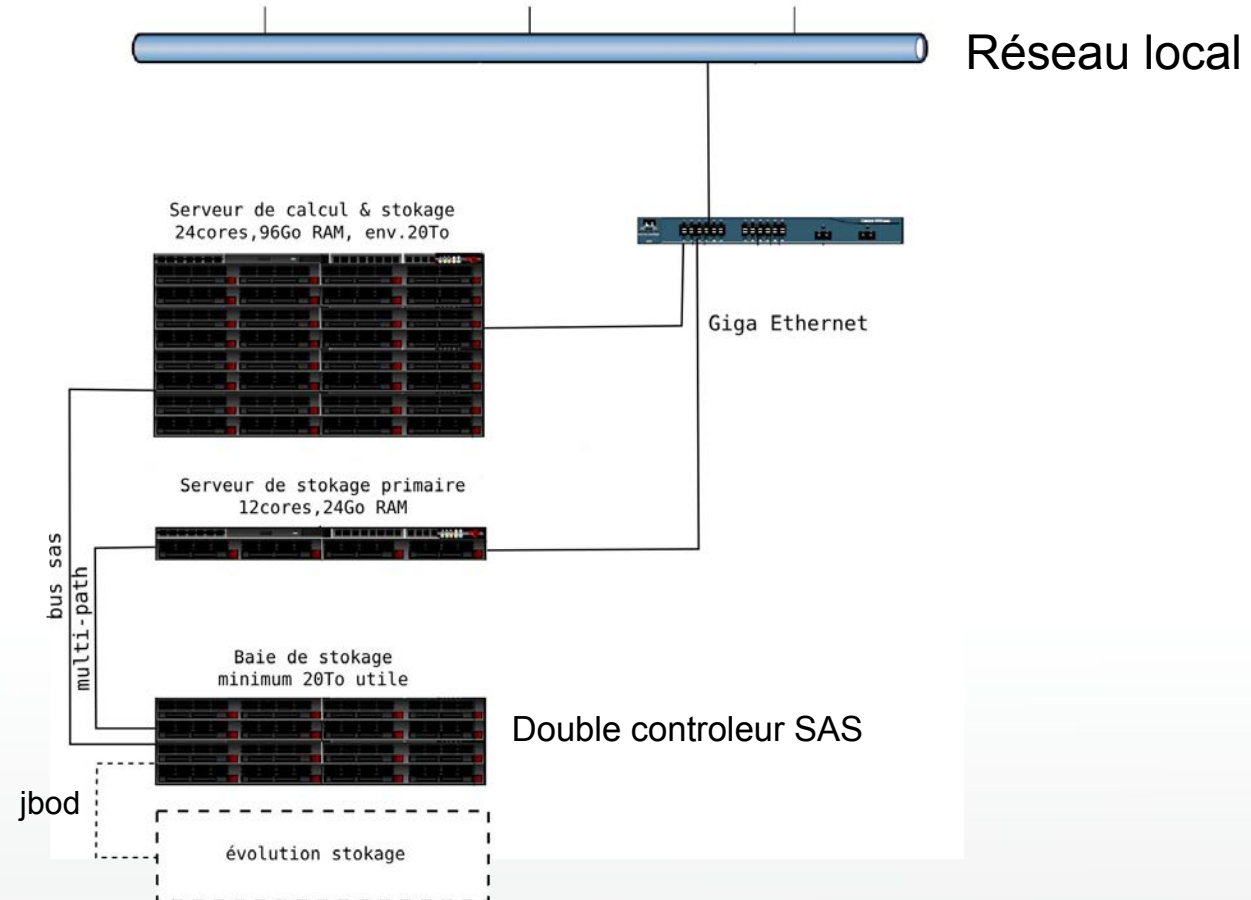
(Cahier des charges simplifié)

- Calculateur haute-performance (HPC) lié à un système de stockage de données évolutif alliant la performance et la sécurité.
- Administration et infrastructure simplifiée(*) pour du calcul massivement séquentiel.
- Quantitatif minimum : 24 cpu , mem : 4Go/cpu, stockage 20To rapide.
- OS & Softs 100% Open source.
Linux + Batch Scheduler (HTCondor) + Monitoring (Ganglia)
- Stockage primaire indépendant et évolutif de 20To sécurisé (raid6)
- Budget de 44K€ HT clé en main.

(*) Architecture "SAN" exclue.



Système de Calcul HPC et de stockage



TRANSTEC

**Lynx CALLEO
Datacenter R440**

**Lynx CALLEO
Application 1260**

**Infortrend
DS16S-R2240**

**Eaton Pulsar
MX 5000 RT**



Serveur de calcul & stokage (545Gflops)

4x Xeon E7-4830 de 8 cores = **32 cpu**
128 Go ram _ 4Go/cpu
2x HD SAS 600Go–15 000tr/mn 1Ctrl Raid1-OS
8X HD nearline sas 3To raid5 = **21To**
2X contrôleurs SAS raid5 - Data.
1X carte sas attachement baie(multipath)
Dual gigabit Ethernet (bonding)

Serveur de stokage primaire

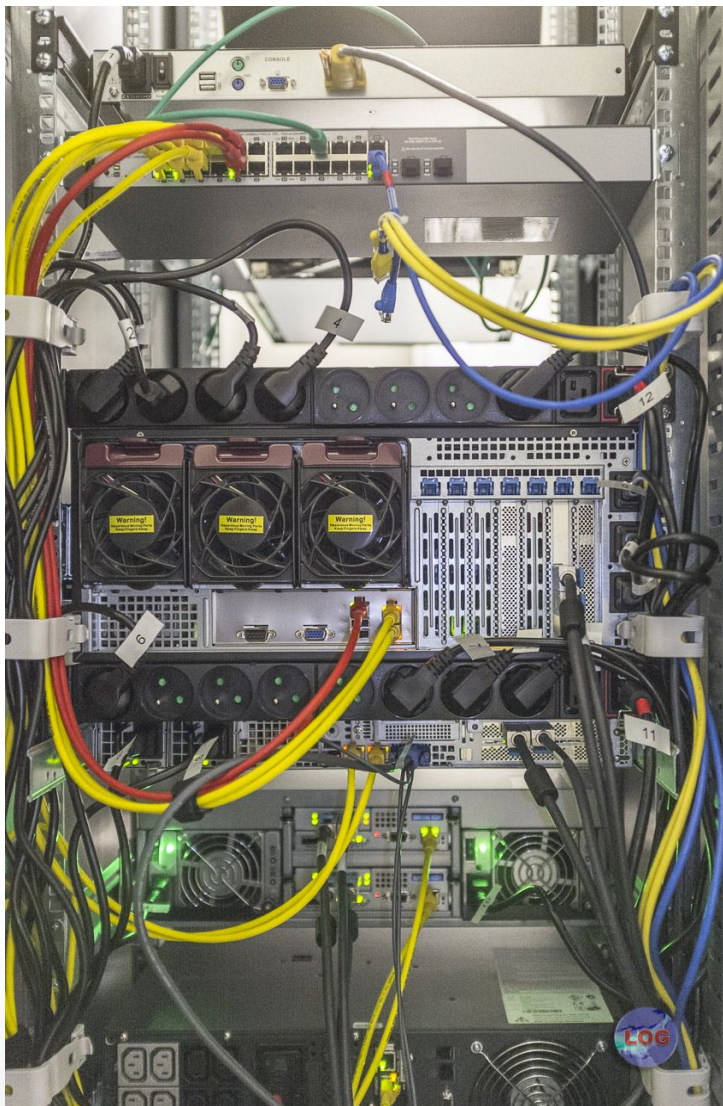
2x Xeon E5-2650 de 8cores - 64 Go ram
2x HD SATA 500Go - 1 Ctrl Raid 1 - OS
1x carte sas attachement baie (multipath)
Dual gigabit Ethernet (bonding)

Baie de stockage primaire

16x 2To = **28To**
2x contrôleurs SAS 6Gb/s redondant hot/swap
Raid 6
Extensible à 112 disques. (Jbod)

Onduleur (charge 60%)

Console RackAccess 19" TFT + Armoire Rack



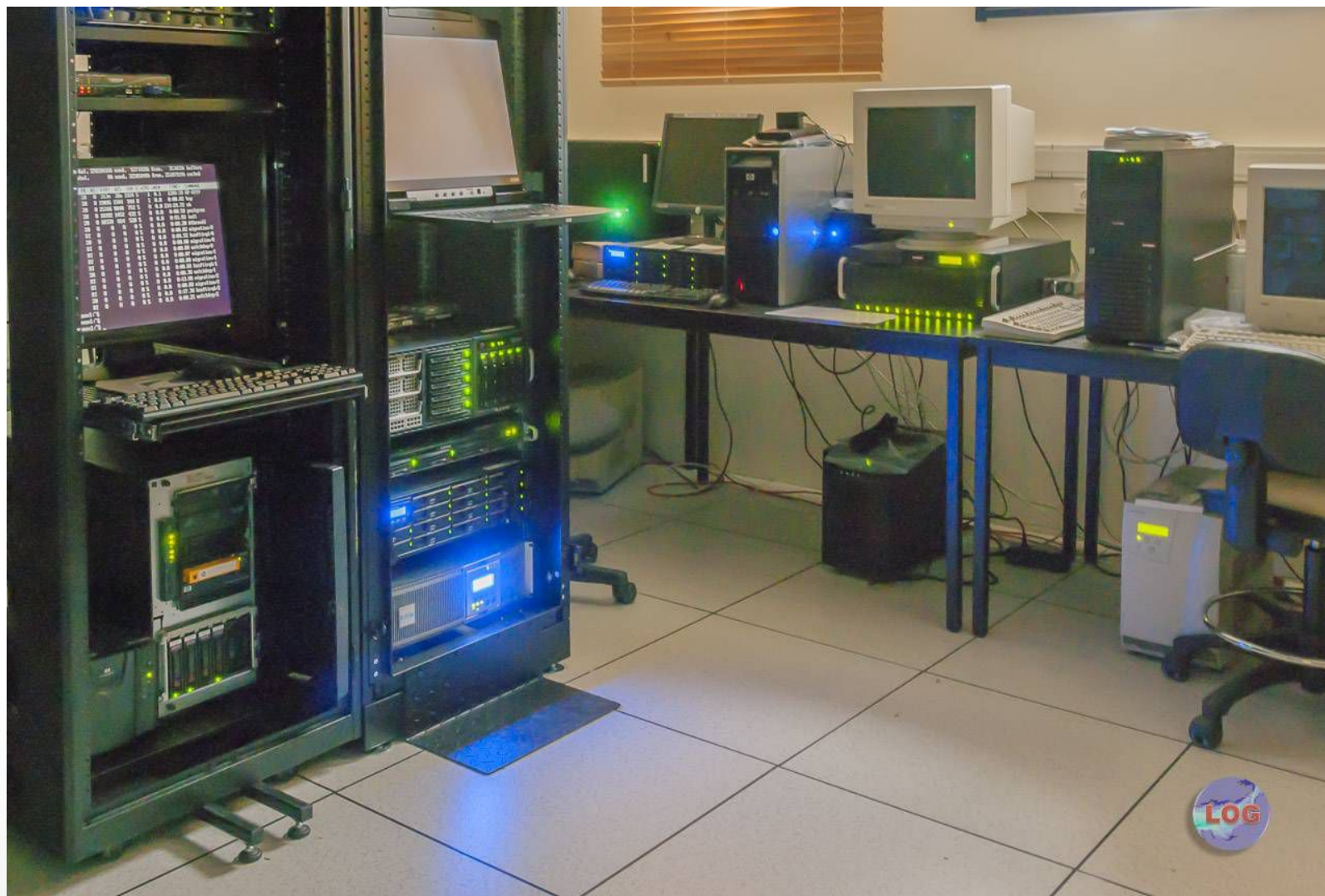
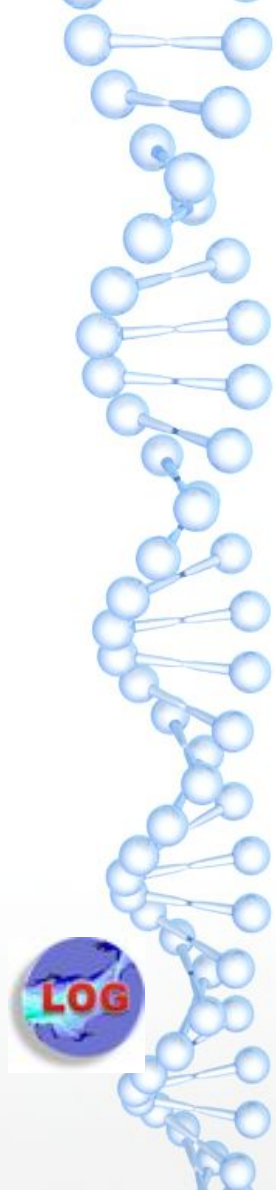
Switch Giga Ethernet

Serveur de calcul

Serveur de stokage primaire

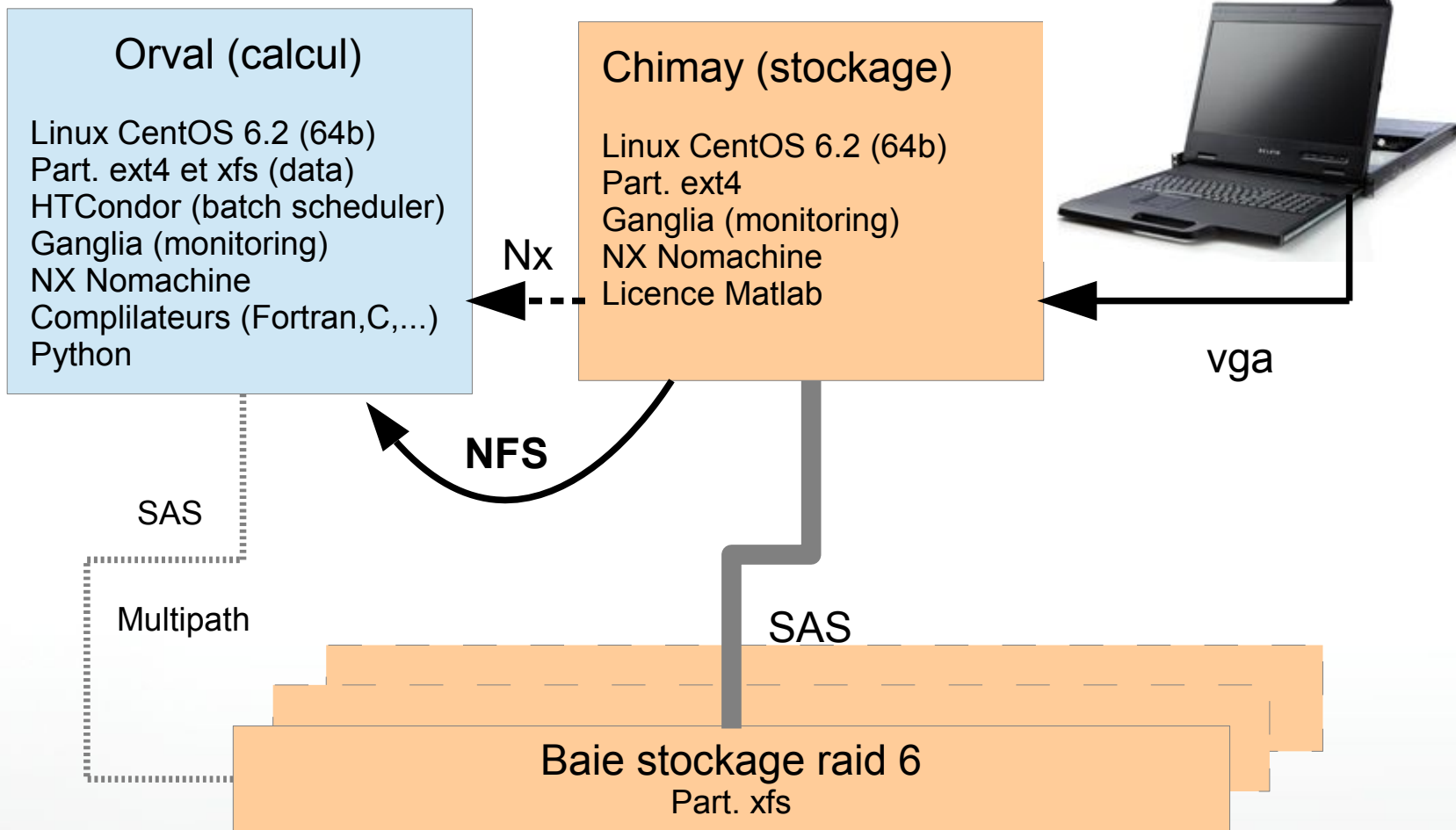
Baie de stockage primaire

Onduleur



MIn2RIEN – Mise en place d'un cluster – 13/02/2014

Configuration logicielle des serveurs



Configuration logicielle des serveurs

Orval (calcul)

Linux CentOS 6.2 (64b)
Part. ext4 et xfs (data)
HtCondor (batch scheduler)
Ganglia (monitoring)
NX Nomachine
Compilateurs (Fortran,C,...)
Python

~~Ormay (stockage)~~

~~Linux CentOS 6.2 (64b)
Part. ext4
Ganglia (monitoring)
NX Nomachine
Licence Matlab~~



vga

SAS

Multipath

Baie stockage raid 6

Part. xfs



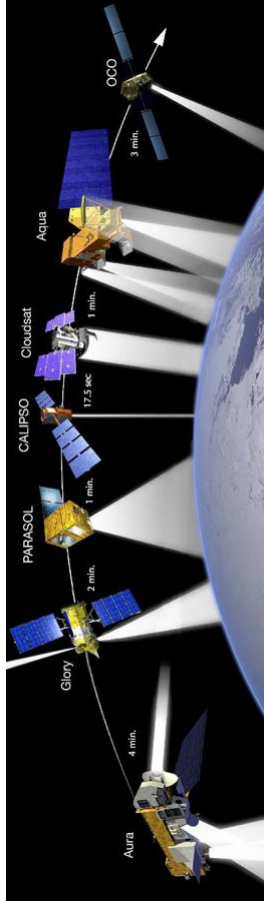
TEST

```
sion@orval:~  
  
1  [|||||||97.4%] 9  [|||||||##*78.8%] 17 [|||||||100.0%] 25 [|||||||96.8%]  
2  [|||||||100.0%] 10 [|||||||99.4%] 18 [|||||||96.8%] 26 [|||||||***** 44.9%]  
3  [|||||||100.0%] 11 [|||||||100.0%] 19 [|||||||100.0%] 27 [|||||||92.4%]  
4  [|||||||94.2%] 12 [|||||||87.8%] 20 [|||||||100.0%] 28 [|||||||100.0%]  
5  [|||||||100.0%] 13 [|||||||100.0%] 21 [|||||||100.0%] 29 [|||||||100.0%]  
6  [|||||||100.0%] 14 [|||||||100.0%] 22 [|||||||100.0%] 30 [|||||||100.0%]  
7  [|||||||100.0%] 15 [|||||||* 69.4%] 23 [|||||||100.0%] 31 [|||||||93.5%]  
8  [|||||||80.1%] 16 [|||||||94.9%] 24 [|||||||100.0%] 32 [|||||||100.0%]  
Mem[||||||#***** 13108/129046MB] Tasks: 223, 153 thr; 31 running  
Swp[ 0/8191MB] Load average: 31.82 31.92 31.07  
Uptime: 172 days(!), 01:46:11  
  
PID USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command  
73747 david 30 10 340M 318M 1132 R 99.0 0.2 0:31.81 condor_exec.exe /mnt/nfs/data/swfs_v2010/L1_MANCHE-MDN/2004/S200409512271  
73809 david 30 10 355M 334M 1132 R 99.0 0.3 0:28.28 condor_exec.exe /mnt/nfs/data/swfs_v2010/L1_MANCHE-MDN/2004/S200409812504  
73904 david 30 10 326M 307M 1132 R 99.0 0.2 0:20.10 condor_exec.exe /mnt/nfs/data/swfs_v2010/L1_MANCHE-MDN/2004/S200410212160  
73953 david 30 10 292M 275M 1132 R 99.0 0.2 0:18.08 condor_exec.exe /mnt/nfs/data/swfs_v2010/L1_MANCHE-MDN/2004/S200410411584
```

\$ htop

```
$ more calc.sh  
#:/bin/sh  
echo 32768^65536 | bc -l  
  
$ condor condor_submit job.cmd
```

```
$ more job.cmd  
Universe = vanilla  
Executable =/home/sion/condor/calc.sh  
Error = error.log  
Log = test.log  
output = output.log  
request_cpus = 1  
request_disk = 20K  
request_memory = 100M  
Queue 100 ( Nbr. de boucles)
```



Contexte

Les utilisateurs

Equipe : *Océanographie physique, transport et*
Télédétection

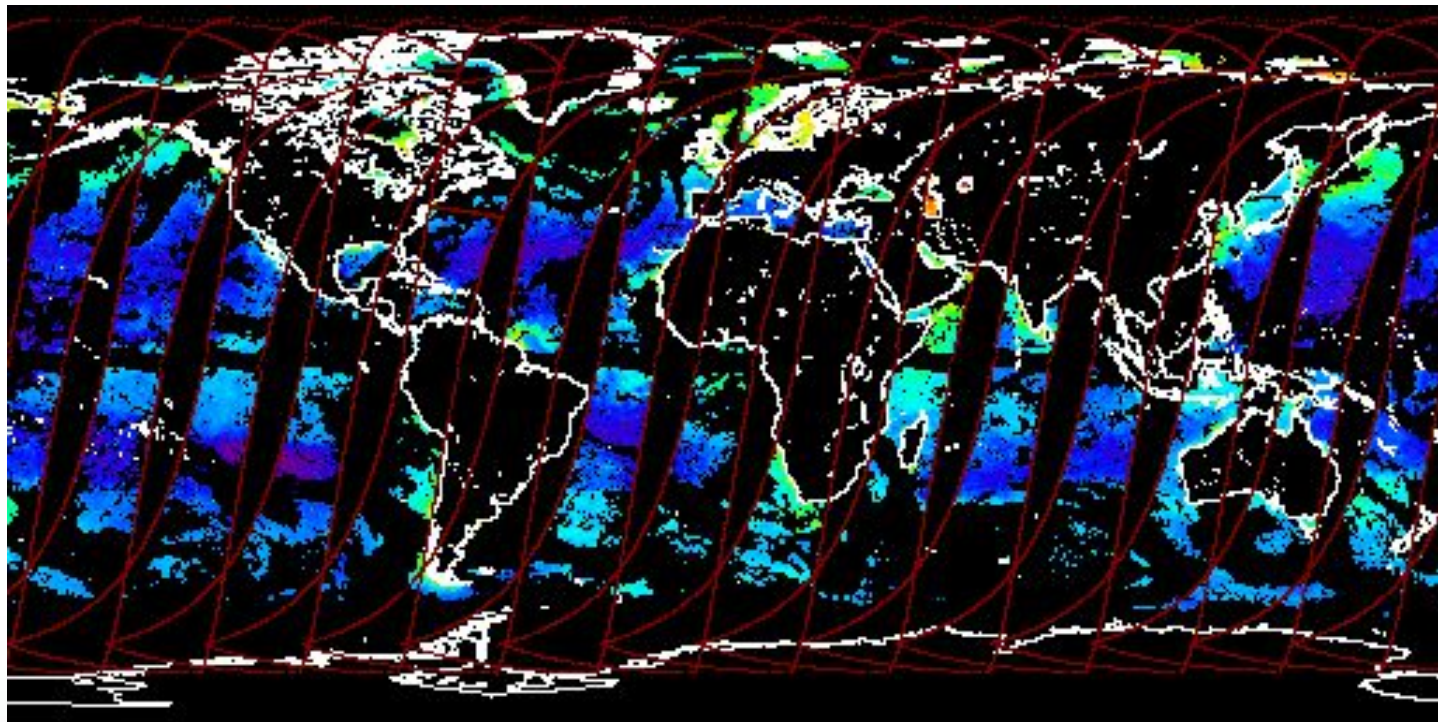
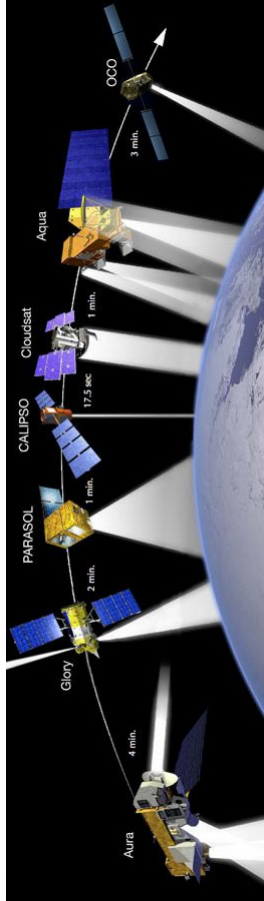
- 7 chercheurs permanents
- 5 Ingénieurs et post-doctorants
- 5 Thésards.

Tous utilisent les ressources informatiques du laboratoire, une majorité a besoin de données satellitaires.

1 Ingénieur Informaticien.

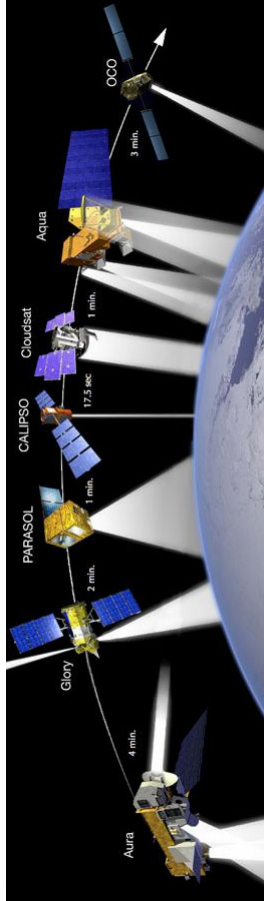
Contexte

Données Satellitale en Couleur de l'Eau



Une journée de mesure SeaWiFS





Contexte

Données Satellitale en Couleur de l'Eau

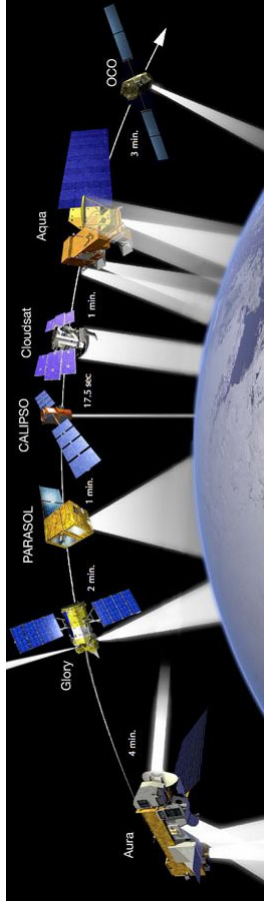


2002 01

Traitement de l'archive GlobCoast :

- ~ 2 000 000 pixels côtiers / jours
- ~ 60000 fichiers
- 500Go d'input
- 1,5To d'output

30"~45" le traitement d'une image =>
25 jours pour l'archive complète



Contexte

Données Satellitale en Couleur de l'Eau

→ Fichiers au format HDF
(avec compression gzip incluse)



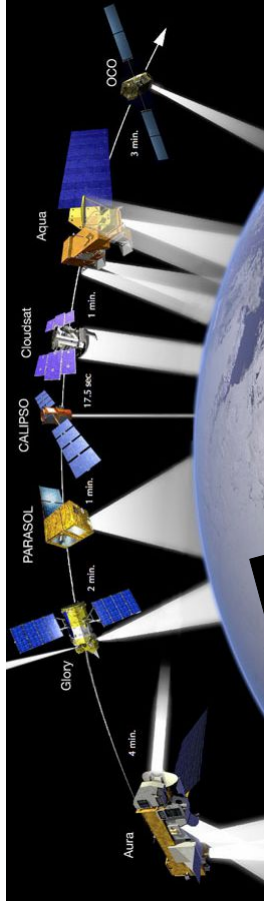
→ Programmes principaux codés en **C**

→ Script en bash ou python

BASH



Traitement individualisé pour chaque fichier image
(temps de process en 30sec et 10min)



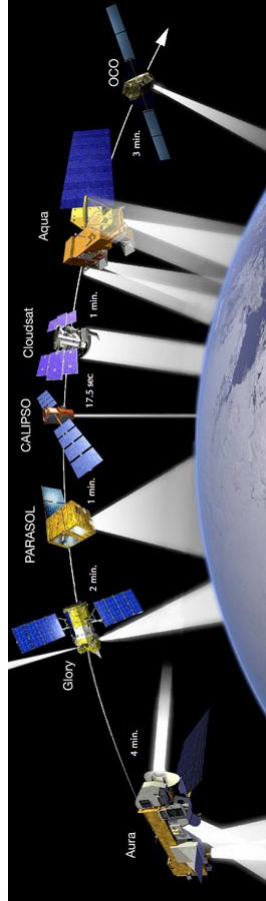
Contexte

Données Satellitale en Couleur de l'Eau

→ Fichiers au format HDF
(avec compression)

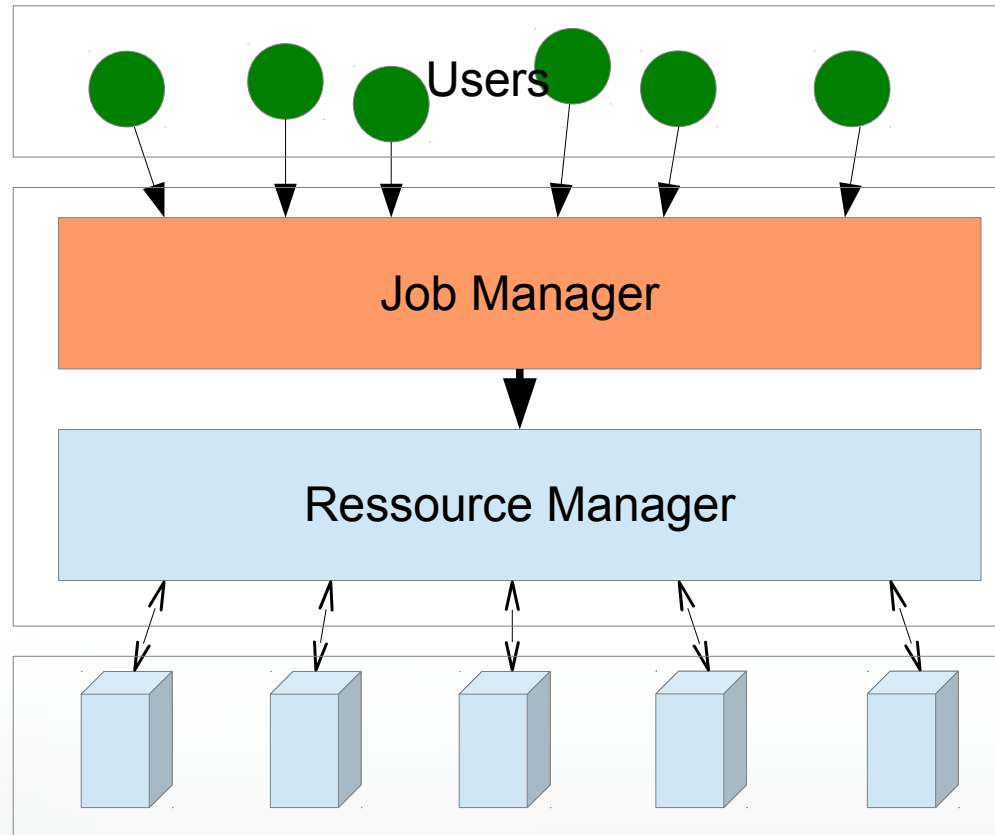
Nombreux processus
Input/Output important
Peu de pression de vitesse de calcul
=

Un candidat idéal pour le
High-throughput computing
(process en 30sec et 10min)



HTC

C'est quoi?



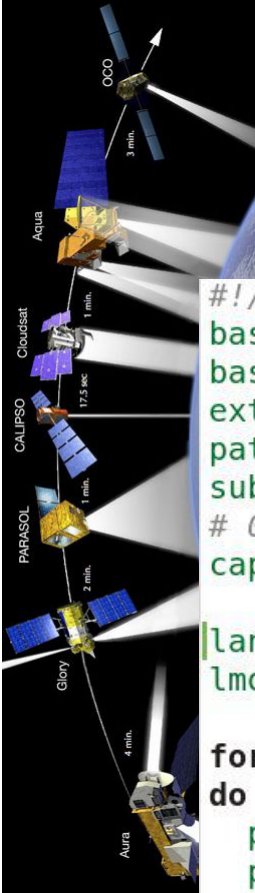
Clients
submission

Server
*Scheduling
Log, Accounting*

*Launching, control
Monitoring*

Computing nodes

- Utilisé par un de nos partenaires  **HYGEOS** ayant une bonne experience du HTC
- Open source (Apache licence V2)
- Simple et adapté à nos besoins
- Evolutif (prévu pour le multi-node)



```
#!/bin/bash
basepathin=/mnt/nfs/data/GLOBCOAST/SEAWIFS_HILT
basepathout=/mnt/data/david/GLOBCOAST/SEAWIFS_HILT
extension='globcoast.L3_MLAC_OC.hdf'
pathcode=`pwd`
submit=submit_processL3.cmd
# 0 = SeaWIFS
capteur=0

lannee=(1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010)
lmois=(01 02 03 04 05 06 07 08 09 10 11 12)

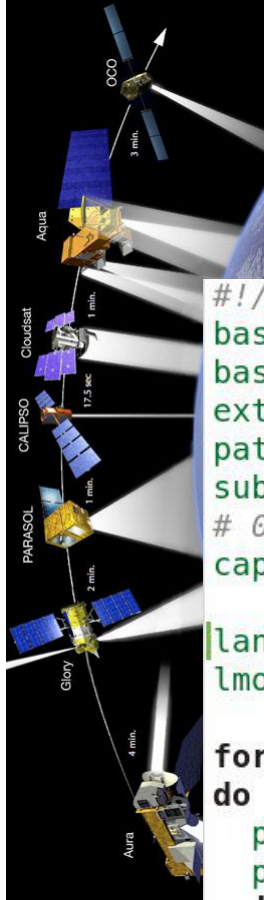
for A in ${lannee[*]}
do
    pathin=${basepathin}/${A}
    pathout=${basepathout}/${A}
    if [[ ! -d $pathout ]]; then mkdir -p $pathout; fi
    for M in ${lmois[*]}
    do
        process_L3_GC ${capteur} ${pathin} S${A}${M} ${extension} $pathout/S${A}${M}_${extension}
    done
done
```



```
#!/bin/bash
basepathin=/mnt/nfs/data/GLOBCOAST/SEAWIFS_HILT
basepathout=/mnt/data/david/GLOBCOAST/SEAWIFS_HILT
extension='globcoast.L3_MLAC_OC.hdf'
pathcode=`pwd`
submit=submit_processL3.cmd
# 0 = SeaWIFS
capteur=0

lannee=(1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010)
lmois=(01 02 03 04 05 06 07 08 09 10 11 12)

for A in ${lannee[*]}
do
    pathin=${basepathin}/${A}
    pathout=${basepathout}/${A}
    if [[ ! -d $pathout ]]; then mkdir -p $pathout; fi
    for M in ${lmois[*]}
    do
        process_L3_GC ${capteur} ${pathin} S${A}${M} ${extension} $pathout/S${A}${M}_${extension}
    done
done
```



```
#lannée=(1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010)
lannée=(2002)
lmois=(01 02 03 04 05 06 07 08 09 10 11 12)
```

```
for A in ${lannée[*]}
do
    pathin=${basepathin}/${A}
    pathout=${basepathout}/${A}
    if [[ ! -d $pathout ]]; then mkdir -p $pathout; fi
    for M in ${lmois[*]}
    do
```

```
        echo "getenv          = True" > $submit
        echo "executable     = ${pathcode}/process_L3_GC" >> $submit
        echo "arguments      = ${capteur} ${pathin} S${A}${M} ${extension} $pathout/S${A}${M}_${extension}" >> $submit
        echo "universe         = vanilla" >> $submit
        # echo "output          = $pathout/S${A}${M}_output.txt" >> $submit
        # echo "error           = $pathout/S${A}${M}_error.txt" >> $submit
        echo "output            = /dev/null" >> $submit
        echo "error            = /dev/null" >> $submit
        echo "log              = /dev/null" >> $submit
        echo "queue" >> $submit
        condor submit $submit
```

```
    done
done
```


Un marteau piqueur pour ecraser une mouche?

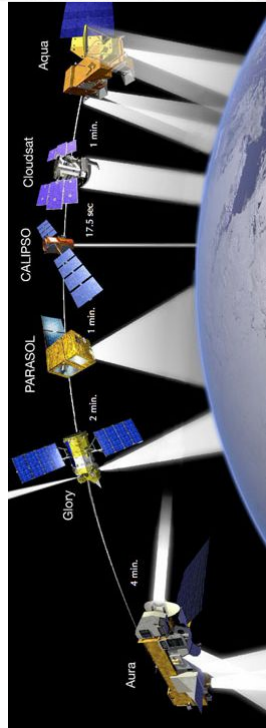
Prise en main ultra rapide

Il s'occupe de tout (scheduling, priority, ressource management)

Potentielles evolutions importantes

Beaucoup plus léger qu'un marteau piqueur





Merci



Min2RIEN – Mise en place d'un cluster – 13/02/2014

R. Sion
2010