



# Update on HPC configuration at Météo-France

Alain BEURAUD  
HPC project manager, Reading, 27 October 2014

## Outline

« HPC 2013 at Meteo-France » : our initial objectives

Data centers preparation

The HPC procurement

The new HPC configuration


Milestones of the project, lessons learnt

Next steps



**The starting point : MF's HPC configuration until 2012**

- HPC configuration based on NEC systems (SX8R and SX9), 42 Tflops peak :
  - 2 SX9 computers of 10 vector nodes each (1 TB of memory per node)
  - 1 SX8R computer of 32 vector nodes (128 GB of memory per node)



- One single data center, located on our site (Toulouse West)
- The computing center was quite saturated : impossible to run simultaneously the new and the previous HPC configuration in our « historical » facility

**METEO FRANCE**  
Toujours un temps d'avance

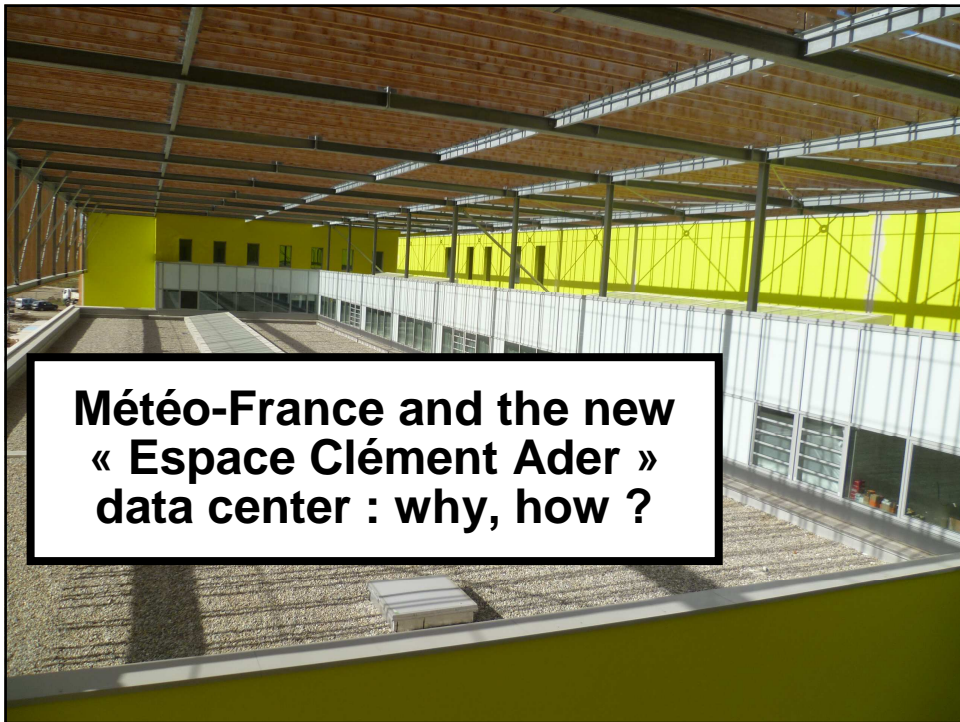
## HPC 2013 : initial objectives

- The aim is to fulfill the HPC requirements of MF for 5 years (2013 - 2018)
- The HPC upgrade has to be carried out in 2 steps, with an upgrade at mid-term contract : in relation to scientific commitments, the efficient ratio has to be at least from 10 to 12 in phase 1, at least from 2 to 3 in phase 2.
- HPC configuration has to be divided in 2 clusters (1 machine dedicated to operations, the other one devoted to research activities), with the possibility to run operationnal suites on the « research » cluster. This topology saves a 24 x 7 maintenance from the provider.
- Since our computing center was almost saturated, one of the 2 clusters will be installed in a new data center (Espace Clement Ader, new research building owned by the University of Toulouse, delivery scheduled in 2013)
- The new HPC configuration has to run the operationnal suites by the end of Januray 2014 (end of NEC contract : 2014, February 6)

## HPC2013 : a large range of activities

Consequently, this project was divided in several kinds of activities :

- the preparation of the data centers (building works in the old computer hall, and supervision of the new data center building works)
- the lead of the procurement for the choice of the new HPC solution
- the migration of the applications to the new architectures (several millions of lines to analyse – to check or to rewrite)
- the monitoring of the different commitments included in the contract (performances, schedule, trainings, user support, maintenance, ...)
- the move of the operationnal suites on the new HPC configuration.



## ECA : the story started without us

2010 :The University of Toulouse and the Midi-Pyrénées region planned to build a new facility for hosting different research teams previously dispersed in different places :

- offices and technical platform for more than 200 searchers
- laboratories for structural mechanics and strenght of material studies (particularly for aircraft technology)
- laboratory for microcaracterisation and nano technology studies (installation of new electron microscope, microprobe,
- data center shared between different partners (University, Public Research,... but also public agency or industrial ..).

The new building, called ECA (Clement Ader is one of the pioneers of aviation, born in Muret, 15 km South from Toulouse), should be delivered in 2013.



## In the meantime, thinking about future HPC infrastructure at Météo-France ...

2009 : the data center is almost saturated with the full NEC configuration (necessity to add 7 SX9 nodes to reach the level of performances).

2010 : our market analysis let us think that our budget would allow us to install in 2013 a new HPC configuration whose peak rate could be around 1 Petaflops.

In our « old » data center :

- space floor quite saturated
- installation not ready for liquid cooling
- electricity configuration to be upgraded and rationalized

⇒ Many solutions were studied :

- ⇒ Installation in « containers » ?
- ⇒ Service contract for renting HPC hours directly from a vendor site ?
- ⇒ New data center on our own site ?

2010 : our questions to vendors concerning infrastructure were :

- 1 Petaflops for 1 MW IT in 2013, will it be OK ?
- water cooling with 13°C / 18°C temperature, will it be OK ?

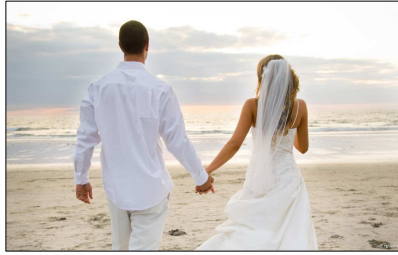
At that time, the answers were not very clear ...

## And their paths crossed ...

Mid 2010 :

- Météo-France hears about the University Project
- the University hears about Météo-France's search for a new infrastructure

... they met and took the decision to unite their destinies.



The requirements of the groom :

- a new data center available on 2013, January the 1st.
- a useful space floor of at least 500 m<sup>2</sup>, with 1MW IT
- a further possible extension

## The difficulties of the young couple

Complex financial closure :

- important gap between initial financial estimation and actual architectural costs
- final agreement occurred only mid-2011

Evolutions of electricity configuration :

- only one external power supply link (instead of 2 initially planned)
- generators only usable in case of failures / tests (no « EJP » mode)

Changes in the hydraulic design :

- initially, only one hydraulic circuit for liquid cooling (upgraded to 4)
- difficulties to design the final pipes before the choice of the supercomputer

Many difficulties around the agenda for the delivery of the building :

- the initial target of 2013, January 1st, appeared as unrealistic
- analysis led to have a partial delivery (for the data center) in March
- global delivery finally delayed in November ...

## Summer 2011 : final strategy concerning infrastructures

The vendors requirements (energy, space floor) become less important, so it appears possible to install one of the 2 clusters in our « old » data center (even if important works have to be conducted).

The « old » data center will have to be ready in March 2013, in order to host the first cluster. The aim is to be able to migrate the full operational suites on this new system before the end of the NEC contract.

The second cluster will be installed in the new data center (Espace Clément Ader) as soon as the new facility will be available => if an additional delay occurs, it will not have any impact on the duration of the NEC contract.

The gap between the 2 installations has an important consequence : the first cluster will have to host operations without any backup cluster => special emphasis on high level of reliability.

Since reliability becomes more and more important, both data centers had to evolve toward a « N + 1 » architecture for generators, inverters, chillers, ...

## Preparing the infrastructures : the new ECA data center



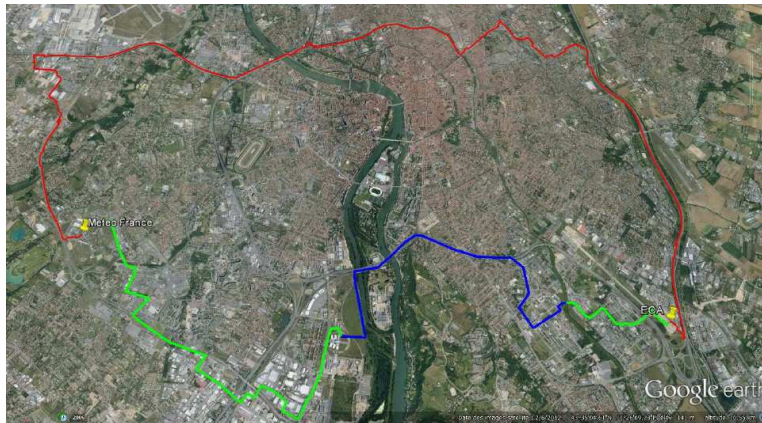
June 2012



ECA building delivery : 2013, Nov 27th



## Preparing the infrastructures : the links between the 2 data centers



Direct link : blue + green  
Backup link (not activated for the moment) : red

## Preparing the infrastructures : the « old » data center (CNC)

Necessity to take into account the previous systems (not only HPC systems, but also data-handling systems, servers, switches, networks, ...) => impossible to upgrade the air temperature of the data center

Works have to be done a few meters from operational computers

No possibility to have a cable path below the ceiling

Only 52 cm to arrange electricity cable, chilled water pipes, Infiniband cables and Ethernet networks

And no chilled water pipe available in the false floor at the beginning of the project

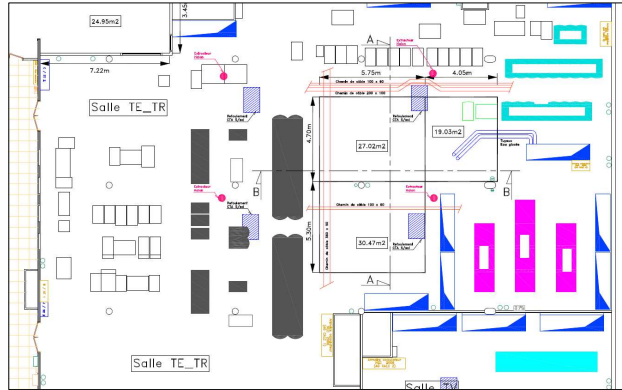


## Preparing the infrastructures : the « old » data center (CNC)

Upgrade and rationalization of electricity configuration

Addition of a new local for inverters, and air cooling installation for this local

Installation of a liquid cooling solution



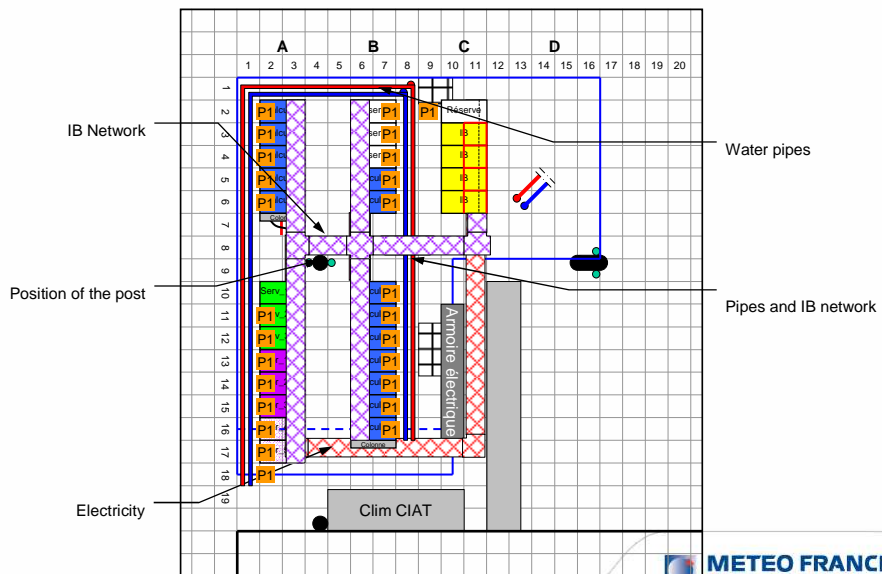
Works had to be done at a few meters only from the NEC computers running the operational suites

Delivery committment :

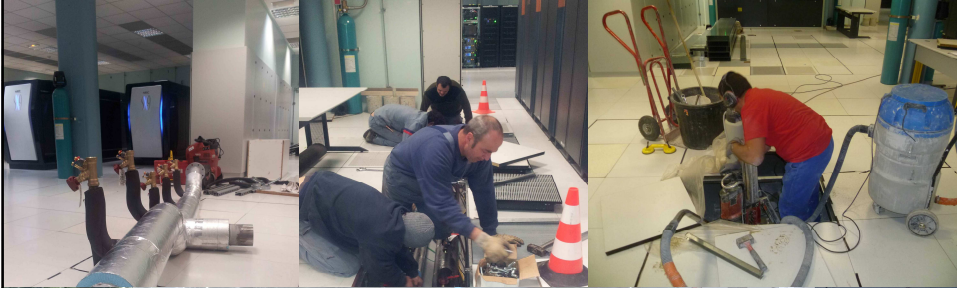
End of February 2013



## Preparing the infrastructures : the « old » data center (CNC)



## Preparing the infrastructures : the « old » data center (CNC)



Very complex building works, to be performed while the operations continue to run on other systems.

The challenge was met,  
the schedule was respected.



The procurement for the choice  
of a new HPC provider

## HPC Procurement : major guidelines

Nature of the contract : service contract

Duration of the contract : 5 + 1 years

Phases : 2 (2 computers per phase, located in 2 sites)

Phase 1 :

Starts with only one computer (ECA building not available in March 2013)  
This computer has to host both Research and Operations during 6 months  
The migration of operations have to be completed by January 2014  
The second computer has to arrive 8 months after the first one

Phase 2 :

Vendors had the possibility to extend or to replace the materials installed in Phase 1

Annual fees indicated to vendors

Maintenance : 12 hours a day – 5 day / 7.

Initial training and support are included in the contract

## HPC Procurement : performances analysis

Operational set : 3 applications in their 2014 expected configuration

(AROME HR Forecast, ARPEGE HR Forecast, ARPEGE HR 4D-Var) to be run in a limited time frame : 2100 seconds in Phase 1, 1200 seconds in Phase 2.

Research set : 7 applications (6 « meteo », 1 « Mercabr »), 2011 version

(Assim. ARPEGE MR, ARPEGE HR Forecast, ARPEGE Ensemble Forecast, ARPEGE ensemble assim, Arome MR, AROME ensemble, ORCA12)

Additional commitments

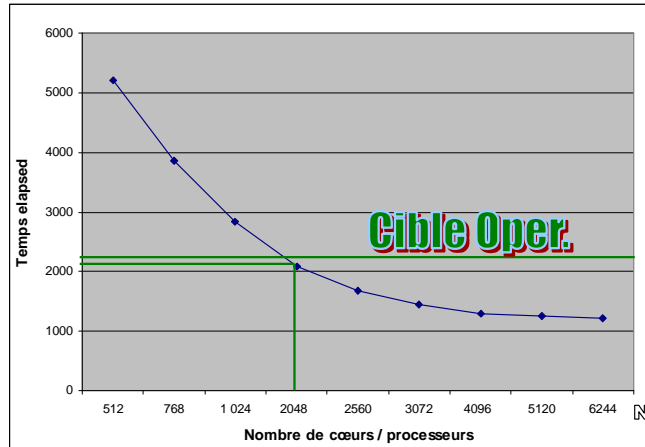
Oceanography (Mercator) operational performances

Code should run in less than 20 mn and use a max of 30% of the operational computing facility

## HPC Procurement : performances analysis

Performances in operational mode :

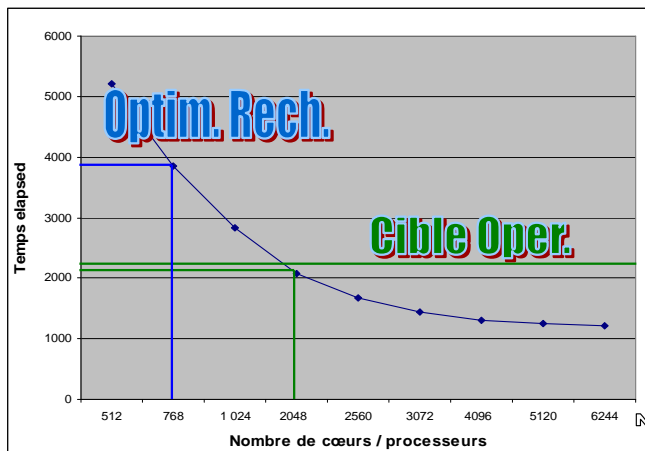
maximum efficiency for the 3 codes of the “operational set”



## HPC Procurement : performances analysis

Performances in research mode :

maximum throughput for the 7 codes of the “research set”



## HPC Procurement : criterias for the final choice

The selection was based on a set of criterias, through a public procurement :

- ▶ **Technical analysis :**
  - benchmark on "operational set of applications"
  - benchmark on "set of research applications"
  - technical architecture (memory, interconnect, I/O)
  - robustness and operational constraints compliance
  - tools for administrators and for users
  - project management
  - quality of services
  - maintenance strategy
- ▶ **Cost analysis**
  - direct costs – including training, maintenance, services,
  - indirect costs – electricity consumption, )

All the major HPC companies applied for this procurement

4 companies made a complete final proposal (full benchmark results, complete commitments, ...)

## The new HPC solution

## And the winner is : ATOS



## Phase 1 : key figures per cluster

- 522 and 513 TFlops peak
- Bullx DLC racks
- Intel® Ivy Bridge EP CPUs (**12 cores, 2.7 Ghz, 130 W**)
- 32 Gb memory/node (928 or 910 nodes), 128 Gb (72 nodes) ou 256 Gb (8 nodes), DDR3 RAM at 1866 Mhz (except Beaufix's large nodes)
- Interconnect InfiniBand FDR
- A set of service nodes
  - 2 Management and administration nodes
  - 6 Login nodes : access to computer / compiling platforms
  - 4 I/O nodes
- Software bullx Supercomputer Suite Advanced Edition
- Batch Management Software : SLURM
- MPI : IntelMPI used by almost all the different applications
- Debugging tool : DDT and MAP (from Alinea)

## Phase 1, cluster 1 : Beaufix



The new HPC system :

- BULLx DLC710
- Intel/Xeon Ivy Bridge
- no GPU, no Xeon Phi
- 1080 nodes (32, 128 or 256 GB)
- 24192 cores (2.7 Ghz)
- 469 Tflops Linpack (89.7% peak)
- DLC technology (15°/ 20°C)

Temporary storage based on parallel file system Lustre :

2 PB (NetApp E5500), bandwidth of 69 Gbps

NFS Permanent storage : 209 TB (NetApp FAS6280)

Interconnect based on Infiniband fat-tree architecture FDR fat (3 Mellanox top switch SX6536)

After code migration and porting, and strict control of the results provided on NEC and BULLx, operations have been switched on the new system on 2014, January the 14.

NEC SX8R and SX9 have been definitively stopped on 2014, February the 6.

 **Toujours un temps d'avance**

## Phase 1, cluster 2 : Prolix



The new HPC system :

- BULLx DLC710
- Intel/Xeon Ivy Bridge
- no GPU, no Xeon Phi
- 990 nodes (32, 128 or 256 GB)
- 23760 2.7 Ghz cores
- 464 Tflops Linpack (90.7% of the peak)
- DLC technology (13°/ 18°C)

Temporary storage based on parallel file system Lustre :

1,5 PB (NetApp E5500), bandwidth of 46 Gbps

NFS Permanent storage : 135 TB (NetApp FAS6280)

Interconnect based on Infiniband fat-tree architecture FDR fat (3 Mellanox top switch SX6536)

Acceptance tests for performances were achieved on February, 15.

Acceptance tests for reliability (duration : 30 days) finished on Mars, 31.

Operational runs started on April, 23.

 **METEO FRANCE**  
Toujours un temps d'avance

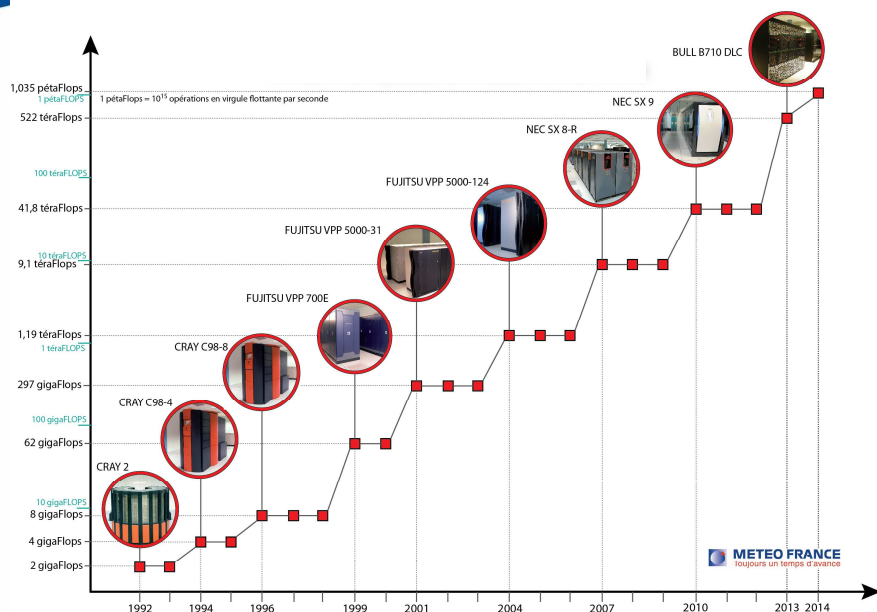
## Phase 1 : ratio flops/watt

Prolix Linpack: 465 Tflops (89.7% peak),  
and electricity consumption measured : 395 Kw  
=> almost 1175 Mflops/watt

CPU Idle (services nodes + disks subsystem + IB racks)  
electricity consumption measured : 184,5 Kw

Standard operations (nodes load around 70 %) :  
electricity consumption measured : 220 Kw

## Evolution of HPC peak power for Météo-France







**Main milestones of the project,  
lessons learnt**



<b>HPC Project : year 2011</b>	
<b>S1 / 2011 :</b>	<ul style="list-style-type: none"> <li>preliminary RAPS results analysis ("pre-bench")</li> <li>writing of the first draft of the procurement</li> <li>preparation of the building works for 2 infrastructures</li> <li>evaluation of the porting effort for all the applications</li> </ul>
<b>Hot spots :</b>	<ul style="list-style-type: none"> <li>benchmarks contents : not too simple, not too complex</li> <li>technology watch : very difficult to have relevant informations from vendors about energy requirements, water cooling temperature, ...</li> </ul>
<b>S2 / 2011 :</b>	<ul style="list-style-type: none"> <li>initial proposals : analysis, oral presentations</li> <li>final specifications for the building works</li> <li>benchmark aménagement (repetability / reproductibility)</li> <li>safety of operations with only one cluster (specification)</li> </ul>
<b>Hot spots :</b>	<ul style="list-style-type: none"> <li>proper benchmark specification for a fair comparison of proposals</li> <li>detailed specifications of works without technology</li> </ul>
<ul style="list-style-type: none"> <li>knowing the HPC which will be installed</li> </ul>	

## HPC Project : year 2012

**S1 / 2012 :** customer site visits and suppliers visits  
writing of the final documentation of the contractual documents  
beginning of the building works for the new data center  
final proposals detailed analysis

**Hot spots :** euros/dollar and euros/yen parity  
keeping informations confidential until official signature

**S2 / 2012 :** reports to the steering committee, then to MF council  
completion of building works in the old computing center  
installation of the porting system

**Hot spots :** debriefings with unsuccessful candidates  
building works while operations run  
floor slab resistance : difficulties to find an agreement



## HPC Project : year 2013

**S1 / 2013 :** Beaufix : installation and performances tests  
RFP for the link between the 2 data centers  
Trainings for system administrators, operators, users  
Application per application : status of migration

**Hot spot :** Uncertainties about arrival of Ivy Bridge processors

**S2 / 2013 :** Beaufix : cluster open to end-users  
Application per application : end of the migration  
Installation and monitoring of a « mirror operational suite »  
(no scientific change versus the NEC operational suite)  
Prolix : beginning of the installation

**Hot spots :** LUSTRE and MPI tunings  
Differences of results between BULL and NEC suites  
Impurities in our cooling water on our old site

## HPC Project : year 2014

**S1 / 2014 :**  
14)

**D-day :** operational suites moved on Beaufix (January, 14)

Dismantling of NEC machines (from February, 6)

Prolix : performances and reliability tests

Prolix : installation of a mirror suite

Switch of operational suites on Prolix (April, 23)

**Hot spots :**

mandatory migration before end of January

operational suites run during 3 months without a backup cluster

**S2 / 2014 :**

I/O server in the operational suite

new DHS : beginning of the installation

operational suite installation of ARPGE-HR / AROME – HR pre-suite

**Hot spots :**

stability of large operational jobs duration

DHS : moving from DMF to HPSS

**METEO FRANCE**  
Toujours un temps d'avance

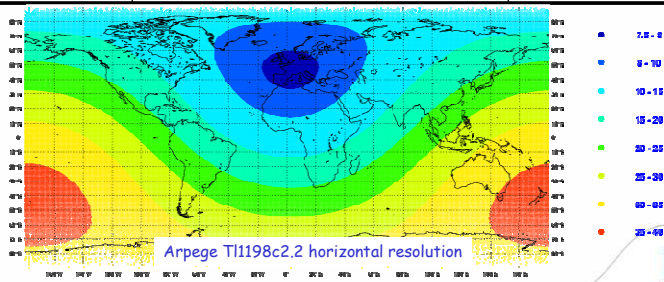


**What next ?**

**METEO FRANCE**  
Toujours un temps d'avance

## Scientific program : global forecasting systems

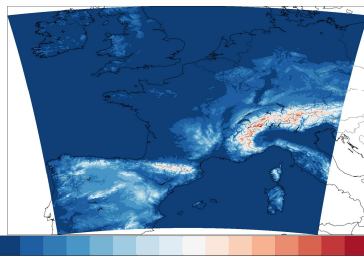
	2014	2015 / 2016
<b>ARPEGE</b> <i>Deterministic</i>	TI798c2.4 L70 (10km on W Europe) 4DVar (6h cyc): TI107c1L70 & TI323c1L70 Forecast hours : 102 / 72 / 84 / 60h	TI1198c2.2 L105 (7.5km on W Europe) 4DVar (6h cyc): TI149c1L105 & TI399c1L105 Forecast hours : 102 / 72 / 84 / 60h
<b>AEARP</b> <i>Ensemble Data Assimilation</i>	TI399c1 L70 ; 6 members 4D-Var (6h cycle): TI107c1 L70	TI479c1 L105 ; 25 members 4D-Var (6h cycle): TI149c1 L105
<b>PEARP Ens.</b> <i>Prediction System</i>	TI538c2.4 L65 (15km on W Europe) 35 members ; Forecast hours : 72 / 108h	TI798c2.4 L90 (10km on W Europe) 35 members ; Forecast hours : 72 / 108h



**METEO FRANCE**  
Toujours un temps d'avance

## Scientific program : regional forecasting systems

	2014	2015 / 2016
<b>AROME</b> <i>Deterministic</i>	2.5km L60 (750 x 720 pts) 3DVar (3h cycle) 5 forecasts per day up to 36h	1.3km L90 (1536 x 1440 pts) 3DVar (1h cycle) 8 forecasts per day up to 42h
<b>ALADIN-OM → AROME-OM</b> <i>Overseas Regional systems</i>	7.5km L90 - 3DVar (3h cycle) 2 forecasts per day up to 84h	2.5km L90 - Dynamical adaptation Forecast hours to be determined
<b>AROME-PI</b> <i>Nowcasting</i>		1.3km L90 (1536 x 1440 pts) 24 forecasts per day up to 6h
<b>PEARO</b> <i>Ensemble Prediction System</i>		2.5km L90 - 12 members Forecast hours to be determined



Arome 1.3km orography

**METEO FRANCE**  
Toujours un temps d'avance

## New DHS : HPSS will replace DMF

Joint proposal BULL / IBM :

- servers : BULL R423 (Intel Xeon EP Ivy Bridge E5 2650 V2)

- HSM software : HPSS 7.4.2

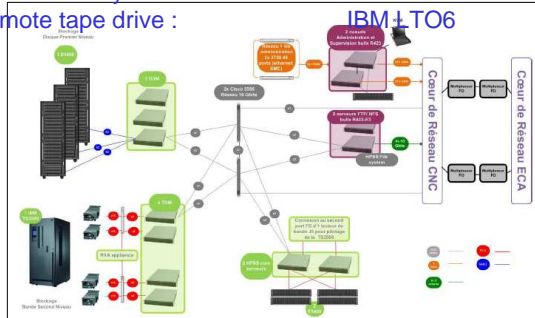
- disk solution : NET APP E5500 (with 6 TB disk NL SAS)

- library technology : IBM TS3500 + TS4500

- tape drive technology : IBM TS1150 (Jaguar 5)

- remote library : IBM TS3310

- remote tape drive : IBM LTO6



**METEO FRANCE**  
Toujours un temps d'avance

## New DHS : one upgrade per year until 2018

	2014	2015	2016	2017	2018
<b>Storage capacity (PB)</b>	26	42	72	120	180
<b>Number of files (Millions)</b>	266	345	478	744	1130
<b>Mean daily throughput (TB/day)</b>	226	452	904	1017	1130
<b>Double copy storage capacity (TB)</b>	200	240	280	465	800

DHS Migration schedule :

- October : Hardware installation, HPSS pre-installation
- November : Jaguar5 installation, and HPSS fonctionnal tests
- December : Acceptance tests, test of DMF metadata insertion in HPSS

HPSS solution

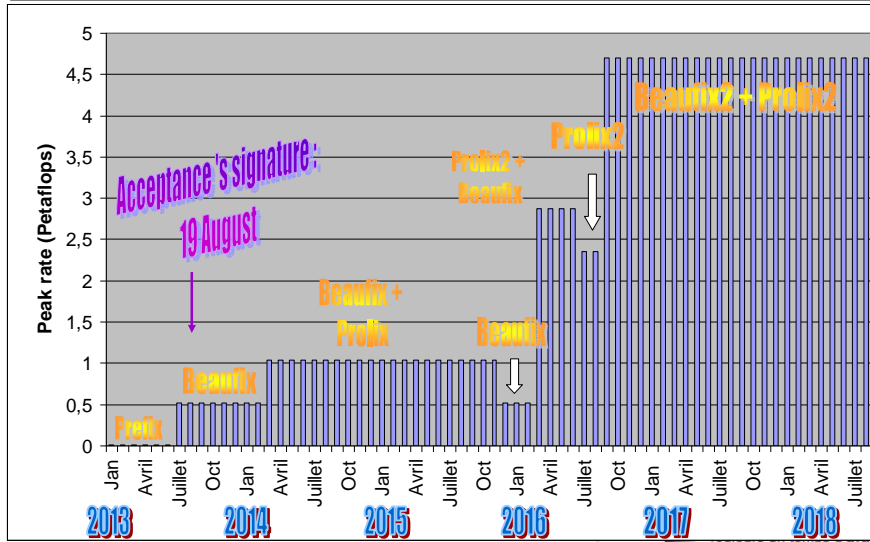
- Jan. 2015 : Moving from SGI/DMF to BULL-IBM/HPSS

- Dec. 2015 : End of data migration, end of the SGI contract

**METEO FRANCE**  
Toujours un temps d'avance

## HPC : Towards the Phase 2

Evolution of peak rate performance during the BULL contract



## HPC : Towards the Phase 2

**522 TFlops peak**  
 56 frames **bullx DLC**  
 1008 compute nodes  
 Intel® Ivy Bridge EP  
 Fat Tree InfiniBand FDR  
 Lustre 2 Po, 69 Go/s  
 Beaufix (06/2013)



**2,2 PFlops peak**  
 56 +45 frames **bullx DLC**  
 1800 compute nodes  
 Intel® Haswell or Broadwell EP  
 Fat Tree InfiniBand FDR  
 Lustre 3,57 Po, 138 Go/s

Upgrade Beaufix  
 (06-08/2016)



**513 TFlops peak**  
 55 frames **bullx DLC**  
 990 compute nodes  
 Intel® Ivy Bridge EP  
 Fat Tree InfiniBand FDR  
 Lustre 1,53 Po, 46 Go/s



Toulouse East

**2,2 PFlops peak**  
 55 +45 frames **bullx DLC**  
 1800 compute nodes  
 Intel® Haswell or Broadwell EP  
 Fat Tree InfiniBand FDR  
 Lustre 2,55 Po, 92 Go/s

Upgrade Prolix (12-2015/  
 02-2016)

## ECA and MF : a couple that will last ?

DHS : the library hosting the second copy of very important files has just been installed in the ECA data center

HPC : at the end of 2015 (or at the beginning of 2016), the upgrade of Prolix will be done inside the remote computing center



## Future use of the ECA remote data center : towards a complete backup solution ?

Météopole, main building, 2013 January 22, 16h50 :



## Future use of the ECA remote data center : towards a complete backup solution ?

Météopole, main building, 2013 January 22, 16h50 :



FRANCE  
Toujours un temps d'avance

## ECA and MF : a couple that will last ?

DHS : the library hosting the second copy of very important files has just been installed in the ECA data center

HPC : at the end of 2015 (or at the beginning of 2016), the upgrade of Prolix will be done inside the remote computing center

Recent incidents (fire, guard held hostage and threatened with a gun ...) on the Météopole could lead Meteo-France to build a complete disaster plan

Agreement between the University and Météo-France allow us to use the remote data center during at least 20 years

METEO FRANCE  
Toujours un temps d'avance



A photograph of a server room. In the foreground, there is a large, messy pile of orange cables on a light-colored tiled floor. In the background, there are several server racks with perforated metal doors. Some of the racks have green indicator lights. The ceiling is a standard drop ceiling with recessed lighting.

**Thanks for your attention**

**Any questions ?**