Tools on new MF computer: how partners will work on NEC

And many other system-related informations !

R. El Khatib and C. Fischer

Input to Toulouse interim cycles up to CY33
C.F. 15/03/2007

CY32T0; declared February 27th 2007:

- bugfix for limited area plane geometries (Aladin, Arome, Hirlam versions)
- bugfix for Aladin 3D-VAR
- B matrix fix (Loïk)
- Bugfix for Climate physics
- NEC portability and optimization (so called "necplus" branch)
- Catch-up of operational branch (CY31T1_op?)

CY32T1; declared March 12th 2007:

- ALARO0 update and rationalization of plug-ins (Note: call to ACDIFUS for implicit treatment of surface fluxes in vertical diffusion computations => *delayed for CY32T2*)

CY32T2; deadline for contributions on May 10th :

- adaptation of Climate model to new features of the Arpège-NWP model: prognostic cloud scheme (AA)
- 1D model
- other ALARO0 contributions: harmonization of cloud computation + call to ACDIFUS1/2 (to allow for implicit call to the surface from ALARO) + cleaning of some repetitive computations (RB+FBs)
- further improvements to the DDH tool and interface (TK+JMP) + bugfix identified by Mats
- cleaning in the tendency aggregation routines with respect to the Catry/Geleyn rules ? (CPTEND, CPUTQY) => most likely delayed
- a new simplified condensation scheme for TL/AD (only adjustment process related to condensation, without explicit microphysics) ? (CL+FBs)
- finalize the Arpège and Aladin-France TKE scheme, with a shallow convection facility (EB+FBs)
- possible introduction of a mass flux prognostic variable (related to the shallow convection) (YB+FBs)
- plug-in Bechtold's convection scheme in Arpège (YB+FBs)
- SURFEX technical and scientific plug-in: I/O, Arpège geometry, physics adaptations (JDG, GH, FBs)
- AROME: MASDEV4.7 from Méso-NH and new cloud and shallow convection scheme (YS + SMal)
- catch-up of phasing for Hirlam's physics plug-in (for CY32) => T. Wilhelmsson + S. Niemela

- dynamics: code cleanings by Karim: GFL setup in SUDIM/SUDYN, better rearrangement of LARCINBTL/AD, obsolete keys, removal of useless interpolations for (gw)_surf in the case NH+LGWADV=T
- LTRAJHR (high resolution trajectory) for Arpège surface fields ? (KY)

- tri-diagonal semi-implicit operator for large domain map factor variations (key "LESIDG" in LAM, only if ready via Hirlam collaboration) ?
- LLONEM in Aladin ESPCM/AD (also by Hirlam collaborator ? – Isabel Martinez -, pre-requisite to LESIDG)
- NFLEVL in Aladin ESPCHOR/AD (Aladin phaser)
- adjoint of the LAM SL advection scheme => F. Vana (Prague) via Karim's branch

- adaptation of the Variational Bias Correction scheme to Arpège and Aladin (VG)
- optimization of CPU for 3D-VAR FGAT (add an explicit increment to the time loops in CNT4TL/AD + switch off useless calls to STEPO) (BC)
- adaptation of the modified (normalized RH) humidity control variable to Arpège and Aladin (LB)
- adaptations in e131 for Arome dataflow (PBr)

- assimilation of METOP sensors
- monitoring of ASCATT (C. Payan)
- assimilation of IASI data
- assimilation of microwave channels over land – code for dynamical surface emissivity calculation (FK)
- assimilation of radial Doppler radar winds (EW+TM+CF)
- adaptation of "bator" in order to handle SEVIRI radiances without using MF GRIB formats (for Aladin partners) => G. Boloni and A. Trojakova (LACE)

- optimization in Full-Pos: some computations in post-processing mode are done in the first Full-Pos stage (in the departure grid) but require data from the arrival grid. These can then be poorly distributed over processors (arrival points but cast in the departure grid distribution), which can lead to load imbalance. The idea would be to add some transposition already in the Full-Pos first part in order to optimize these computations with respect to the arrival grid distribution. This development is a target for Arome post-processing. (REK+KY)

- changes in the FA Arpège/Aladin file system and GRIB interfacing in order to increase compression (useful for archiving and transmitting the LAM coupling files) (work by REO, via JMA's branch)

- improvements in configuration 901 for converting IFS/Tessel surface fields into Arpège/ISBA fields (work by L. Descamps and B. Chapnik, enters via BC's branch)

CY32T3 (September)

- first code towards a finite element non-hydrostatic semi-implicit operator ? (depending on the maturity of this topic in the LAM community)
- optimization in Full-Pos: bugfixes and further improvements. (REK+KY)
- modularization of LATTEXTL/AD (not urgent)

Switch to NEC: snapshots for external users
Marion Pithon, J.A. Maziejewski, other contributors …

Introduction:

The intensive computing system at Météo-France is composed of ***2 NEC systems***, ***each having 16 SX-8R vector nodes, a linux TX7 scalar front-end, a HP-UX batch handler and a GFS file server*** (also called "NAS head"). These two systems are fully identical. One of these system, called "sumo", is dedicated to operational tasks while the other one, called ***"tori" hosts R&D tasks***.

Access point open to users:

Users can work on the "tori" front-end and the 16 computing vector nodes.
Each of these vector nodes is a symmetrical multi-processor composed of 8 vector processors sharing a 128 Gb memory. The characteristics of a vector node are as follow:

- operating system: Super-UX
- 8 processors at 35 Gflops theoretical maximum performance each. ***The R&D theoretical peak performance is therefore of 4,5 Tflops***.

The nodes interconnect via a very efficient and high-performanceinternal crossbar network (IXS) with a maximum transfer rate of 2*8 Gb bidirectional/nodes.
The scalar front-end is the unique access entry point for users. The mainframe characteristics are as follow:

- Linux operating system (Suse)
- 16 cores Intel Itanium2 & 32 Gb memory.

Interactive work is only possible on the scalar front-end.

This front-end shall be used for compiling (cross compiler for SX8), for the submission of tasks on the vector nodes and for file transfer with either "cougar" or other access points of Météo-France's network. Tasks will preferably be launched in batch mode on the front-end (e.g. in the "compile" class), in the "ft" class for file transfers. The "vector" class is the only one for submitting jobs to the vector server.

Disk configuration:

The total user disk space is of 19 Tb. It is shared by all the vector nodes and the scalar front-end through the Global File System (*GFS*).
A local disk space of 256 Gb is available on each vector node, which can be used as a temporary working space for a "single-node" batch task (/localtmp). In this working space, data from a batch request will be automatically destroyed at the endof the request.
Anyone has accessto this disk space, for node-local specific use.
HOMEDIR: user's permanent separate storage area. DSI/SC/CC provides the back up of data stored there. HOMEDIR user's data are saved by the software «Time Navigator». Total volume amounts to 6Tb.
TMPDIR: temporary disk space, accessible from every node, which a user can use either during the lifetime of his job or his interactive session. This space is therefore not saved. Its total volume amounts to 9Tb.
TMP_LOC: specific to a computing node: temporary disk space specific to anode. Be cautious, this space is unseen by other nodes and to the front-end. It may be of interest to use this space for a single-node job. The I/O efficiency there will be a lot better than on the shared TMPDIR.

FTDIR: buffer zone to be used when transferring through ftserv. This space is monitored by the system (automatic cleaning).
WORKDIR: intermediate (without back up) working space where data is stored as for long as possible. It is quicker to save files on WORKDIR than to go and retrieve them on «cougar». The oldest files are automatically destroyed as soon as the file system has reached a certain level. Total volume amounts to 4Tb.

Each of these disk spaces can respectively be reached using the variables $HOME, $TMPDIR or $tmpdir, $TMP_LOC or $tmp_loc, $WORKDIR or $workdir.

Task submission:
It is done directly from tori (the front-end).
A batch task (or batch request) can be single-processor, single-node (up to 8 processors) or multi nodes.
The task scheduler being based on real time (or elapsed time), it is absolutely necessary to specify this limit when setting the "qsub" submission options. ***Please note that this is new, compared to the VPP procedure: CPU time is optional, while elapsed time is mandatory***.

For an optimal working of the scheduler, ***it is very important to describe as precisely as possible the task resources (number of nodes, number of processors per node, CPU time, elapsed time, node memory) and to make sure that options made through "qsub" are consistent with the mpirun command for jobs requiring MPI***.

The main standard NQS instructions are as follow**:**
job submission**:**

qsub [options] myjob
for submitting jobs (myjob = submission script)
ex:
qsub -q vector -b 2 -l cputim_job=1200, cpunum_job=4, elapstim_req=600, memsz_job=12gb -j o . /test .sh
submits script "myjob" to the "vector" class, on 2 nodes, 4 procs per node, 12Gb memory per node, 1200 sec of CPU time per processor and 10 min total elapsed time.

File transfer software to storage mainframe:
The « ftserv » software has been implemented on the front-end only.
The local commands (ftpassword, ftget and ftput) have been implemented in order to regulate and to secure transfers between "cougar" and "tori". The transfers use the "ftp" protocole, login password on « cougar » is stored in an encrypted form in the config file « .ftuas » in the $HOME of tori.

« ftget » and « ftput » commands must be used from this machine to transfer files between "cougar" and $HOME or $WORKDIR or $FTDIR. ***The use of FTDIR is recommended should you not wish to keep the files after computing has been done***. These temporarily stored files will be seen from the vector nodes. $FTDIR, contrary to $TMPDIR, will be kept after the retrieval step, therefore stored files will be accessible by the request running on the vector nodes.

During batch jobs, it is of the utmost importance that jobs/tasks should have the following structure (split into 3 NQS secondary jobs):

1. **Pre-processing step (jobs running on the front-end "tori" in the "ft" class)**:
   i. file retrieval under GFS on $FTDIR buffer space (ftget)
   ii. qsub -q vector job_calcul
2. **Computing step (job running on vector nodes)**:
   i. links of input files from $FTDIR to $TMPDIR
   ii. computation inside $TMPDIR
   iii. move output files (to be transferred to cougar) to $FTDIR
   iv. qsub -q ft post_processing
3. **Post-processing step (job running on the "tori" scalar front-end in the "ft" class)**:
   i. file archiving on cougar or any other local platform (ftput)

To achieve a more user-friendly split into 3 parts of the jobs (ie : splitting almost transparent for the users), Eric Sevault has developed a special tool box named ***MTOOL***. It is highly recommended to use this tool !
Eric Sevault has also developed a user-interface to "mpirun", named "***xmpirun***". To be used as well !

The FORTRAN Compiler:
FORTRAN 95 is the norm for the FORTRAN compiler (sxf90). It is a "cross compiler" which generates codes for vector nodes from the front-end "tori".

« sxf90 » enables to get to the cross compiler, either run « man sxf90 » or read the guide to know the different options or the full documentation « cookbook ».

Main differences with the VPP:

- *User's work (editing, compiling, linking, library creation…) should be done on the front-end (TX7). All the cross-commands are prefixed with "sx" : use sxf90, sxar, sxld, ... instead of f90, ar, ld ...*
- *For making mpi executables, use sxmpif90 instead of sxf90..*
- *Contrary to the VPP, a MPI parallel executable must be launched with the « mpirun » command.*
- *Equivalence of compilations options between the several systems (source E. Gondet /MERCATOR)*

Move your $HOME data from VPP to NEC:
NEC is to be switched off end of June 2007. The move of data must be the opportunity to filter and only select the really useful data (NO BINARY EXECUTABLES !!).

CNRM/GMAP/COOPE (E. Escalière) have asked for opening of all VPP-identified users on NEC/tori (March 29th). Remote 'external) access validation already has been provided to those who have renwed their "parme" (firewall) access since January 2007. Those who'd need to renew their access from beginning of July onwards will get their "tori" access validated on a case by case basis.
(refer to email by E. Escalière and C. Fischer)

Basic reference scripts will be provided to all Aladinists (R. El Ouaraini):

- E923

- E927 and forecast

Test of coupling files and clim files produced on NEC:

Basically 3 types of tests:

1. one set of coupling files for every transmission domain seen from MF
2. local tests of ee927 using the ("new") NEC-generated clim files
3. a blending E-suite run in Prague

Status of validations:

Domains	Formal reply	Test 1	Test 2	E-suite test	Done in Toulouse
Algeria	--	--	--	--	ok
Belgium	ok	ok	ok	none	--
LACE	ok	ok	ok	ongoing	--
Morocco	ok	ok	ok	--	--
Poland	ok	ok	ok	--	--
Portugal	ok	ok	ok	--	--
SELAM					
Bulgaria	ok	ok	ok	--	--
Romania	ok	ok	ok	--	--
Tunisia	ok			--	--

Test 1: a set of specific coupling files
Test 2: re-run local ee927 using the NEC-generated clim file

LACE: reference (and coordination) in Prague/CHMI; results on website:

http://www.chmi.cz/meteo/ov/aladin/climtest07/

Switch of the "DSI" development platform from the "andante" to "merou+largo" servers
E.E and C.F.

"andante" =>

- "largo" or "triolet" for operational database acquisition and extraction (BDM, SOPRANO, general Unix environment and user applications). In early 2008, "largo" will again be replaced by a Linux cluster system. All users and $HOME will be transferred straight from "andante" to "largo", in a user transparent way.
- "merou" for clearcase usage. Between mid-March and mid-May => GCO move their working environment ; clearcase application already installed on "merou" ; CNRM staff need to switch to "merou" for their CY32T2 contributions. Access only via "ssh" commands !

Impact on partners:

/utemp might not be transferred ! => ideally, clean it and save relevant data on cougar

remote (including firewall) access validation will automatically granted (for users having valid access to "andante") to "largo" or "triolet"

user accounts for "merou" will be open on demand, only for clearcase users => please send your request for merou+cc by email to Eric Escalière as soon as possible in May ! remote access will be granted only at the next general remote access validation of a given user (to all his/her acceding platforms at MF). Thus, first work with merou+cc will be by access via "largo".

Later, to come: … security partitioning within the MF computer domain, mostly between "development" and "operational" (NEC, "largo/triolet", "merou" ?) platforms.

GMKPACK
Ryad El Khatib

GMKPACK has been ported on the front-end "tori" server and supports the cross compilation as well. Usage is the same as on other machines, and the compilation and links are much faster than on the VPP5000.

Further details => see on the GMKPACK documentation on-line :

http://www.cnrm.meteo.fr/gmapdoc/