NARVAL a Modular Distributed Data Acquisition System with Ada 95 and RTAI

Xavier Grave, Rogelio Canedo, Jean-Francois Clavelin Sylvie Du, and Eric Legay

***Abstract*—NARVAL stands for *"Nouvelle Acquisition temps-Réel Version 1.2 Avec Linux"*. It is developed in a object-oriented language with Ada 95. It is also an acquisition system partitioned with the help of Ada Annex E. All the Unix processes running the acquisition system are seen as an unique Ada 95 program with multiple tasks running on different computers. The dataflow handling is done with TCP/IP socket connections between tasks running on different computers and with UNIX fifo for tasks running on the same computer.**

All processes are based on a class that inherits from an abstract class named Actor. There are three main categories of actors : - producers, which typically collect data from hardware - intermediaries, which act as NxM software routers - consumers, which are mainly used to store or histogram data.

All these actors are managed with a main task that concentrates configuration information and a state machine that gives the state of the system and the possibilities to change it.

Narval currently supports real time access to VME, VXI and PCI busses using custom Linux drivers based on RTAI. ATCA is under consideration with Gigabit Ethernet support (Infiniband or PCIXpress support are also envisaged). ATCA board will embed real time linux on Xilinx Virtex II pro.

NARVAL is used in different places and experiments. It is accepted for the European AGATA detector. About 7000 high resolution channels (14bits@100MHz ADC $\Rightarrow$ 14GB/s) will be processed (few stages of analysis) in order to store about 100MB/s of rebuild data.

I. NARVAL

A. NARVAL *history*

NARVAL was developed to replace OASIS [1], the previous IPNO's data acquisition system. NARVAL developement is in Ada 95, an object-oriented language, and is a multi-architectures distributed system (*PowerPC* and *i386*).

The NARVAL architecture was thought to facilitate the deployment for all kinds of acquisitions. For this reason, NARVAL is built from two types of components: the NARVAL core and the NARVAL actors. The core handles the dataflow and the application integrity, while the actors make specific tasks of the acquisition.

B. *Design Overview*

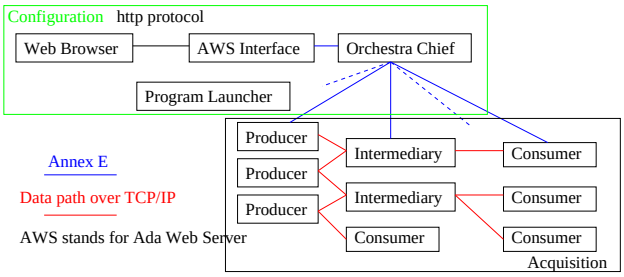
The acquisition software design can be described as follows:

1. A partition named *"Chef d'orchestre"*.
   Handles state machine, configuration.
2. A partition named *"gestion d'erreurs"*.
   Handles errors and recovering from errors
3. Partitions named actors
   Handle dataflow

II. NARVAL CORE AND ACTORS

The following figure shows an example of how an experiment can be implemented using NARVAL.

X. Grave is with the INSTITUT DE PHYSIQUE NUCLÉAIRE D'ORSAY, Université Paris XI, Orsay, 91405, France, Phone: (33)-1-69-15-79-59, email: grave@ipno.in2p3.fr



A. *The Core*

In order to coordinate and implement all NARVAL actors, the acquisition system is based on a few libraries that handle all the configuration phases, manage the DAQ's state machine and the data transport.

Libraries list :

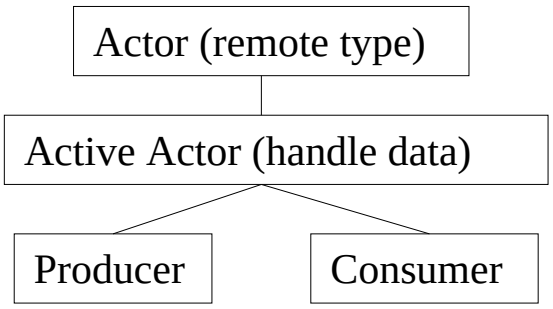
- communication classes
  data transport over TCP/IP or over UNIX fifo,
  MPI in developpement
- error handling (Ada Annex E library distributed)
  storage with a POSGRESQL database
- protected memory handling
  protected object (an Ada 95 specificity) containing three buffers of variable length
- actor class (Ada Annex E object distributed : remote type)
  an abstract class that every object that can change state inherit off

On this bases, three main programs are developped :

- orchestra chief
  elaborate and contain the configuration of the current experiment
  send orders to all experiment actors
- error handler
  collect error messages, store them and execute a recovery task
- program launcher
  running on every node where an actor has to run
  launched by the orchestra chief by a rsh/ssh command
  can kill defective actor if needded

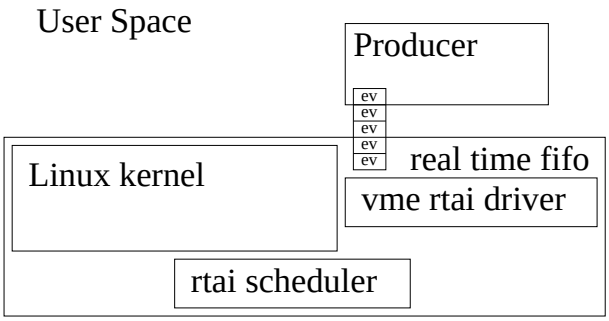
B. *The Actors*

Actor is an object class, three kind of actors inherit from it : producer, intermediary and consumer.



B.1 The producer

This familly of actor collects data from hardware device such as electronic boards embedded in VME crates. This class of process are not real time ones. Instead of this, when real time is needed, they take data from RTAI fifo. This mechanism can be summerized by the following figure :



B.2 The intermediary

An intermediary act as a software switch, it can have N input and M output. It is usually used as a compressor/decompressor, event builder or event dispatcher. Producer and consumer inherit from this class, for the first one N equal zero and M equal one and for the second one N equal one and M equal zero.

B.3 The consumer

The consumer is the endcap of the dataflow, in this way it can store data or display it with or without transformation. Typical use of a consumer is a DLT storage or an histogramer.

III. Some technical point

A. Hardware supported

Theoretically, NARVAL is compatible with all kinds of architectures, if a compliant Ada 95 compiler exists. NARVAL handles different architectures in the same acquisition without problems, it has been tested with *PowerPC* and *i386* (Opteron and Xeon32) architectures.

The real-time part of NARVAL uses *Real-Time Application Interface* [2] (RTAI) as real-time OS. Real-time drivers for VME, VXI and PCIbusses have been developped at IPNO and a driver for ATCA is under development.

NARVAL communications are of two types : Ada Annex E and more standard network protocols like TCP/IP for exemple. The advantages of the Ada Annex E will be discuted in sub-section B.
NARVAL supports Gigabit Ethernet communication and, in the near future, will support Infiniband. To communicate via Gigabit Ethernet link, we use TCP/IP protocol, but in order to improve performances we are implementing MPI protocol. In any case, this protocol will allow us to use Infiniband in the best way.

B. Ada 95 and NARVAL

Ada 95 is a language used by companies in need of huge reliability like *BOEING*, *Nuclear Power Plant* or *Paris Subway*. The features involved in Ada 95 are the following:

- Strong typing.
- International Standard (ISO 8652 :1995) [3]. This standard is available freely, and porting an application to an other exploitation systems needs few more modifications.
- Compiler gcc, multi OS and multi-architectures.
- Thread/task model for expressing concurrent activities, mutual exclusion for shared resources, inter-thread or task coordination, and responses to asynchronous events including hardware interrupts.
- Facilities for assigning priorities to threads/tasks and for establishing appropriate scheduling behavior (for example, controlling priority inversions).

For the NARVAL development, we take advantage from Ada Annex E (CORBA equivalent). This Ada 95 annex simplifies the communication between processes. It allows to export to others authentified processes objects or any structures allowed by Ada 95.

IV. NARVAL examples

A. Experiments under NARVAL

Several experiments use NARVAL as DAQ system :

- LAG64 - TANDEM (Orsay, FRANCE), Heidelberg (GERMAN)
  VME crate with LINUX + RTAI embedded on a *PowerPC* + TDC 64 channels
- RECIF - TANDEM (Orsay, FRANCE)
  i386 computer + CPCI crate, no real time support
- AZ4PI - GANIL (Caen, FRANCE)
  same as RECIF with *i386* computer in a VME crate
  NARVAL slave of DAS [1] acquisition
- CHACO - TANDEM (Orsay, FRANCE), Los Alamos (USA), GANIL (Caen, FRANCE)
  VME crate with LINUX + RTAI embedded on a *PowerPC* + TDC, ADC, scalers
- PARRNE - TANDEM (Orsay, FRANCE), CERN ISOLDE facility
  VXI crate with LINUX + RTAI embedded on a *PowerPC* + custom VXI boards (DSP embedded on each board, including slot zero board)
- MUST2[2] (Orsay, FRANCE)
  VME crate with LINUX on a *i386* computer + MXI2 link

[1] DAS is the acquisition system of GANIL
[2] MUST2 is an electronic test system

B. NARVAL *and* AGATA

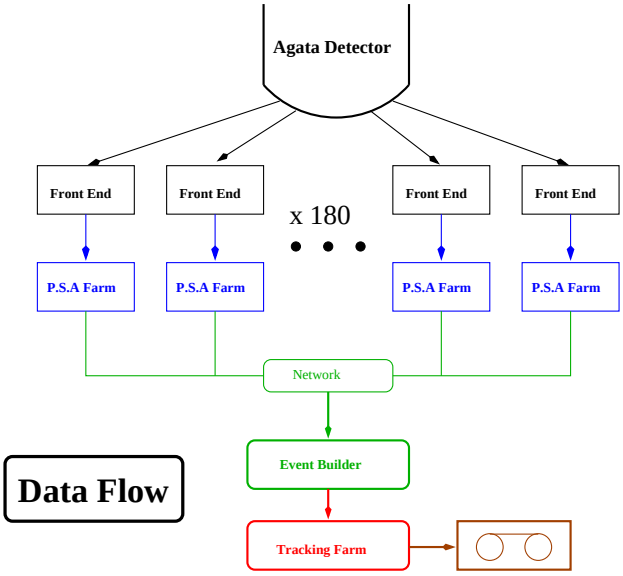
AGATA is a new generation experience. The maximum output of the electronics is evaluated to 67GB/s@50kHz. It's a real challenge to design the NARVAL architectures to this throughput.

B.1 AGATA presentation

AGATA [4] (ADVANCED GAMMA TRACKING ARRAY) is a new generation of 4π photon spectrometers for nuclear spectroscopy studies. The novelty resides in the fact that unlike in the current generation of spectrometers, such as Euroball [5], the Germanium crystals will not be surrounded by Compton suppression shields. The removal of the BGO shields yields a considerable gain (close to a factor 2) in the solid angle which can be covered by Germanium thus increasing the overall detection efficiency of the spectrometer. In order to preserve a good peak-to-total (P/T) and recover the full energy of photons which scatter from one germanium crystal into another, the trajectories of the photons as they interact throughout the spectrometer need to be reconstructed.
The performances of AGATA are strongly related to the capacity to accurately determine the positions (within a few mm) and energies of the individual photon interactions in the Germanium crystals and to the reconstruction efficiency of the photon trajectories.
The first item relies on pulse shape analysis. This is the stage which requires the highest computing power in AGATA. The predicted bandwith for the input of a PSA farm is in order of 110 MB/s@15kHz by crystal. We estimate that the needed computing power represents 90 blades with 14 biprocessors each [3].
The second point relies on tracking photon trajectories [6]. This stage consumes less computing power, but needs higher bandwith.
The sensitivity of AGATA will be superior to any existing array by several orders of magnitude. It will be for a long time a rich source for nuclear structure physics providing the means for new discoveries and opening challenging new perspectives.

[3] Under discussion, actually we are missing a reallistic PSA algorithm

B.2 Dataflow detail description in AGATA

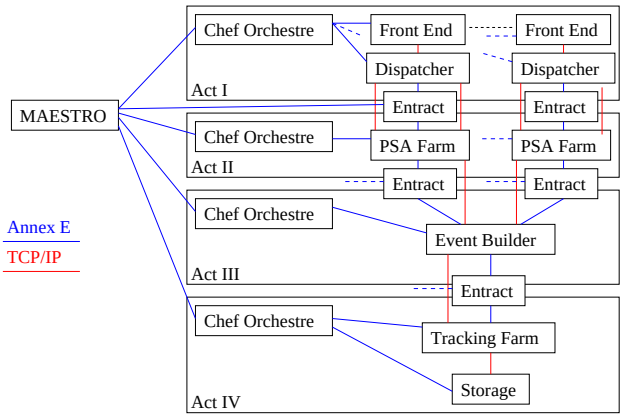


- *Front-End Electronic* : The DAQ Dataflow comes from the Front End Electronic (FEE) which treats each crystal (core+36 segments) as a separate entity. The core signal is formed by the superposition of the charge released by all the interactions in the 36 segments of each detector. This signal can be used as a local trigger for the whole crystal. Each crystal has a 110 MB/s throughput.
  In the FEE, NARVAL should have three kinds of actors:
  1. producer to concentrate the data from a crystal.
  2. intermediary to perform zero suppression.
  3. intermediary to dispatch the data to different PSA treatments.

  The communication between the FEE and the Pulse Shape Analysis computer are handled with Gigabit Ethernet Interfaces.
- *Pulse Shape Analysis* : The PSA farms should be constitued by 90 blades center with 14 biprocessors each. One blade center should be able to handle the data incomping from two crystals. The dataflow from one crystal will be reduced by a factor 75 to reach a throughput of 1.5 MB/s.
  The computing power needed by the PSA strongly depend on the algorithms performances. To date several algorithms investigations are under way (neural networks, matrix inversion, genetic algorithms and wavelets for pattern recognition).
  In the PSA, each blade center will run the same programs in parallel. NARVAL should have one actor by implemented algorithms and by blade.
- *Event Builder* : The Event Builder (EB) receives the events group by detector and send the events group by time slot. So, the EB must accept a throughput of 270MB/s in input and an other 270MB/s in output.
  Actually, we are studying two form of EB : a farm or a single multiprocessor host. In the case of a farm, one can use Ethernet to send the data from the PSA to the EB and towards the tracking farm in the next step. But we'll need a high performance communication system to allow the differents units to collaborate and put together the events

fragments.
In the EB, NARVAL should have one kind of actor : an intermediary which collects the FEE's data, regroup the data by events in buffers and send them to the Tracking Farm. We are studying different architectures (mono or multi nodes), different kinds of network (Ethernet, Infiniband) and different protocols (TCP/IP, MPI, TCPOE[4]).

- *Tracking* : The tracking farm could use the same type of calculation nodes as the PSA farm. Each node will be able to process any event buffer from the EB since each event in a buffer is present in its totality.
  If the EB is a single node, it can send the event buffers through Infiniband to a dedicated node which will then dispatch them by Ethernet.
- *Data Storage* : The Agata data storage architecture can be based on a two-stage hierarchy. The primary storage proposes a large enough disk space to store all the acquired data from two successive experiments (about 2x60 Terabytes). It is made of an array of hard disk drives capable of sustaining the 140 Mbytes/s data throughput from the acquisition. RAID arrays , based on Fibre Channel and Serial ATA are available today. This will provide adequate performances and security for data with low cost. This primary stage involves a NARVAL development. We will develop an interface between NARVAL and the chosen storage system.
  The secondary storage is realized in an offline manner, independent of the data taking and therefore independent of NARVAL.



Since AGATA is an european collaboration, NARVAL has to take into account communication with non Ada 95 programs. So every actor and partition of the system will embedded a web server and publish web services. All of this will be implemented with the AWS[5] library, which able to generate a WSDL file from an Ada 95 specification file. All the skeleton and stubs for client and server are also automaticly generated.

REFERENCES

[1] Borome N., Bossu Y., Douet R., Harroch H., and Tran-Khanh T, "A vme multiprocessor data acquisition system combining a unixworkstation and real-time microprocessors," *IEEE Trans. Nuclear Science*, vol. 37, no. 4, pp. 1514–1519, 1990.
[2] P. S. Hughes and D. Beal, "Rtai: Real-time application interface," *LINUX Journal*, 2000.
[3] "Ada reference manual," *ISO/IEC 8652:1995(E)*.
[4] W. Korten J. Gerl, "Agata, technical proposal," *GSI, Darmstadt*, 2001.
[5] J. Simpson, "Z. phys. a 358 (1997) 139," .
[6] A. Lopez-Martens et al., "γ-ray tracking algorithms: a comparison," *Nucl. Inst. and Meth A 553*, pp. 454–466, 2004.

B.3 NARVAL improvements

In order to adapt NARVAL to the AGATAexperiment where a lot of actors will be needed, the developpement team has started several improvements. To ease the managment of a big number of actors NARVAL will be cut in slides named act. Each slides is an actual NARVAL system with its configuration library and state machine (embedded in the orchestra chief). Only the error handler will be generalized for the whole system. A partition called MAESTRO will handle a **"meta"**configuration and a master state machine, it will coordinate all the NARVAL subsystem. Interconnection between subsystem will be done using Ada Annex E in a partition called entract. In it will be stored by a first subsystem the location (hostname and port) where other subsystem can take data. This will allow dynamic connection to data provider. The following figure shows a possible scheme for the AGATA experiment.

[4] TCP OFF LOAD ENGINE

[5] Ada Web Server from Pascal Obry