Handling of large datasets in Quantum Chemistry

CECAM Lyon, 19/04-21/04 2006

Antonio Monari[a], Stefano Evangelisti[b]

[a] Dip. Chimica Fisica ed Inorganica-Univ. Bologna, [b] Lab. Physique Quantique-Univ. P. Sabatier Toulouse.

Plan of the talk

- Introduction
- Q5cost data format
- Q5Cost library
- Future Developments and Conclusions

Introduction

The context

- Activity carried out within Cost in Chemistry D23
- Involved parties:
  - CINECA, Italy
  - Univ. Bologna, Italy
  - Univ. P. Sabatier Toulouse, France
  - Univ. Ferrara, Italy
  - Univ. Budapest, Hungary
  - Univ. Valencia, Spain
- Codes produced by the involved parties are complementary and often need to be interfaced
- Final goal: To build a grid based distributed laboratory
- Facilitate communication between different QC codes
- First problem to face: Each code works with its own data format

A Common Format for QC codes

- Our decision:
  - To build a Common Format for QC problems
  - To write a converter wrapper for each code in the set

- Common Format should be:
  - as general and complete as possible
  - flexible enough to be interfaced with codes under constant development
  - platform indipendent
  - easy to use for chemical users

QC Data

- We can identify two types of QC data:
  - Small data quantities (mainly ASCII coded)
    - Geometry, Symmetry, Atomic basis set, etc...
    - We devised a XML based format QCML
  - Large datasets (mainly binary)
    - AO or MO integrals, MO coefficients, Wavefunction
    - We devised Q5cost an HDF5 based data format

Q5cost data format

What is HDF5?

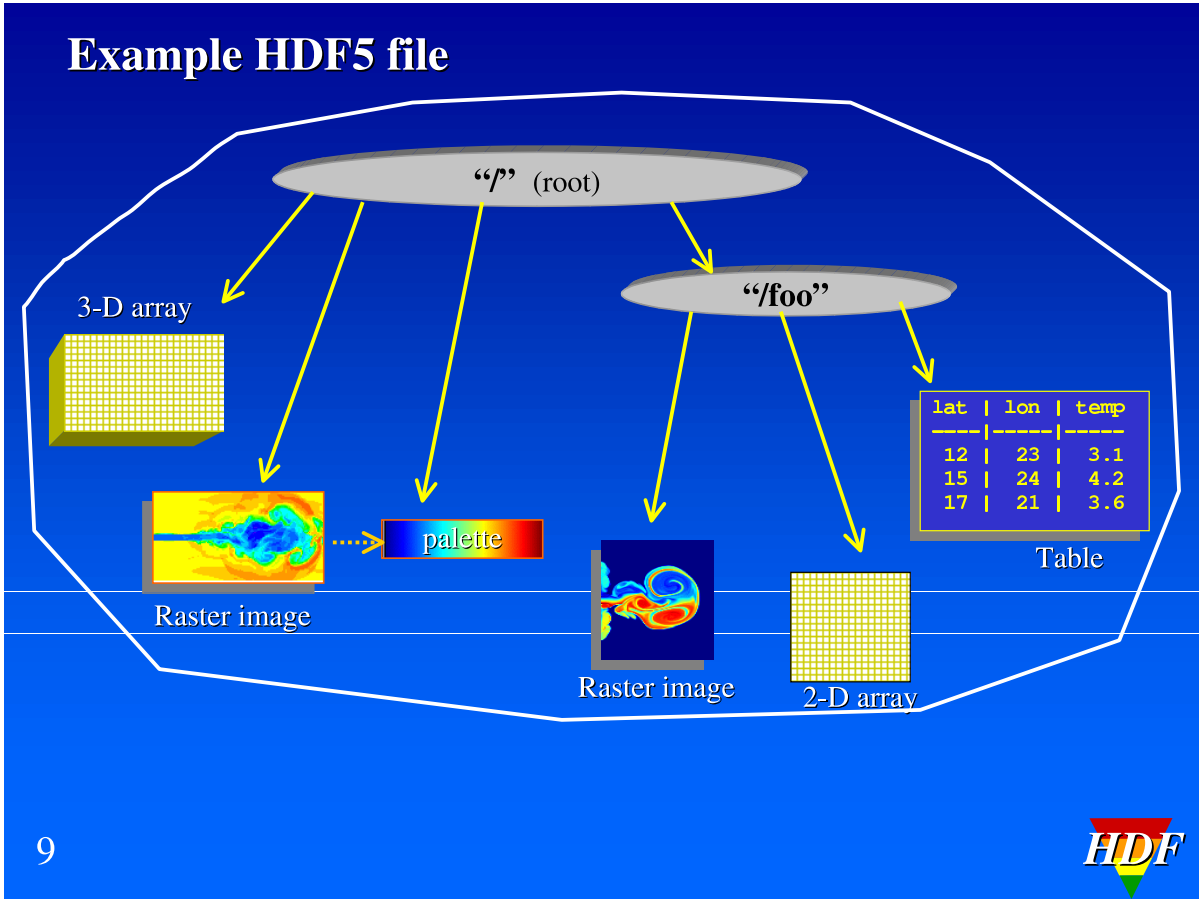
- HDF5 [a] Format and software for scientific data produced by NCSA/University of Illinois
- Support any kind of data for digital storage regardless of their origin and size
- Stores data in a highly organized and hierarchical format
- High efficient chunked I/O
- Allows inclusion of metadata (attributes)
- Platform indipendent file format
- Widely used in scientific or visualization codes

[a] HDF5 a general purpose library and file format for storing scientific data. http://hdf.ncsa.uiuc.edu/HDF5/

HDF5 Data Model

- Datasets
  - Multidimensional arrays of elements together with suporting metadata (attributes)
- Groups
  - Directory like structures containing, datasets, attributes, other groups

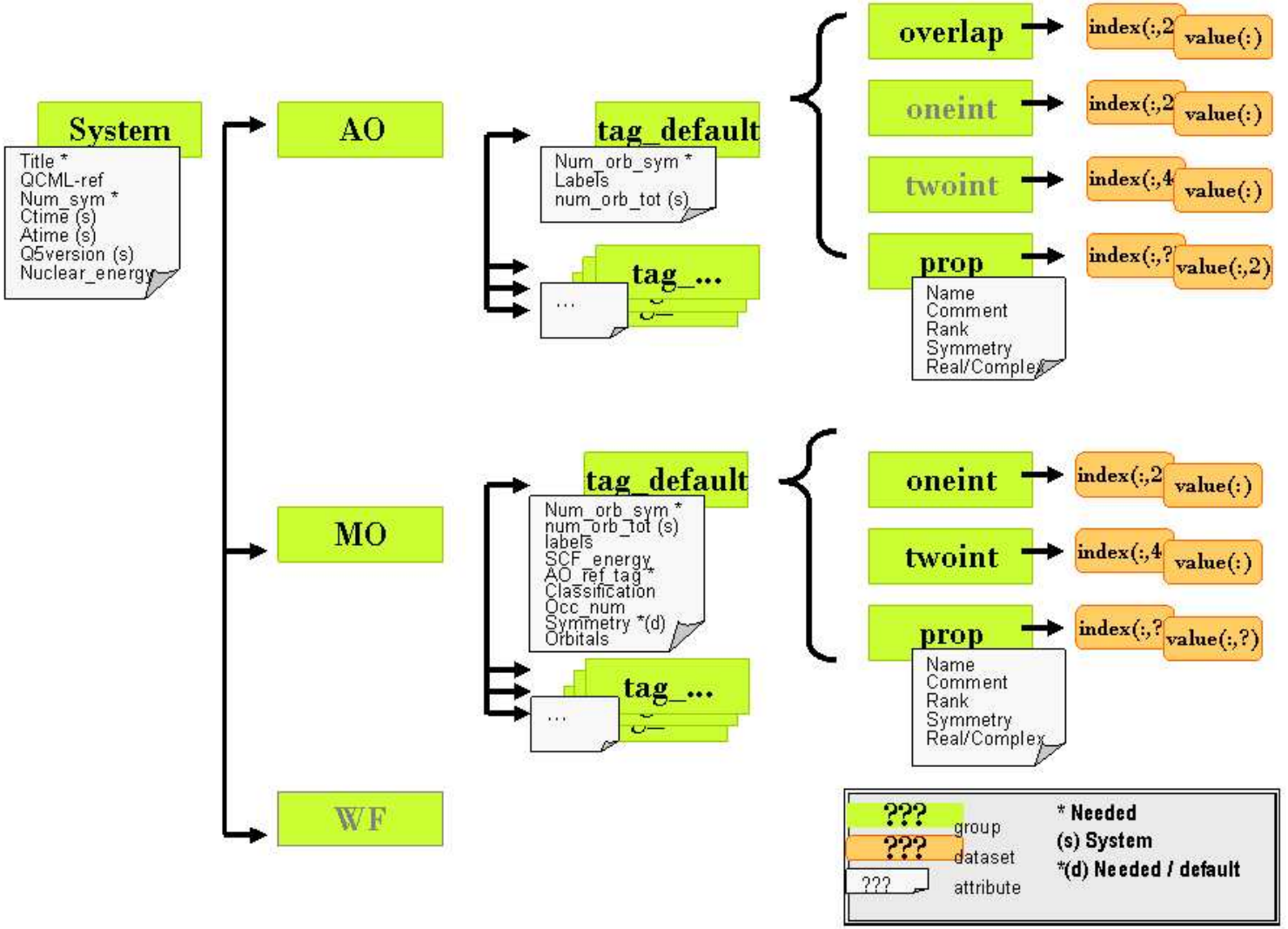
HDF5 Hierarchy



Q5cost file

- Q5cost file stores:
  - large sparse matrices with arbitary number of indeces (AO and MO integrals related to a generic One or Two particles operator). They can be defined as Generic Properties
  - small data (scalar and arrays), called metadata (nuclear energy, orbitals label, MO coefficients, etc..)
- File has a hierachical structure
  - A first root container (**System**) represents the molecular system
  - A System can contain several Domains, grouping together Properties whose indeces conceptually refers to the same kind of functions
    - **AO Domain**
    - **MO Domain**
    - **WF Domain**

Q5 Cost file hierarchy



Q5cost file Conclusion

- Q5cost file stores Atomic and/or Molecular Integrals (only non zero values with indices) and Coefficents.
- Q5cost file contains all binary information one needs to perform a QC computation.
- If some information is missing or still not produced it can be added the file subsequentelly.
- We can define a proper hierarchy, and memorize in a simple accessible way metadata.
- We can logically and physically separate AO and MO even memorizing them all in the same file.
- Different AO or MO can be stored on the same file. They are separate by the use of the identifier tag_ as different occurencies of a given domain
- Due to HDF5 features Q5cost files are platform indipendent.

Q5Cost library

Q5Cost library

- Basing on HDF5 API we wrote a FORTRAN95 high level library
- Provides read and write acces to Q5cost files
- API is based on well known Chemical entities, rather than HDF5 objects
- Provides a high level access for quantum chemist developers

Q5Cost, where can I found it?

- Present version 9.0.3
- The library is free and licensed as LGPL
- It can be downloaded from the net: http://abigrid.cineca.it
- It has been tested on various Unix/Linux architecture, and with different Fortran compilers

Library structure

- The libary is consist of several modules. The most important ones:
  - **Q5Cost**: Defines the high level API to be used by the final programmer
  - **Q5Core**: provides a wrapping facilities for HDF5 routines
  - **Q5Error**: provides error managment. Useful for debugging of library or application codes

Q5Cost module

- Provides high level acces to Q5cost files. Routines are organized in several classes:
  - **Init**: initialise and uninitialise the library
  - **File**: creates, opens, closes the Q5cost file
  - **System**: manages the System object
  - **AO**: manages object refering to the domain of Atomic Orbitals
    - **AOOverlap**
  - **MO**: manages object refering to the domain of Molecular Orbitals
    - **MOOneInt**
    - **MOTwoInt**
  - **Property**:manages a generic property in a given domain. User has to specify domain, rank and name
  - **WaveFunction**: not implemented yet

See what you have: Q5DUMP

Once you have your Q5cost file you may want to know what you have inside:

We designed Q5DUMP

- Allows you to see most significant Metadata
- Allows you to see if AO and MO are present
- Allows you to see which Integrals are present
- Allows you to see if MO Coefficients are present
- Allows you to see which Properties are present

Performance Test 1

- First test: writing time versus Buffer size for Q5cost and binary file

Buffer size	Time Binary (s.)	Time Q5cost (s.)
1024	265.23	226.62
2048	121.13	114.53
4096	62.38	59.02
8192	34.39	31.46
16384	18.86	17.04
32768	8.56	6.09
131072	6.19	4.86
262144	5.84	4.08

Number of integrals: 15000064, binary file size: 343 Mb, Q5cost file size: 346 Mb

Performance Test 2

- Disk occupation and writing time versus number of integrals for Q5cost and binary file. (Fixed chunk 16384 integrals)

Integrals	Q5Cost size	Wrt Q5cost (s)	Binary size	Wrt binary (s)
16384	397 Kb	$5.00 \cdot 10^{-2}$	384 Kb	$5.00 \cdot 10^{-}2$
65536	1.5 Mb	$1.00 \cdot 10^{-1}$	1.5 Mb	$1.00 \cdot 10^{-1}$
114688	2.7 Mb	0.15	2.6 Mb	0.17
507904	12 Mb	0.62	12 Mb	0.68
1015808	23 Mb	1.21	23 Mb	1.37
5013504	115 Mb	5.88	115 Mb	6.41
10010624	231 Mb	11.11	229 Mb	12.12
50003968	1.1 Gb	56.19	1.1 Gb	64.21
100007936	2.3 Gb	125.32	2.2 Gb	148.53

Future Developments and Conclusions

Future Developments

- Add direct support to AO One and Two electrons Integrals
- Introduce standard order or similar alghoritms to avoid storing indeces for two electrons objects
- Find an efficient way to store Wavefunction

Conclusions

- An efficient binary, platform indipendent file format for QC large data has been presented
- An easy to use Fortran library has been written to acces the file format
- Preliminary performance tests show library efficiency regarding disk occupation and writing/reading time
- First applications have been written, and first actual computations have been already performed:
  - Interface from MolCas files to Q5cost file
  - Interface from Dalton files to Q5cost file
  - Interface from Q5cost file to MolCost files (Toulouse format)
  - Bologna FCI code reads data directly from Q5cost file