Strategic Issues for Binary/File Format

ILDG4 May 21 2004, T.Yoshie CCS,Tsukuna

Definition

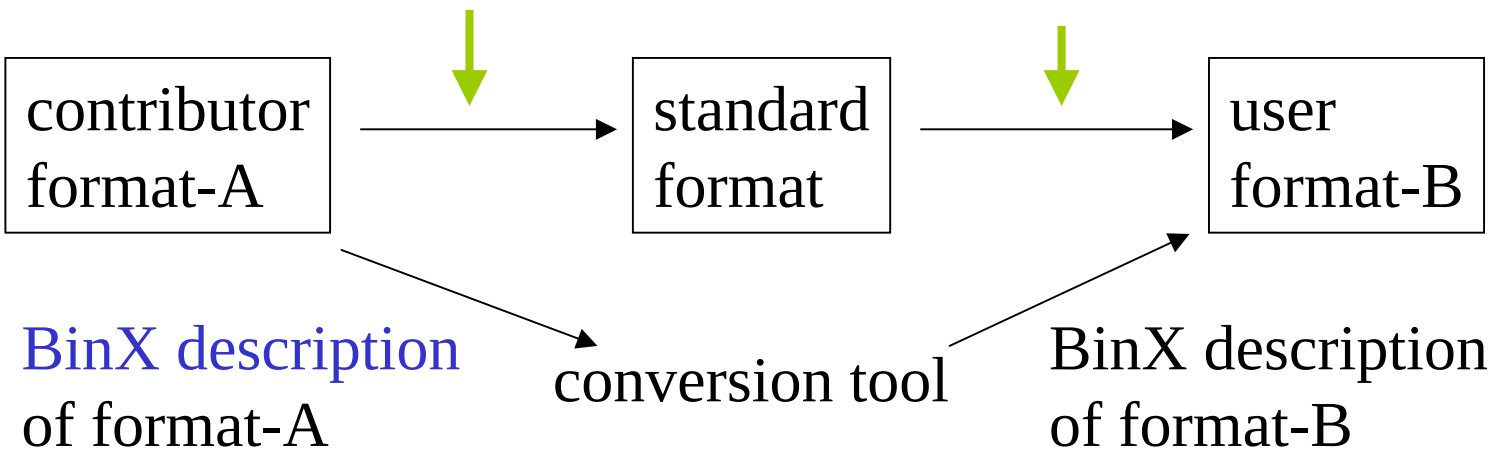
- binary format:
  - array format of config. U(3,3,X,,,)
  - numerical format (precision,IEEE,byte-order)
- file format:
  - format of a file transferred via ILDG
  - the file includes config. and (probably) some information of config.

MDWG Proposals (in QCDML draft 4.0)

- abstract (reference) standard binary format

C-library (provided by contributor) to read config in standard format

convert config. to user's format if necessary



enables us to save

- disk space on contributors side
- cost of conversion

- encapsulation (packing) of one binary config. and corresponding XML IDs in one file
  - this enables us to keep identity of files distributed via ILDG
  - no other header information is added

Issue-1

- Packing has to be done when config. is transferred from Grid to user disk
  - if we want to avoid "doubled disk space"

Possible Solutions and Comments

1. Give up packing
   - can avoid doubled disk space
   - config. identity is lost (user has to remember/record information)
2. Packing in advance
   - doubled disk space
   - config. identity is kept
   - XML IDs are duplicated (those in a packed file and those in metadata catalog), hence they are subject to mismatch

2’ Config.ID header

Add only a short header for unique config.ID

e.g. markovChainLFN+series+update

- (probably) doubled disk space
- config. identity is kept
- no duplicated XML documents

3. Packing when transferring

- no doubled disk space
- config. identity is kept, no duplicated XML IDs
- We have to develop an intelligent transfer software (SRM does not have this function)

note for discussions

- if we take “packing in advance” or “config.ID header” strategy, we may change strategy of “abstract binary format”
- if we take “packing when transferring” strategy, the transfer software has to be so intelligent that it handles multiple files for space-time decomposed configuration.
- in future, we may have “Replica” of configurations among Grids. Our decision has to be compatible with “replica”.
- If SRM supports directory,
  - GDN: top directory of a configuration
  - place a small file including config.ID at the top directory

Issue-2

- contributors binary format is a matter independent of ensemble/configuration
  - machine (#nodes) and code are placed in config.XML => binary format (in general) depends on config
  - a common binary format is used for many ensembles
- We have to develop a mechanism to manage C-libraries and BinX XML IDs
  - if we keep strategy of "abstract binary format"
  - or convert config. to standard binary format in advance

Conclusions

- Decide strategy and assign works to WGs

My personal opinion

- “config ID header “ strategy
  - add in addition, “Lattice Size” header
- convert configurations in advance from contributors binary format to ILDG standard binary format
  - Give up a strategy of “abstract binary format”

Reasons and Drawback

- simple, almost no management issues arise
- doubled disk space
- users have to download XML IDs in advance, before they use config for measurement.