Large Scale Nuclear Physics Calculations in a Workflow Environment and Data Provenance Capturing

Fang Liu and Masha Sosonkina
Scalable Computing Lab, USDOE Ames Laboratory

Outline

- Motivation
- Background:
  - Scientific workflow system
  - Provenance capturing for scientific applications
- Previous work on data management system for MFDn
- LCCI upstream codes and MFDn
- Workflow system in LCCI
- Potential provenance integration for MFDn
- Conclusion

Motivation

- We want to increase nuclear physics scientific discovery in several ways :
  - Sustainability : to ensure the codes, runs, results obtainable by the future users.
  - Provenance : to be able to get the information of the existing large run in a form which the new method can be compared.
  - Easy of entry : to lower the learning curve for getting young people involved , to avoid the inappropriate usage on super computers.
- A trustful provenance system is needed to record a multi-step data generation process, which allows to reuse the data product and data setting.

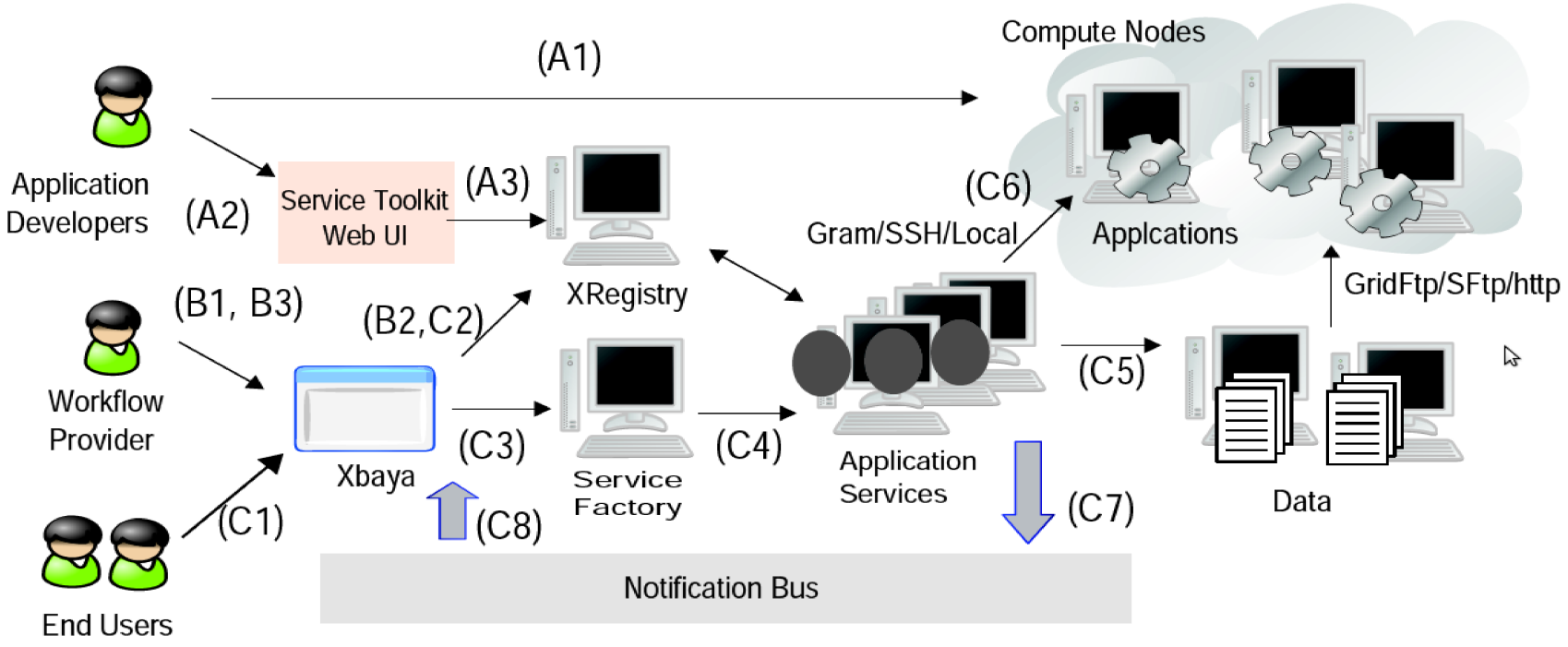
What is a Scientific Workflow?

- A workflow consists of a sequence of connected steps.
- A Scientific Workflow system is designed specifically to compose and execute a series of computational or data manipulation steps, or a workflow, in a scientific application.
- Domain specified workflow systems are:
  - Linked Environments for Atmospheric Discovery (LEAD) Project (http://portal.leadproject.org/gridsphere/gridsphere)
  - GridChem Computational Chemistry Grid (https://www.gridchem.org)

Major components in a workflow system

- Compute Hosts – where the application will be executed
- Application Services – Web service wrappers to command line science applications
- Service Hosts – where application services, service factory and registry service are running
- Scientific Workflows – tasks graphs of a combination of application services
- Application Provider – who constructs the application graphs
- End User – who is interested on the outcome of the workflow

General Workflow System Architecture



Tools – OGCE Workflow Suite

- A domain independent workflow suite facilitates users to securely share their applications as web services and construct workflows with these services.
- Suite includes a toolkit to
  - Wrap tasks as web service
  - A service registry
  - A workflow composition – xBaya
  - Enactment and monitoring GUI
- These tools are supported through Open Grid Computing Environments (OGCE) project (http://www.collabogce.org/ogce/index.php/Main_Page).

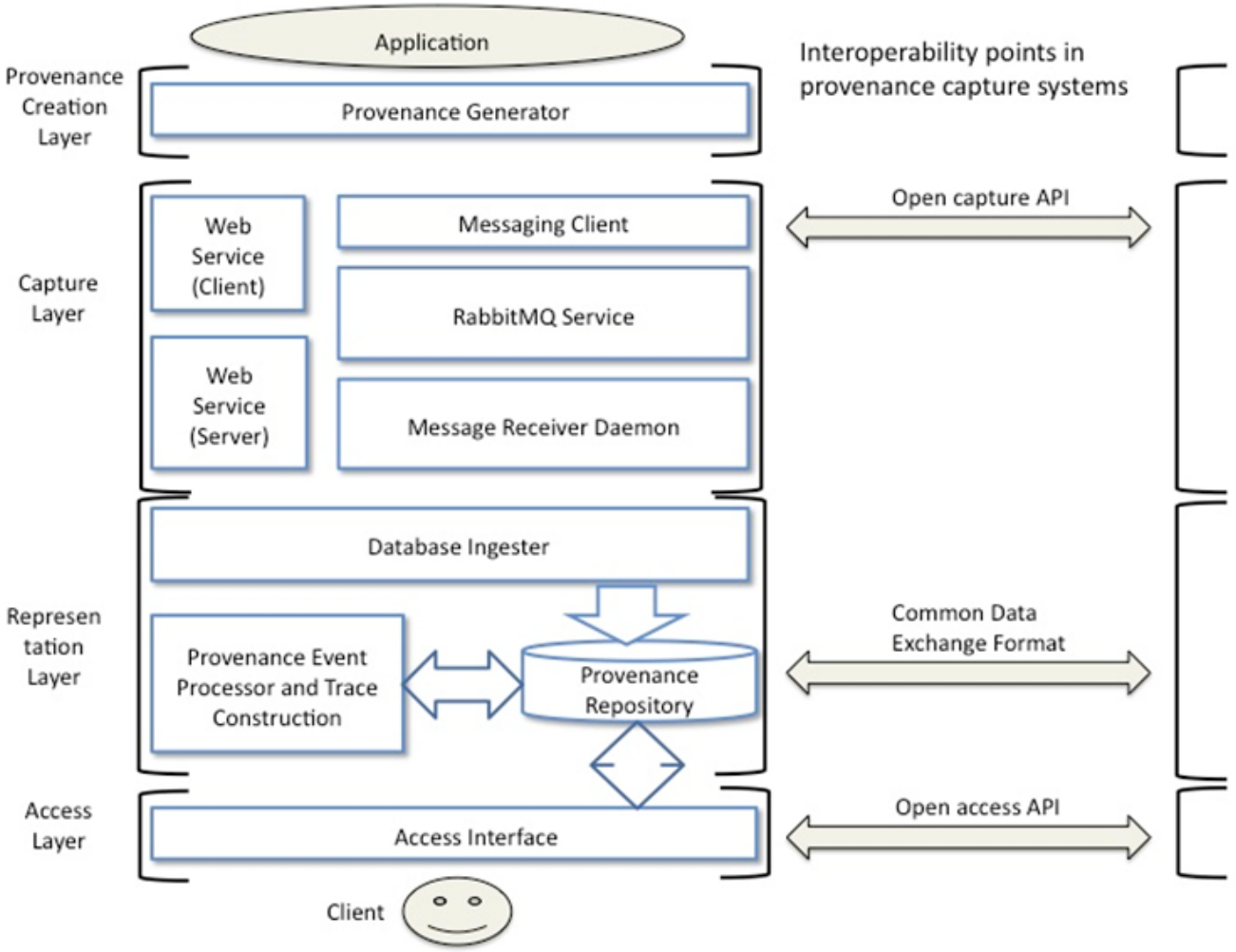
What is the Provenance for a scientific application?

- The provenance of a scientific application is when a collection was created, by whom, where and more experimental related information.
- Provenance
  - can identify event causality;
  - enable broader forms of sharing, reuse, and long-term preservation of scientific data;
  - can be used to attribute ownership;
  - determine the quality of a particular data set ;
  - insure the trust of the data set.

Provenance Tools

- Karma is a provenance capturing system that originally worked in cooperation with the GPEL workflow system in LEAD gateway.
- Karma is developed at Indiana University from the mid-2000's, and it evolves to be a standalone system, independent of any workflow system and free of the workflow systems have.
- Karma supports three ways to collect the provenance:
  - User annotation
  - Scavenging
  - Full Provenance Instrumentation

Logical architecture of a Provenance system



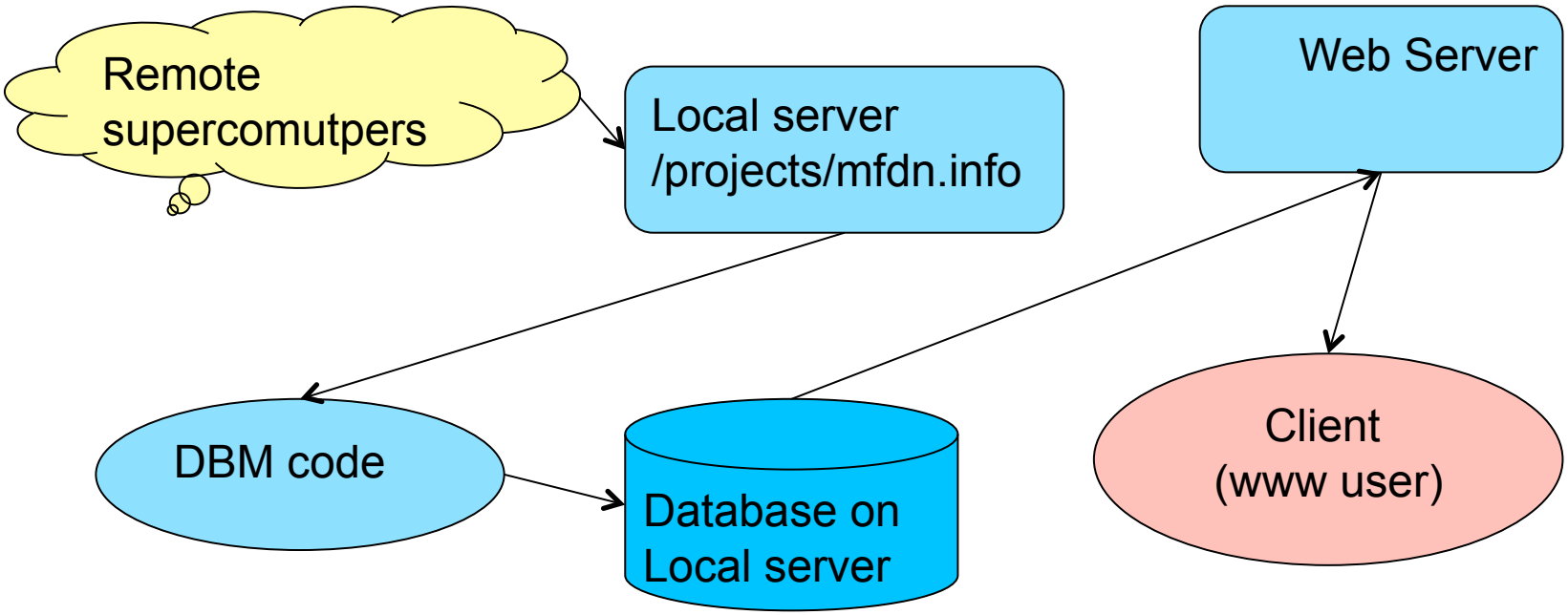
Projects using Karma

- Instant Karma: applying a proven provenance tool to NASA's AMSR-E (Advanced Microwave Scanning Radiometer – Earth Observing System) data production stream. Collect and disseminate provenance of AMSR-E standard data products, initially focusing on Sea Ice. (http://instantkarmatest.itsc.uah.edu/karmabrowser/view).
- netKarma: exploring networks of the future (http://groups.geni.net/geni/wiki/netKarma))

A Data Management System for MFDn

- An efficient tool for retrieving output from Ab-Initio CI calculations, the first step towards the provenance capturing
  - Record not only results, but also now, when and where those results are obtained.
- Record meta-data of every run in database
  - Data: results from LCCI ab-initio codes, typically stored on platforms where runs are performed
  - Meta-data: key information about each run, consolidated and formatted in the .info file

System Design and Components

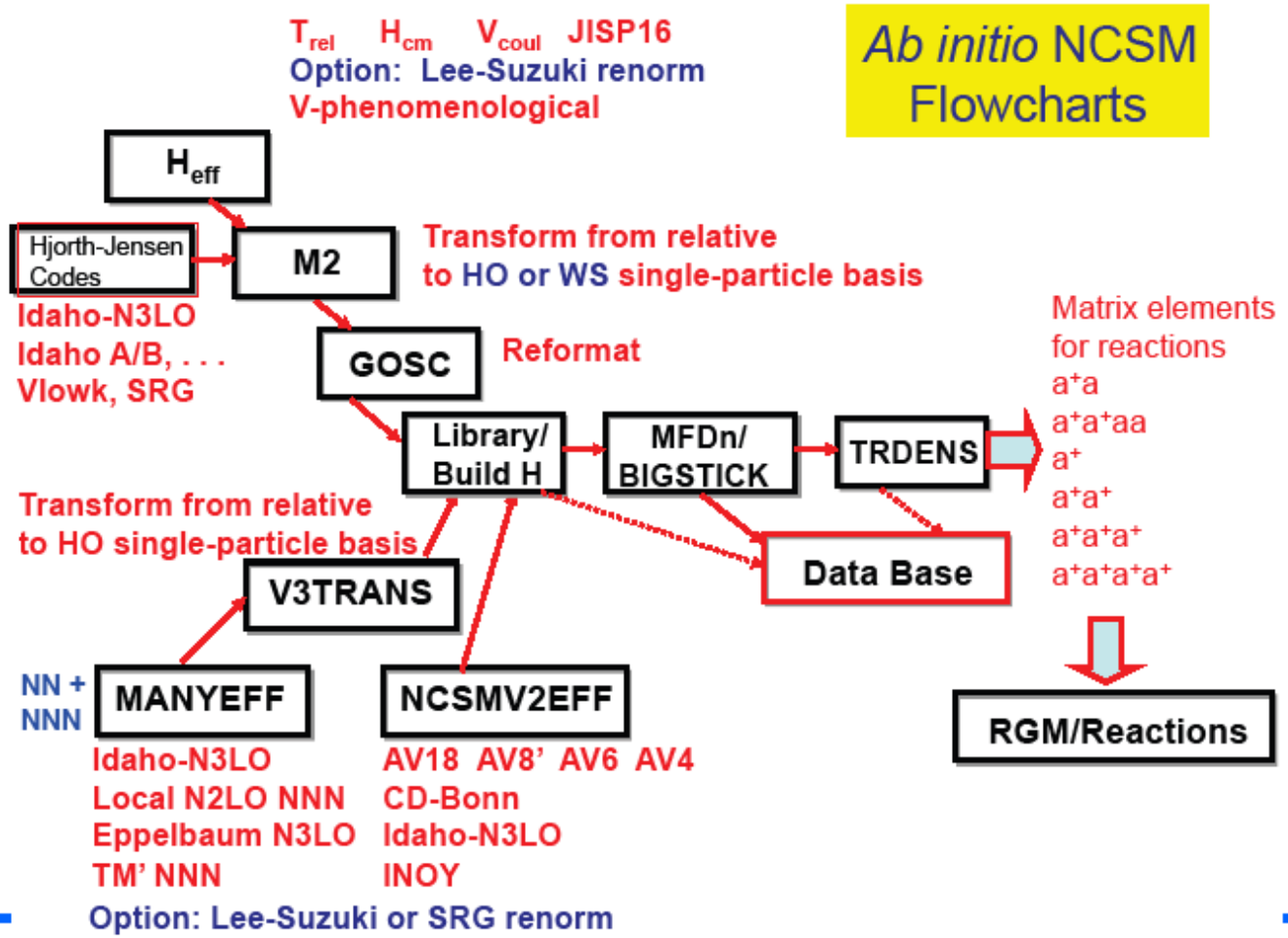


1. DB Manageer: parses the mfdn.info file and inserts the run record to Database
2. Web based front end: searches and lists existing run
3. DB Server: stores all the related metadata for each of MFDn run

Key Contributions

- It addresses provenance issue for computational nuclear physics community.
- It records the experimental information for a reproducible research in which the computational experiments can be run by others.
- The system is online since 08/2010 at http://nuclear.physics.iastate.edu/info
- A paper is published at HPC 2011, Boston, MA

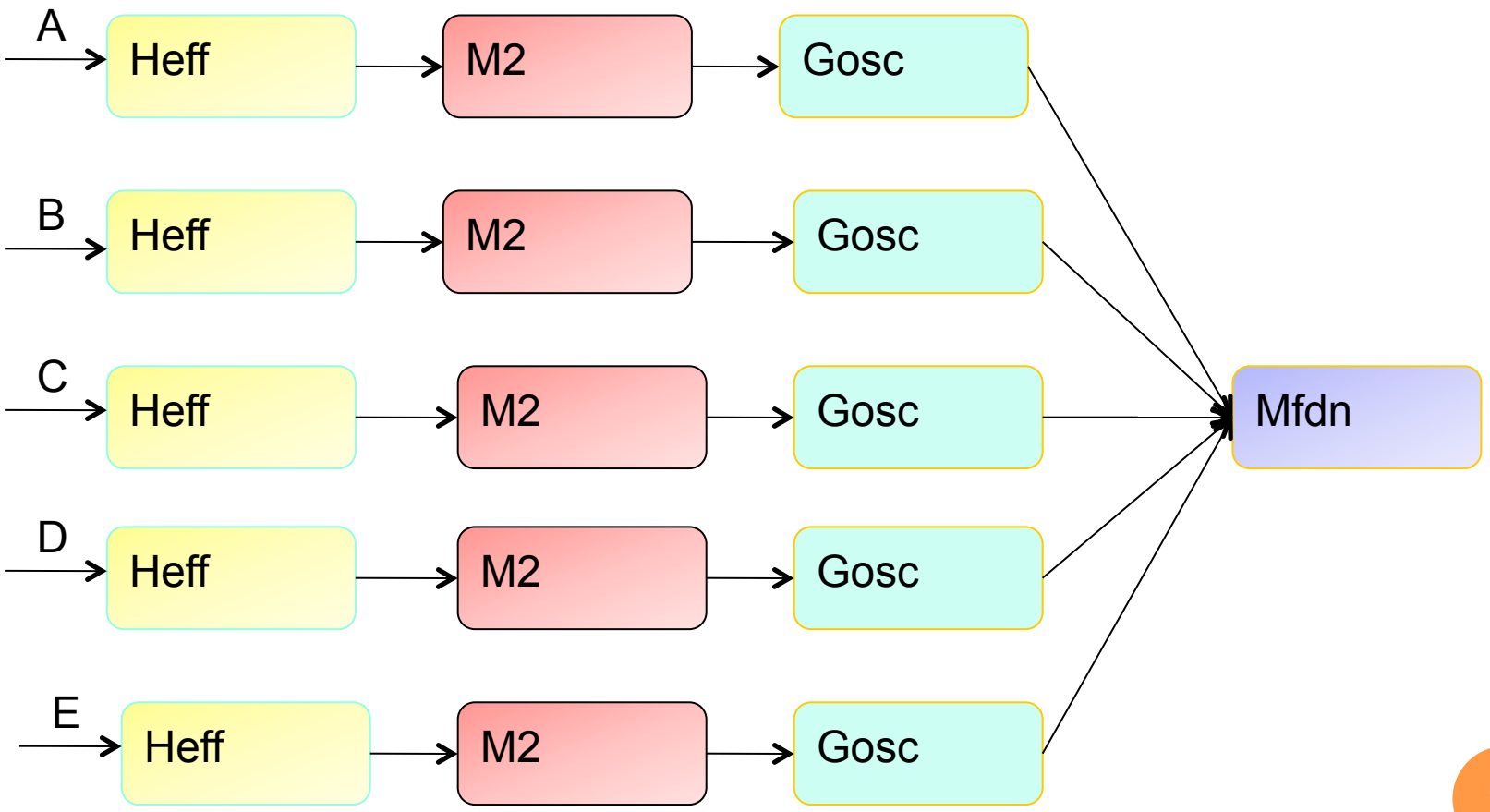
Overview of pre- and post-processing codes



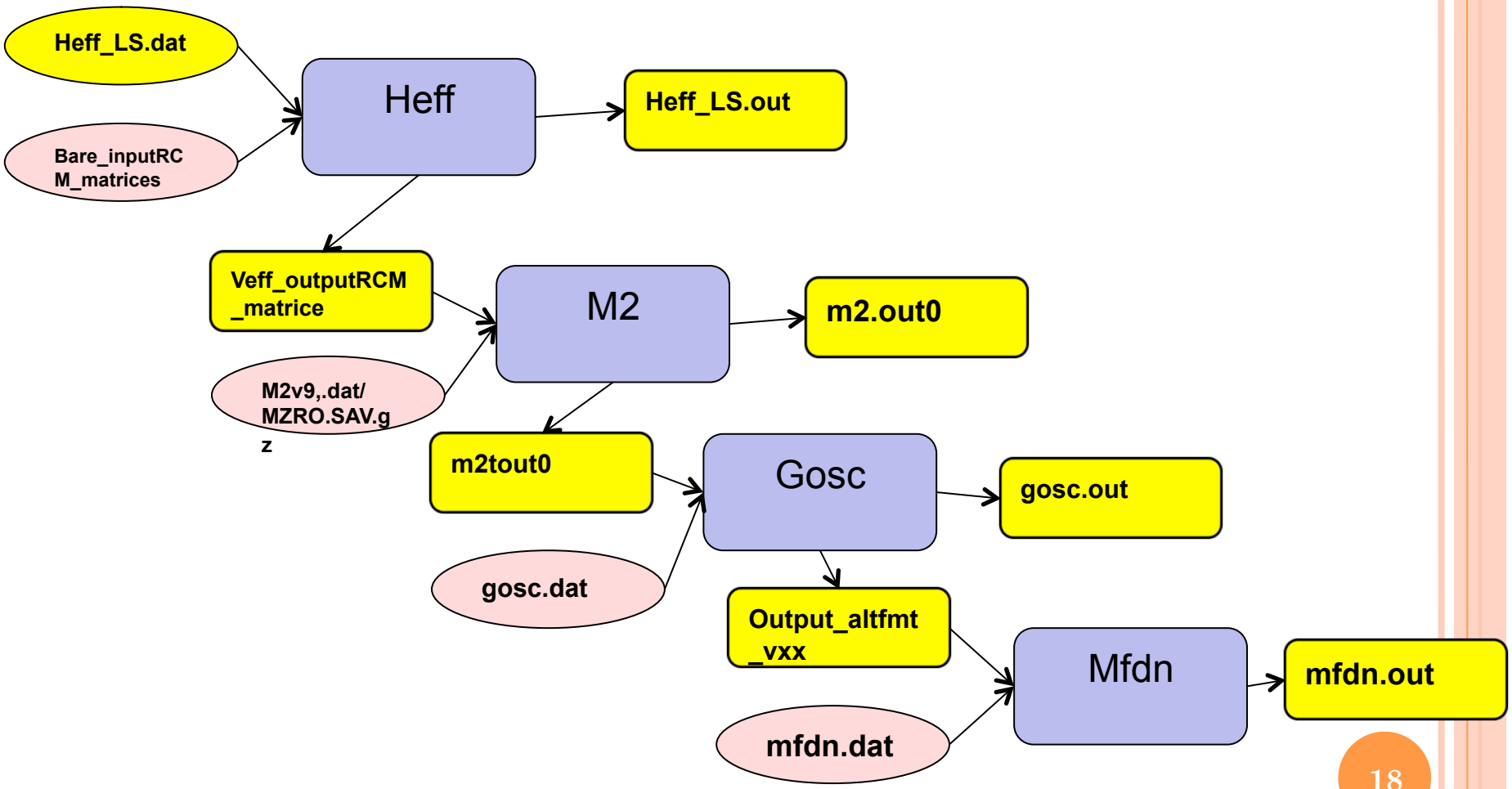
From a Data Management System to a Workflow System

- Using a database to store the provenance information for each participated software components.
- Workflow constructed with those components can also be recorded in a data management system.
- Provenance information will be extended to not only including the activities of individual component but also the combinations of the provenance information of the components.

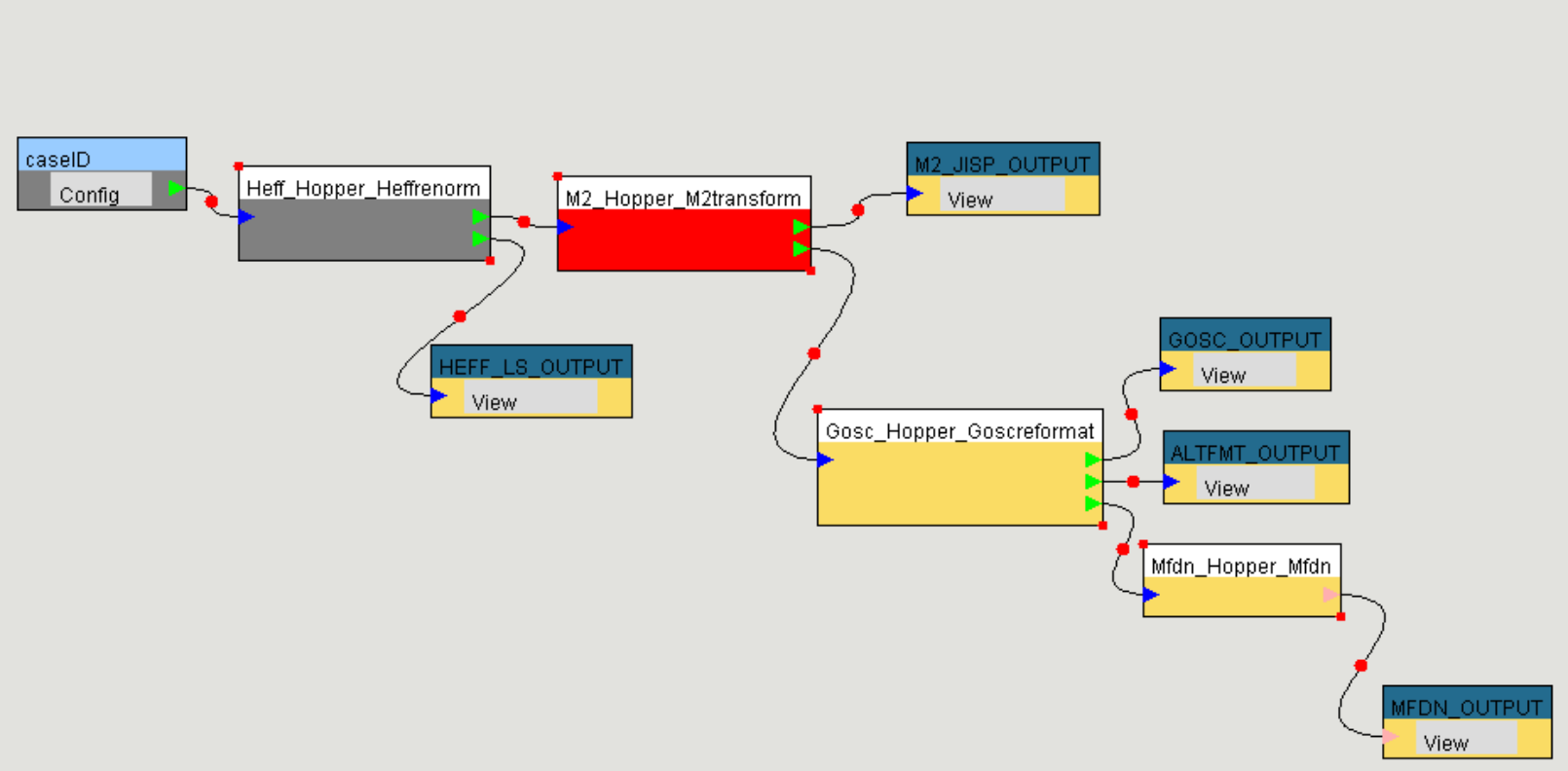
Demo workflow on upstream codes and MFDn



Detailed data flow between components



Turn into a Workflow Mode



Provenance Capturing in MFDn

- Based on existing mfdn.info file, a Karma adaptor will be written to convert the two column text file into Open provenance model compliant database schema
- Other related tools will be used to retrieve the data later.
- To record the provenance for both upstream and MFDn codes including the dependency between those codes, Karma tool will be used with OGCE toolkit.

Conclusion

- A workflow approach is adapted to LCCI code to demonstrate the feasibility of the project.
- It is first step towards the fully provenance capturing in nuclear physics community.
- An introduction of the scientific workflow and provenance capturing system is made to the community.