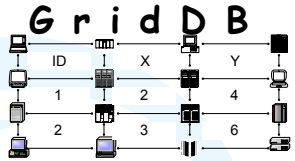
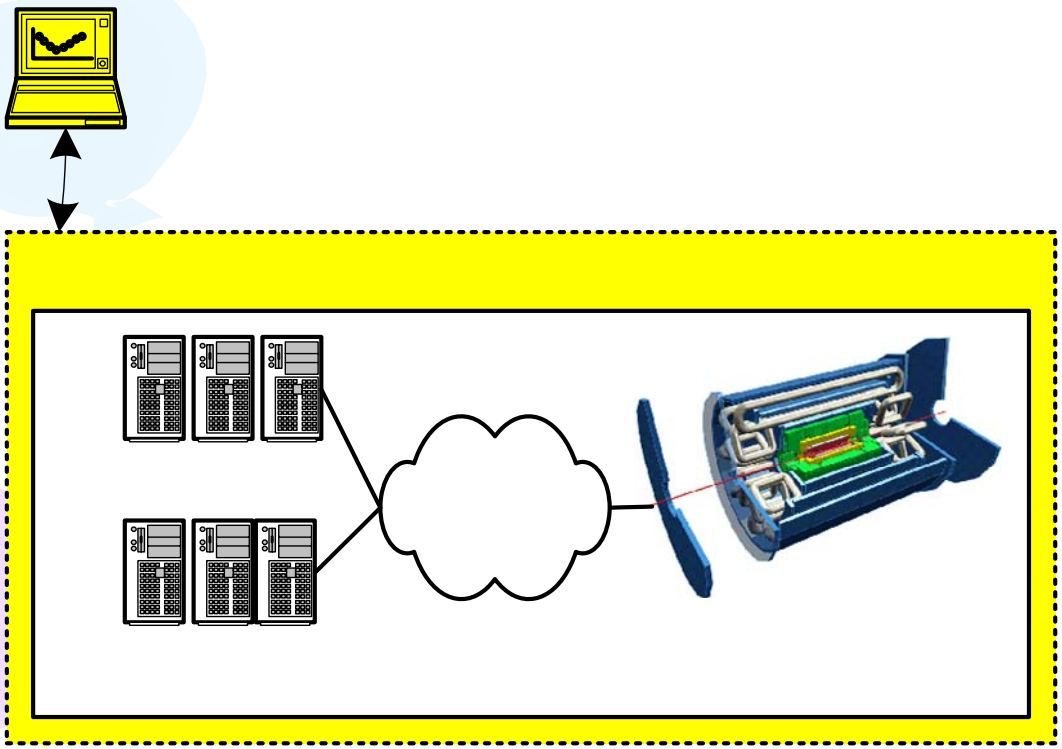


GridDB: Data-Centric Services for Scientific Grids

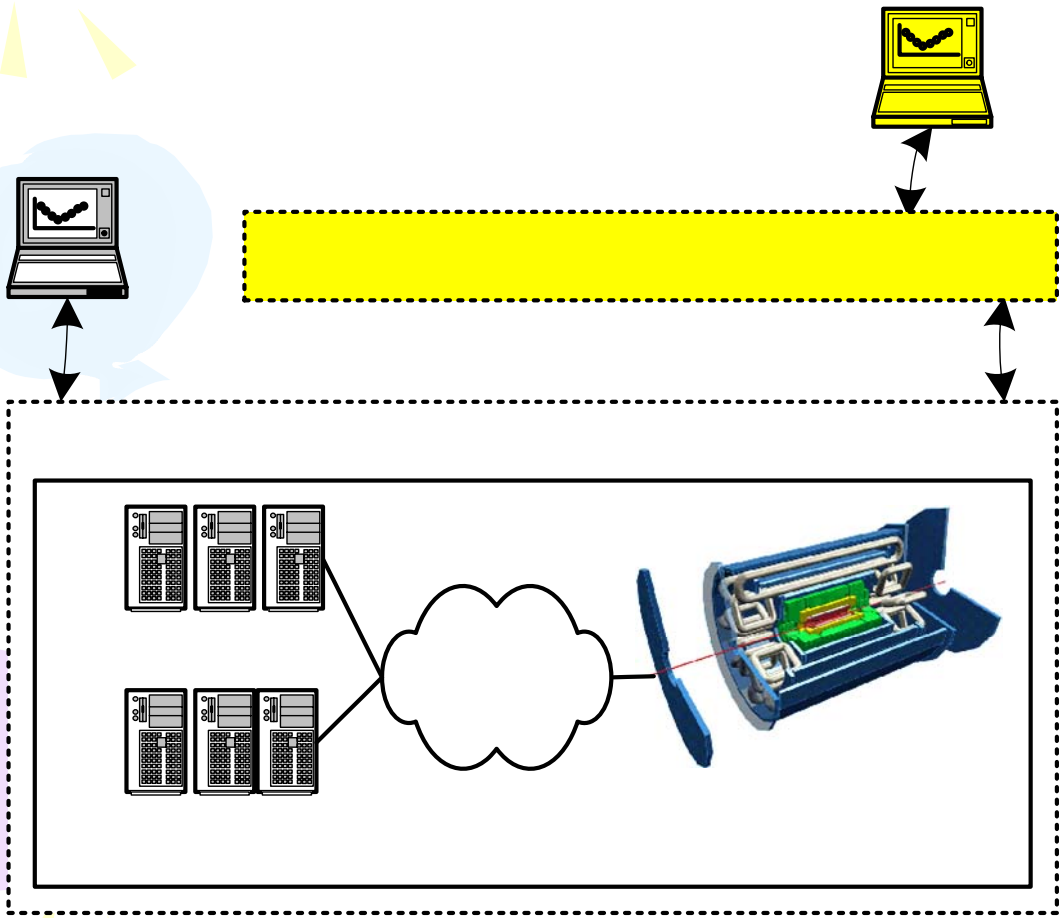
David T. Liu and Michael J. Franklin
UC Berkeley
EECS Department
Database Group

Status Quo: Process-centric middleware



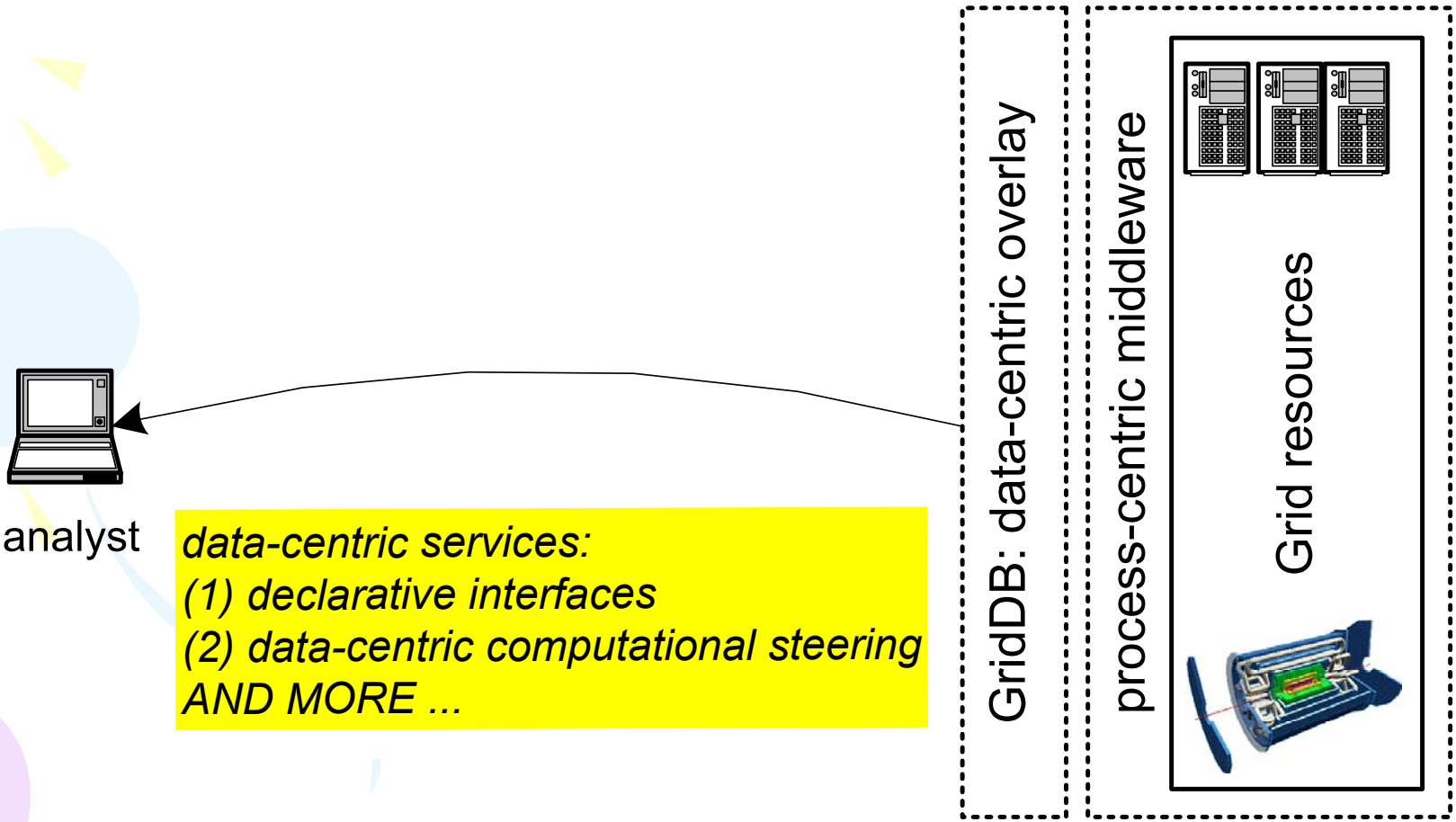
- Status Quo: Process-centric middleware; provides low-level, OS-like interfaces
- BUT: Scientific data analysis requires higher-level data-centric interfaces and services

GridDB: A Data-Centric Overlay

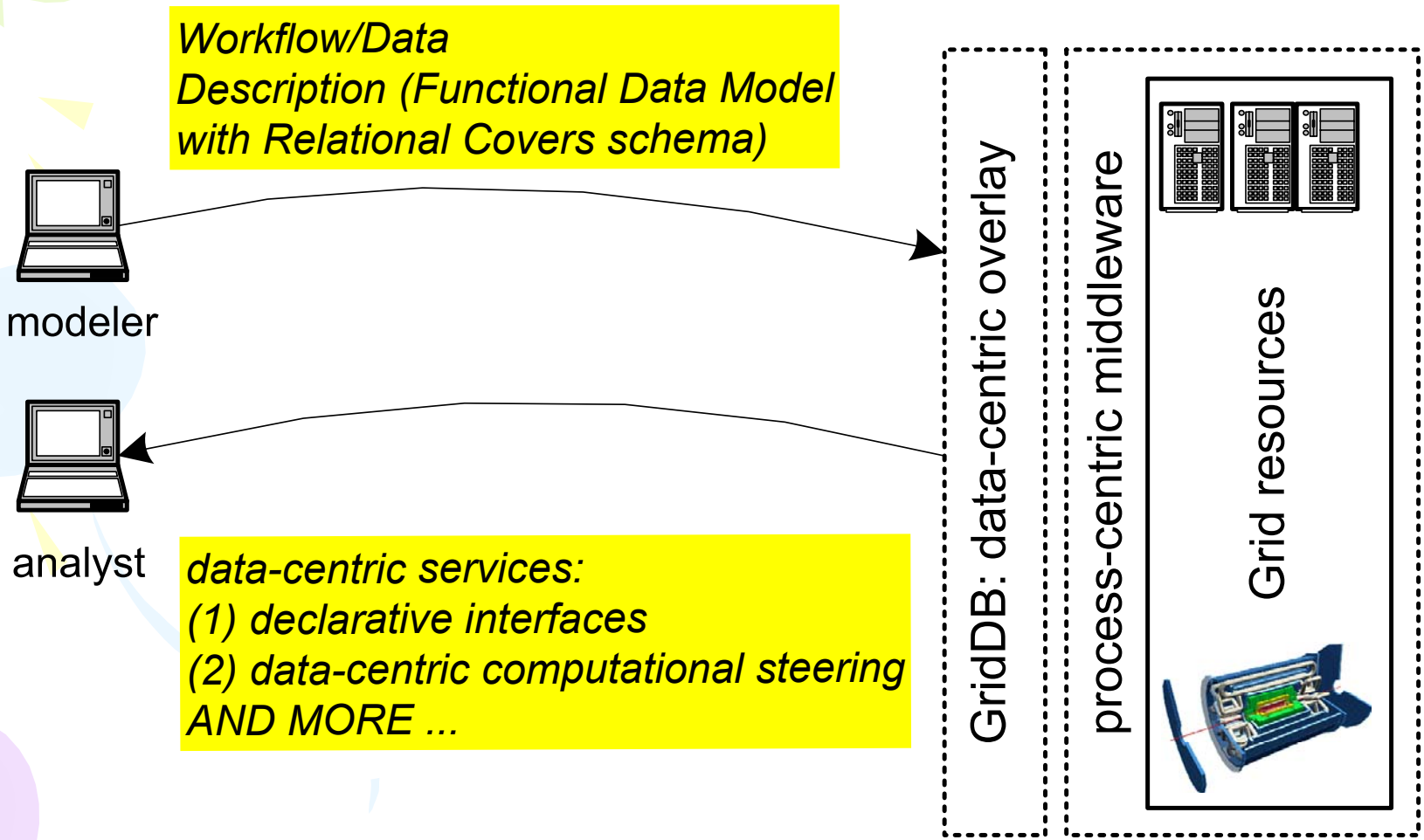


- GridDB: A data-centric overlay
  - Frame requests in terms of data rather than processes
  - Built on-top of process-centric middleware to preserve pre-existing science codes
  - A confluence of workflow and database technology

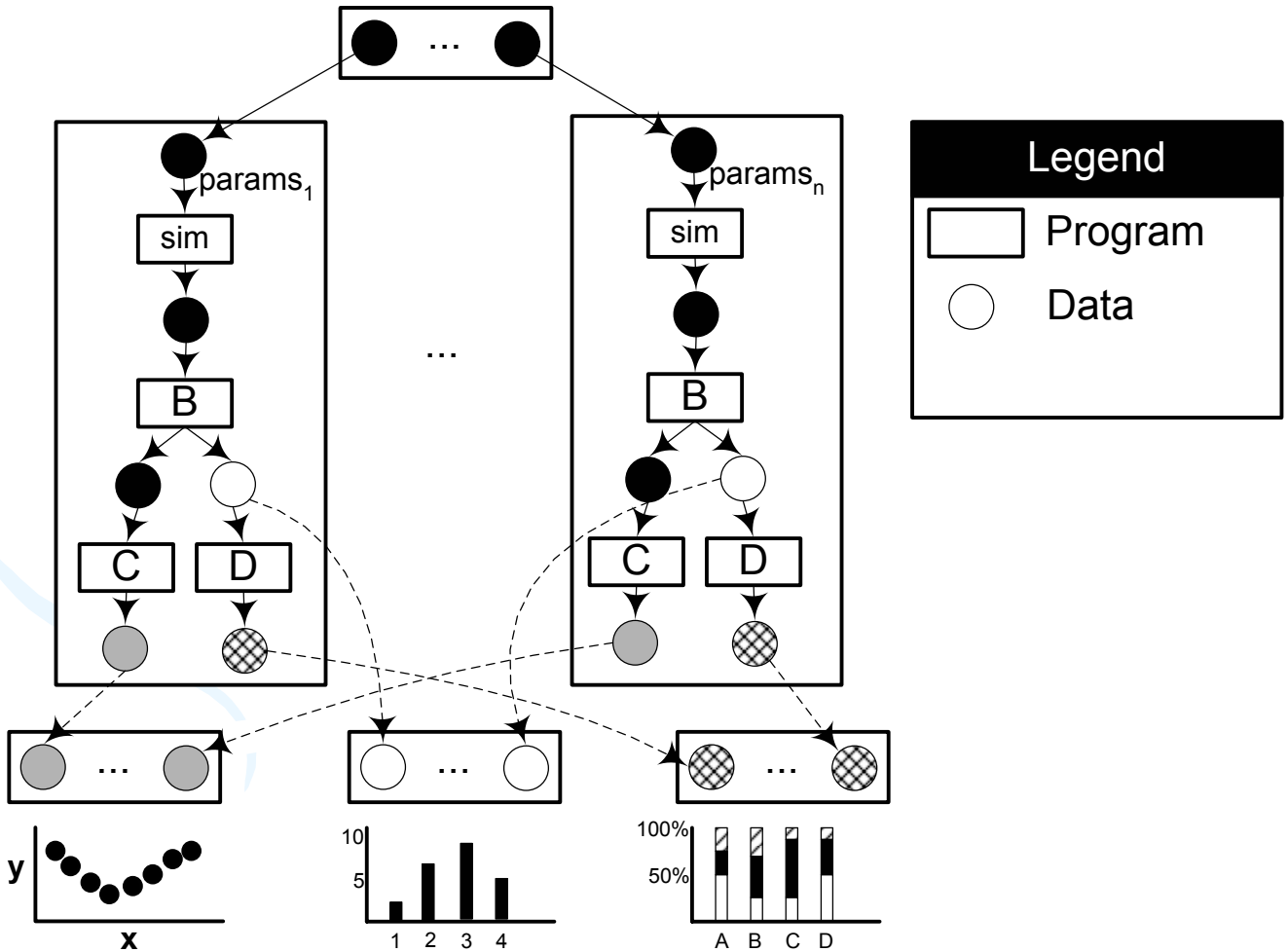
GridDB User Experience



GridDB User Experience

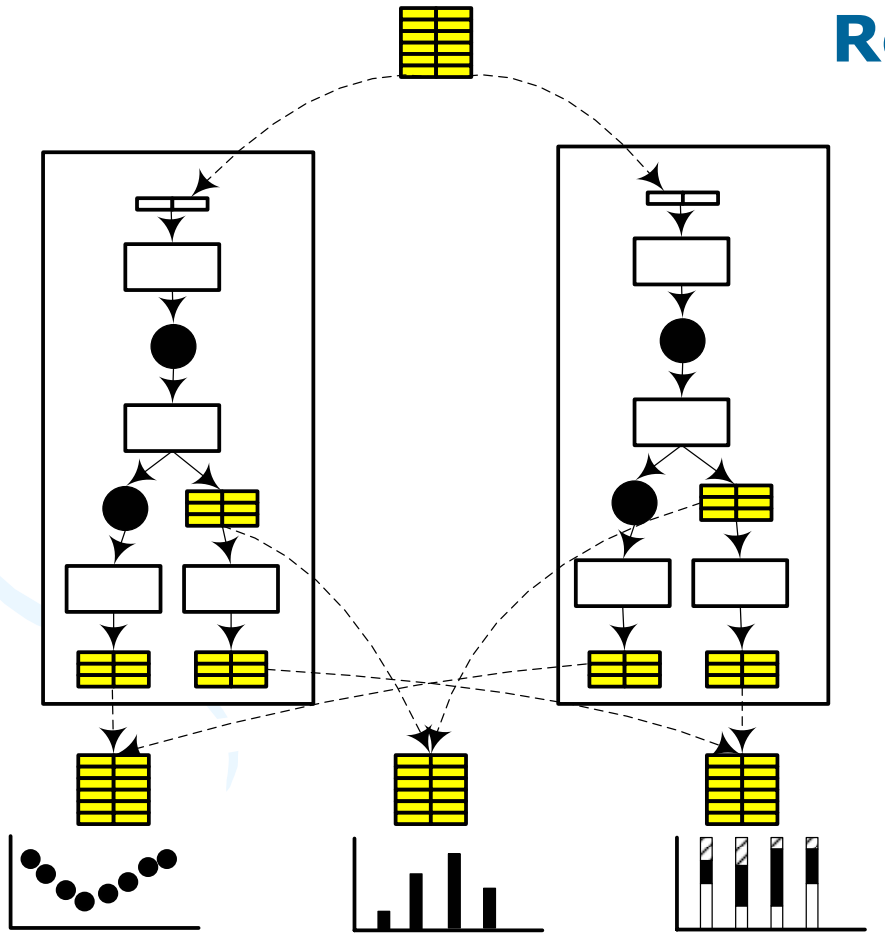


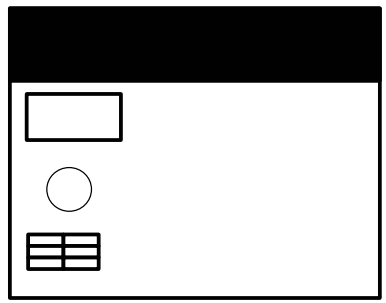
A Grid Workflow



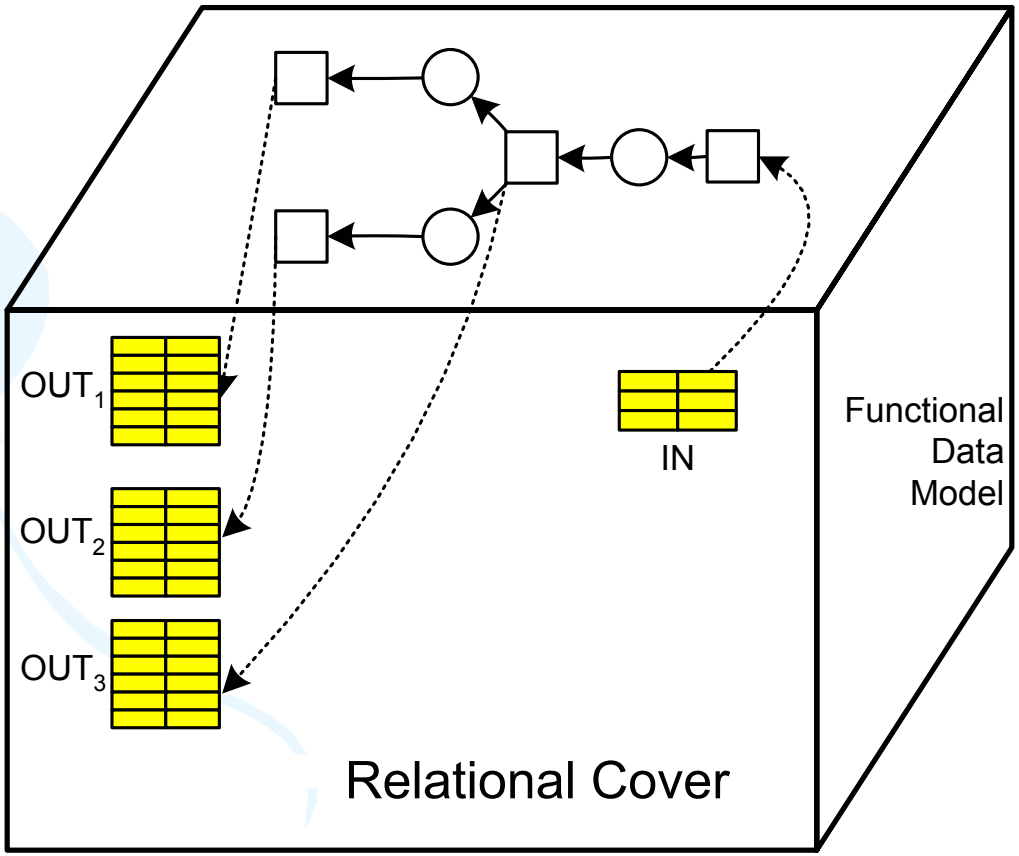
Grid Workflows Have Relations

Inputs and Outputs Relational!

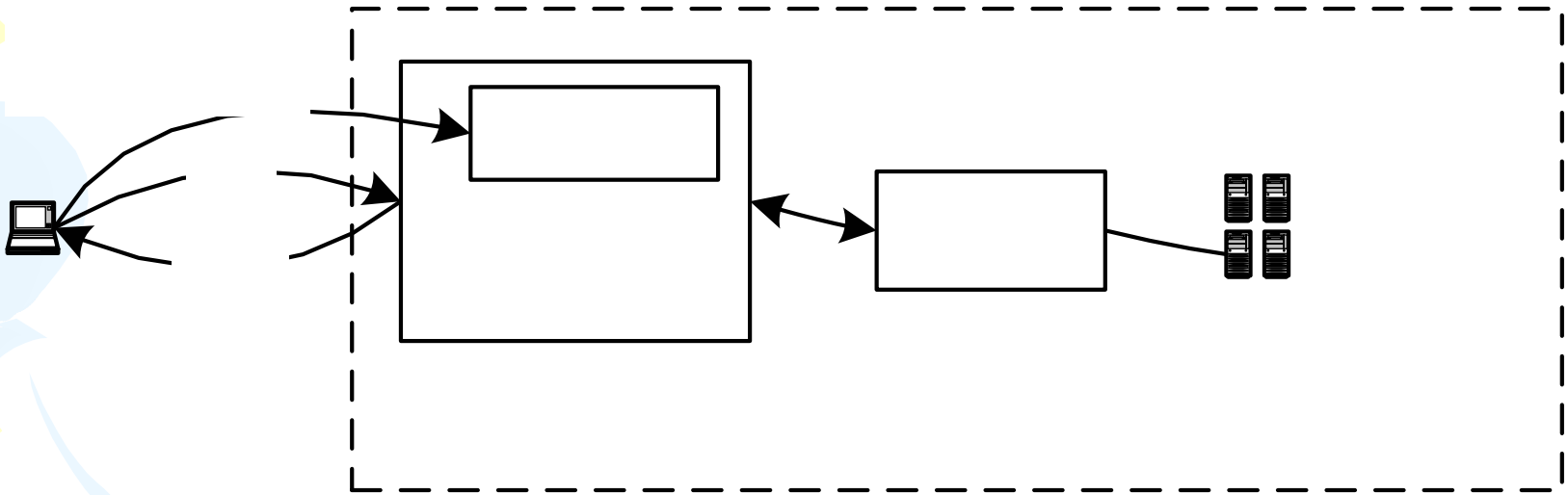




Functional Data Model with Relational Covers (FDM/RC)

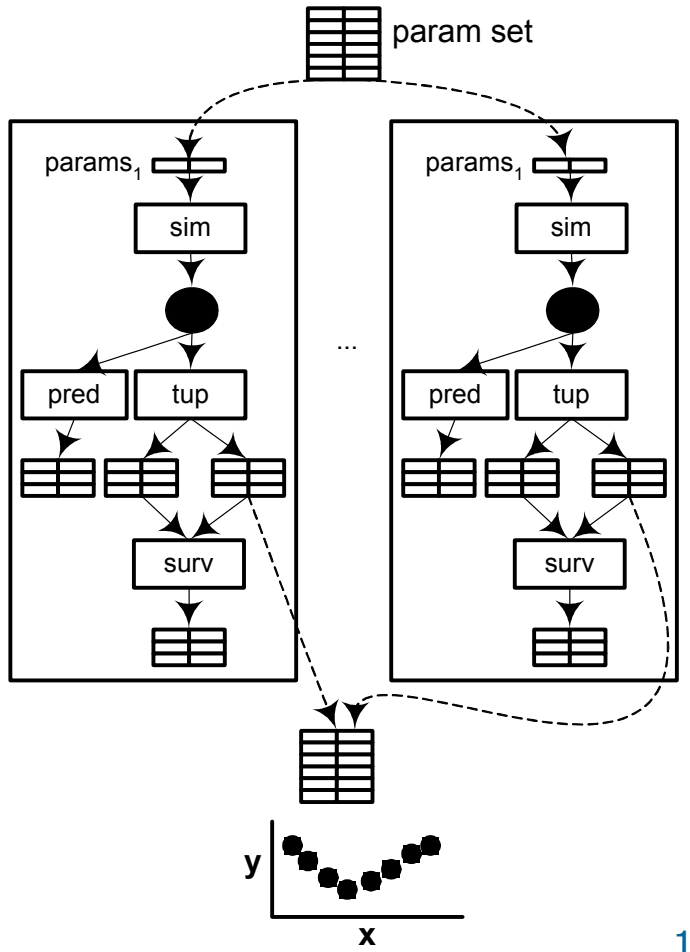


Demo Setup



The Demo: Fish N Sharks

- Setup (Data Definition Language)
  - Create Workflow, define data
- Analysis (SQL)
  - Data Procurement (INSERT)
  - View Data (SELECT)
  - Computational Steering (UPDATE)

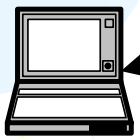


The Demo

Data-Centric Services

Workflow/Data Description (Functional Data Model with Relational Covers schema)

modeler



analyst

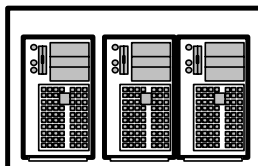
data-centric services:
(1) declarative interfaces
(2) data-centric computational steering

(3) type-checking
(4) work-sharing, memoization
(5) data provenance

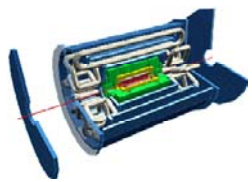
(6) query relaxation
(7) speculative computation
(8) data-centric smart scheduling

GridDB: data-centric overlay

process-centric middleware

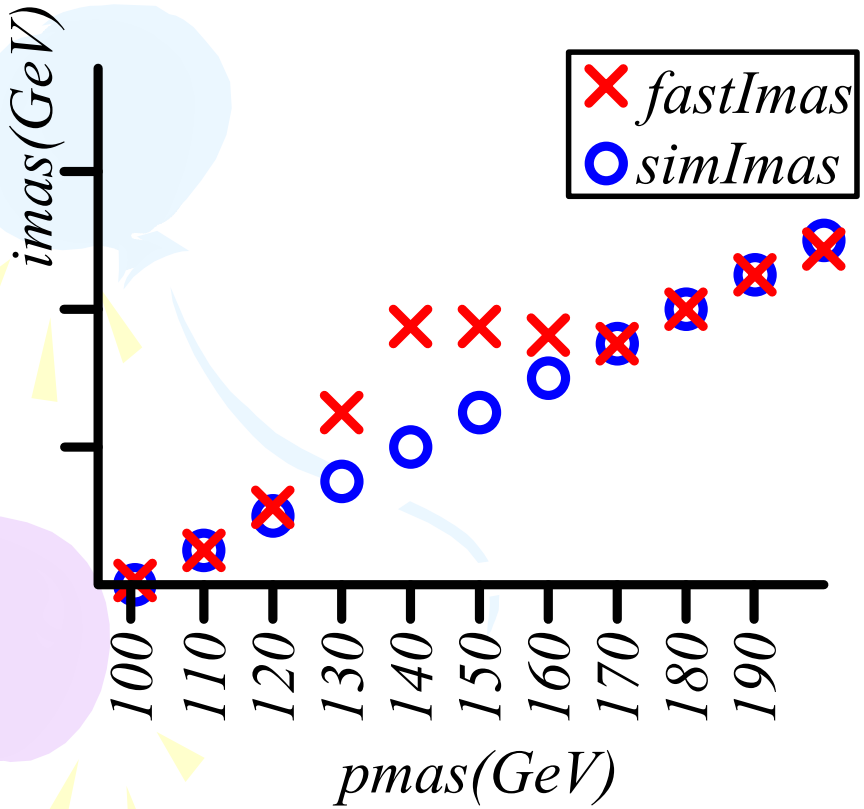


Grid resources

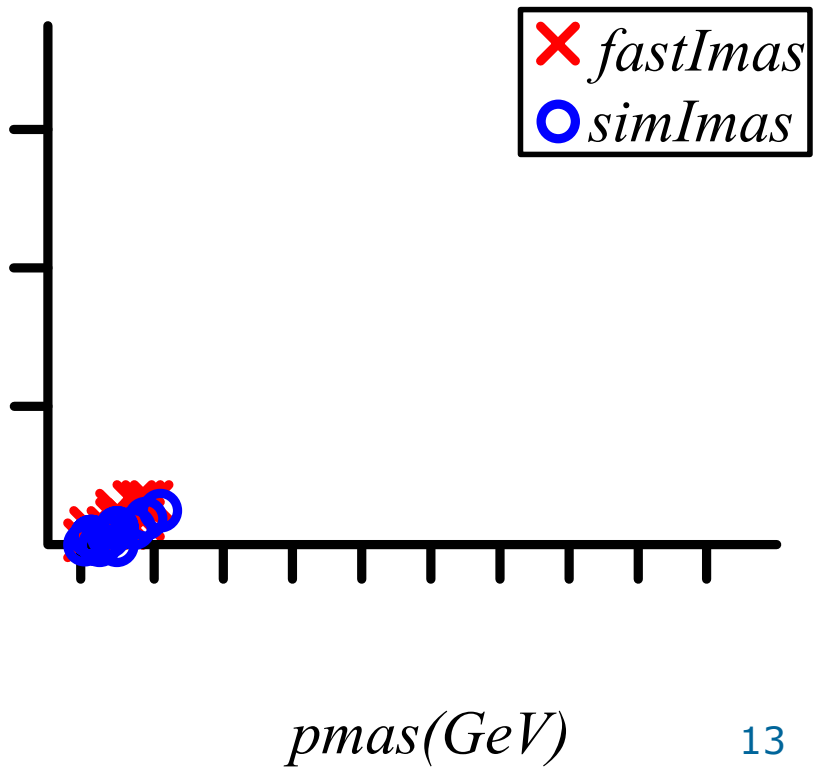


Data-Centric Smart Scheduling

pmas vs. *imas*
(20 point sample
with smart scheduling)



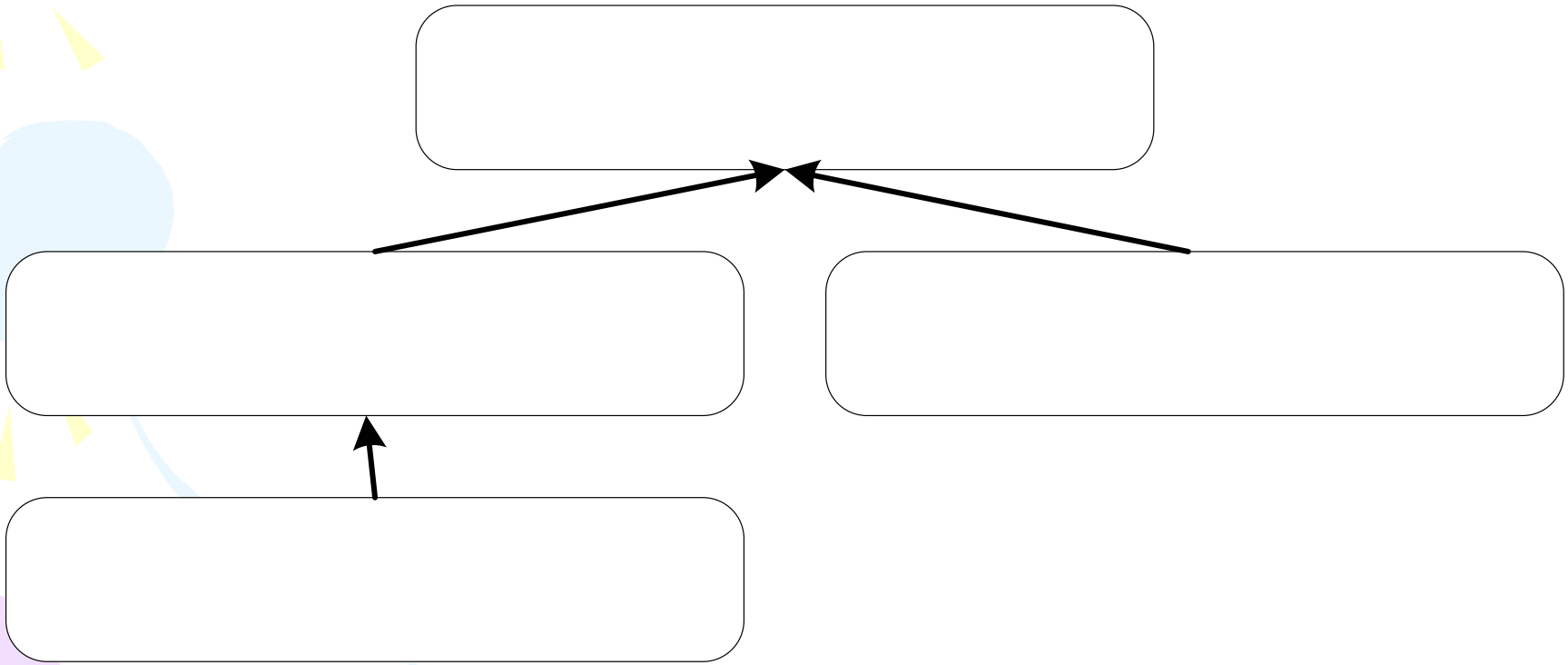
pmas vs. *imas*
(20 point sample
with dumb scheduling)



Status

- Workflow/Data Model, Data Definition Language (DDL), Data Manipulation Language (DML) developed
- Prototype built: ~19,000 LOC
- Use cases:
  - SDSS clusterfinding (with Annis)
  - ATLAS (with Tull, Gardner, Smirnov)
  - LSST (with Ghaleb Abdulla)
- Preliminary Experimental Results
- VLDB'04 paper:
  http://www.cs.berkeley.edu/~dtliu/pubs/griddb_tr.pdf

Related Work/Systems

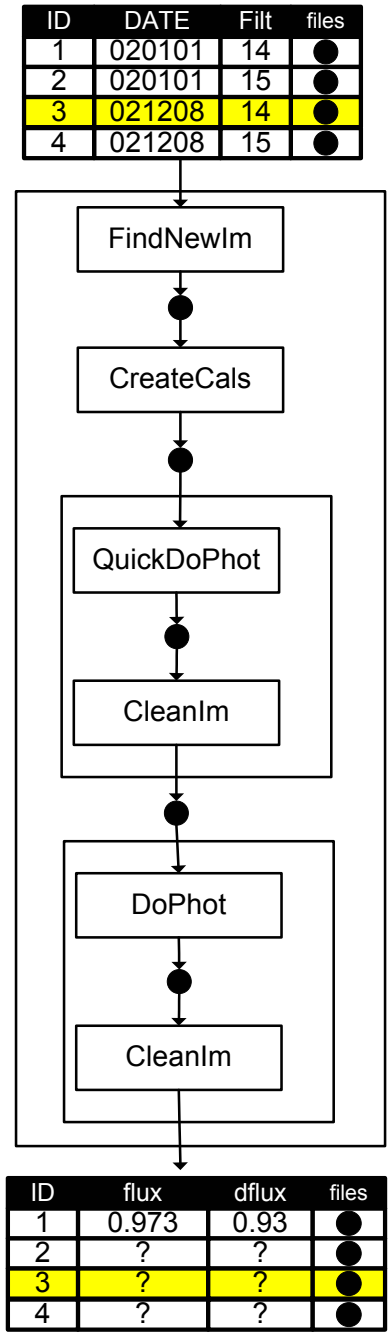


GridDB and LSST Synergies

- GridDB middleware provides services interesting to LSST: modularity, declarative interfaces, real-time monitoring and computational steering
- LSST team will provide a cutting-edge application for GridDB: real pipelines, scales, queries
- Synergy between the two collaborations
- UCB Database Group also at the forefront of Data Stream Processing (ala TelegraphCQ project)

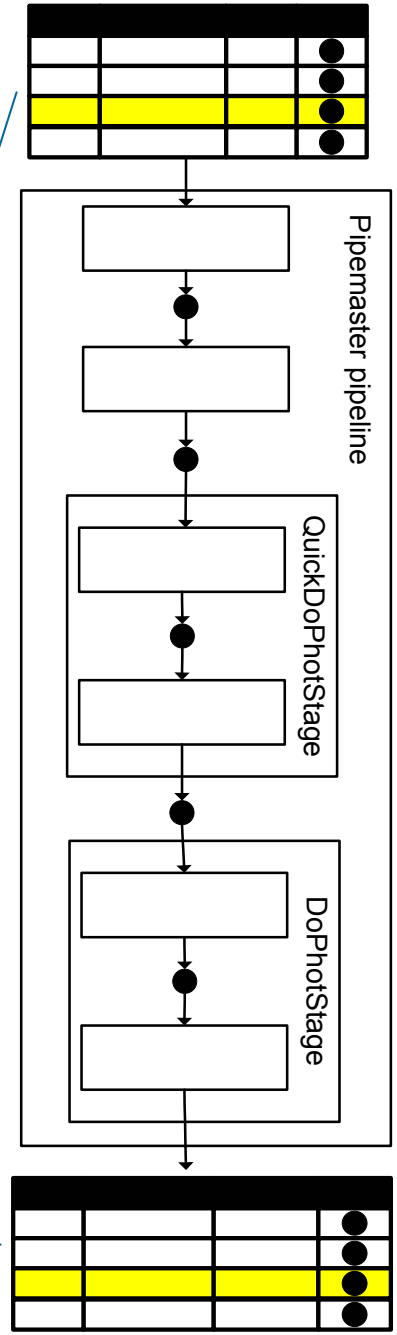
Test-Case: SuperMacho's pipemaster

- Ran SuperMacho a PipeLine through GridDB as a test-case
- Specified a model to encapsulate workflow
- Have working:
  - Data Procurement
  - Partial-Result examination
  - Computational Steering



SuperMacho's pipemaster

- Have working:
  - Data Procurement
  - Partial-Result examination
  - Computational Steering
- SuperMacho pushes GridDB to the limit!



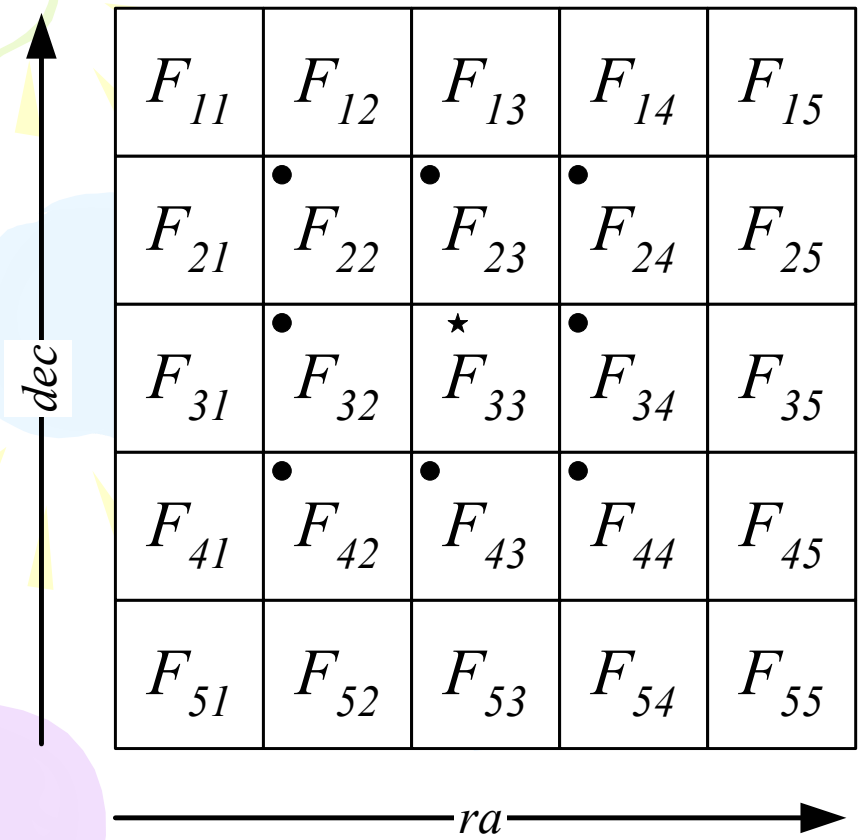
ID	DATE	Filt	flux	dflux
1	020101	14	0.973	0.93
2	020101	15	?	?
3	021208	14	?	?
4	021208	15	?	?

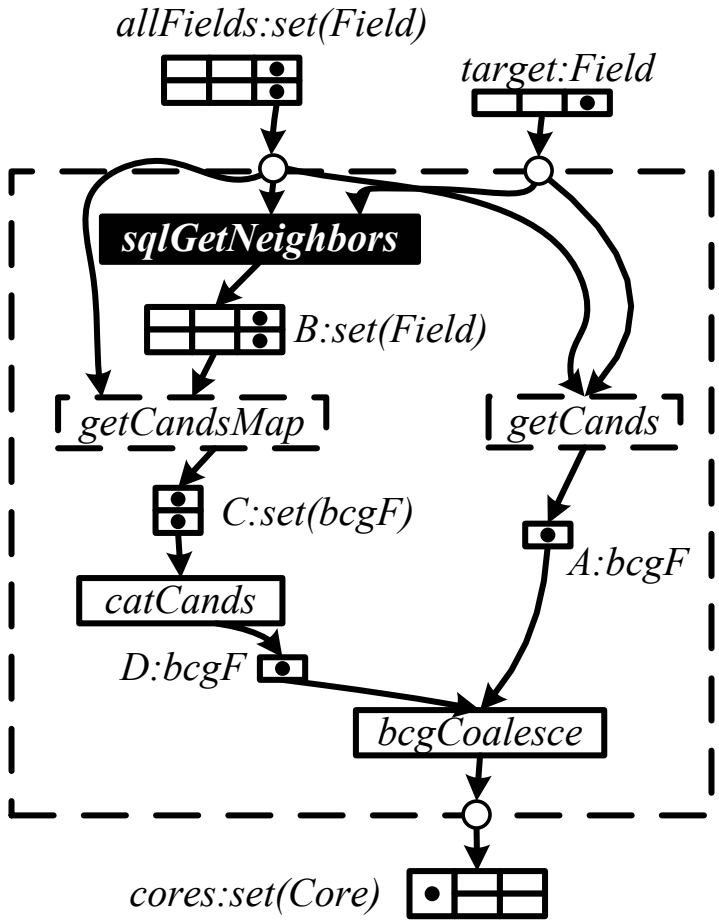
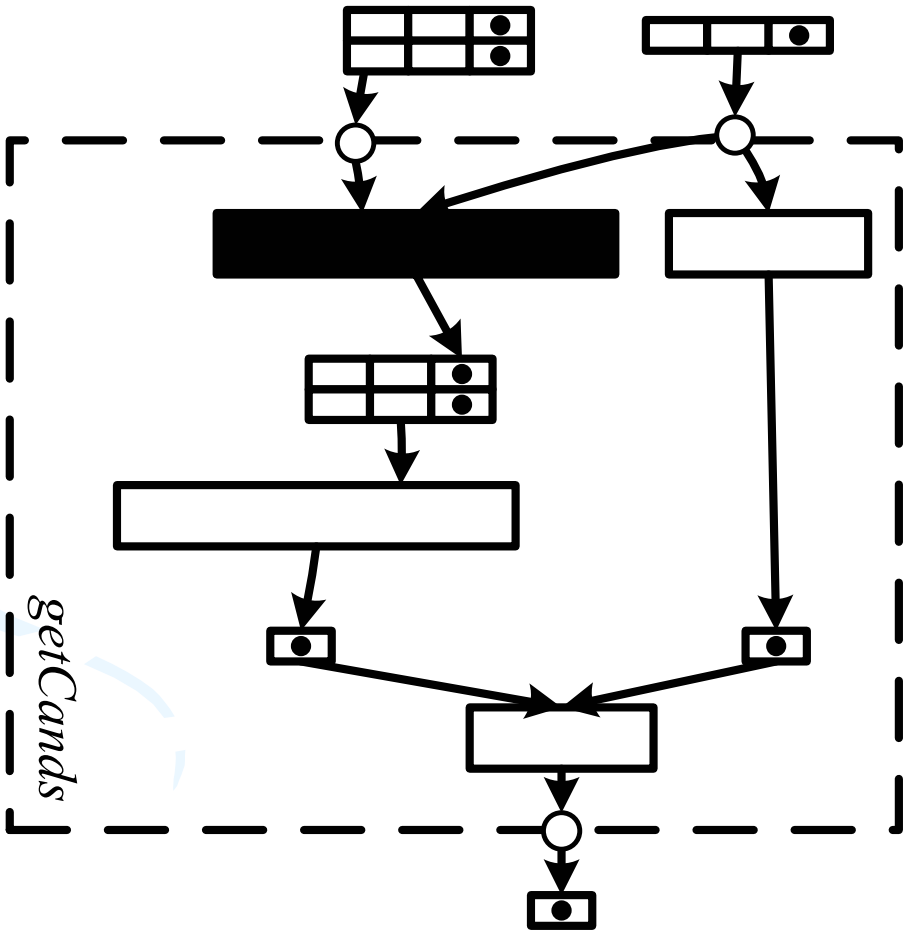
Conclusion

- The FDM/RC models both workflow and data information
- GridDB provides data-centric services by exploiting FDM/RC schema information
- Currently: Working with Ghaleb Abdulla + Livermore LSST group to expand upon our GridDB implementation of pipemaster

Backup

Astronomy Cluster Finding





getCands: a nested composite function