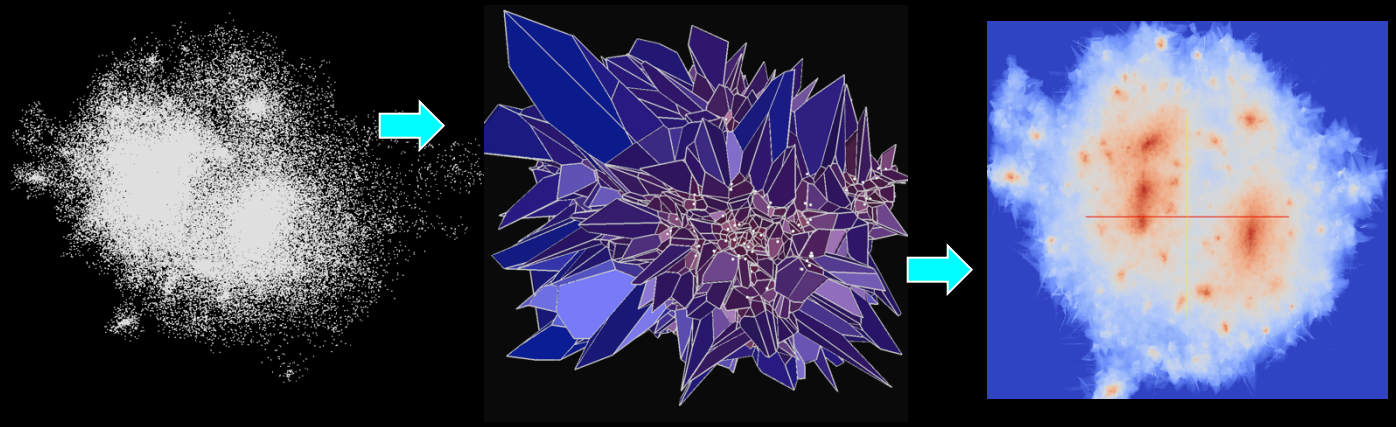
U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

From Particles to Meshes to Grids: Data Movement Within and Between Analysis Codes

Halo particles, Voronoi tessellation, and 2D density estimation



JLESC 14
Nov. 24, 2014

Tom Peterka
tpeterka@mcs.anl.gov

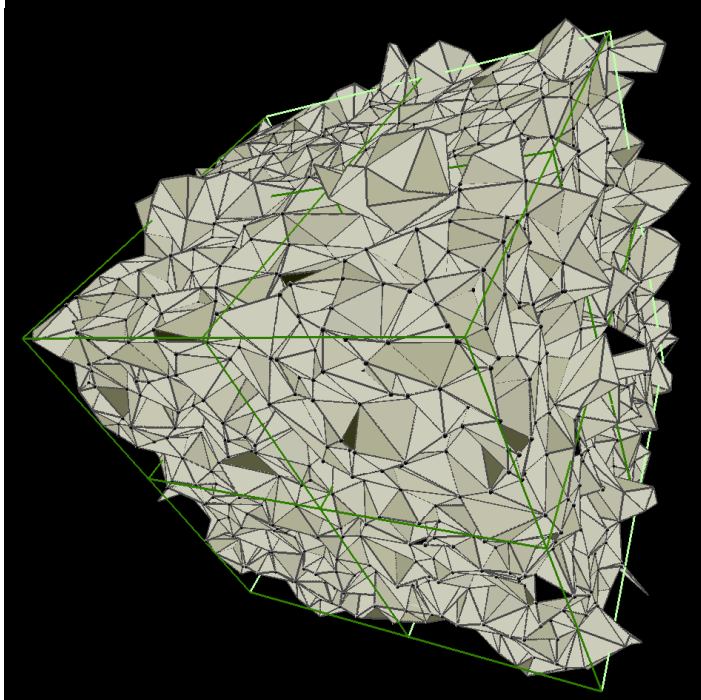
Executive Summary

Analysis of cosmology data motivates a new general-purpose way to couple simulations to analyses and multiple analysis tasks together at very large scale. We begin with a review of two specific analysis tasks, and then preview our recently awarded project to couple such tasks together with simulations.

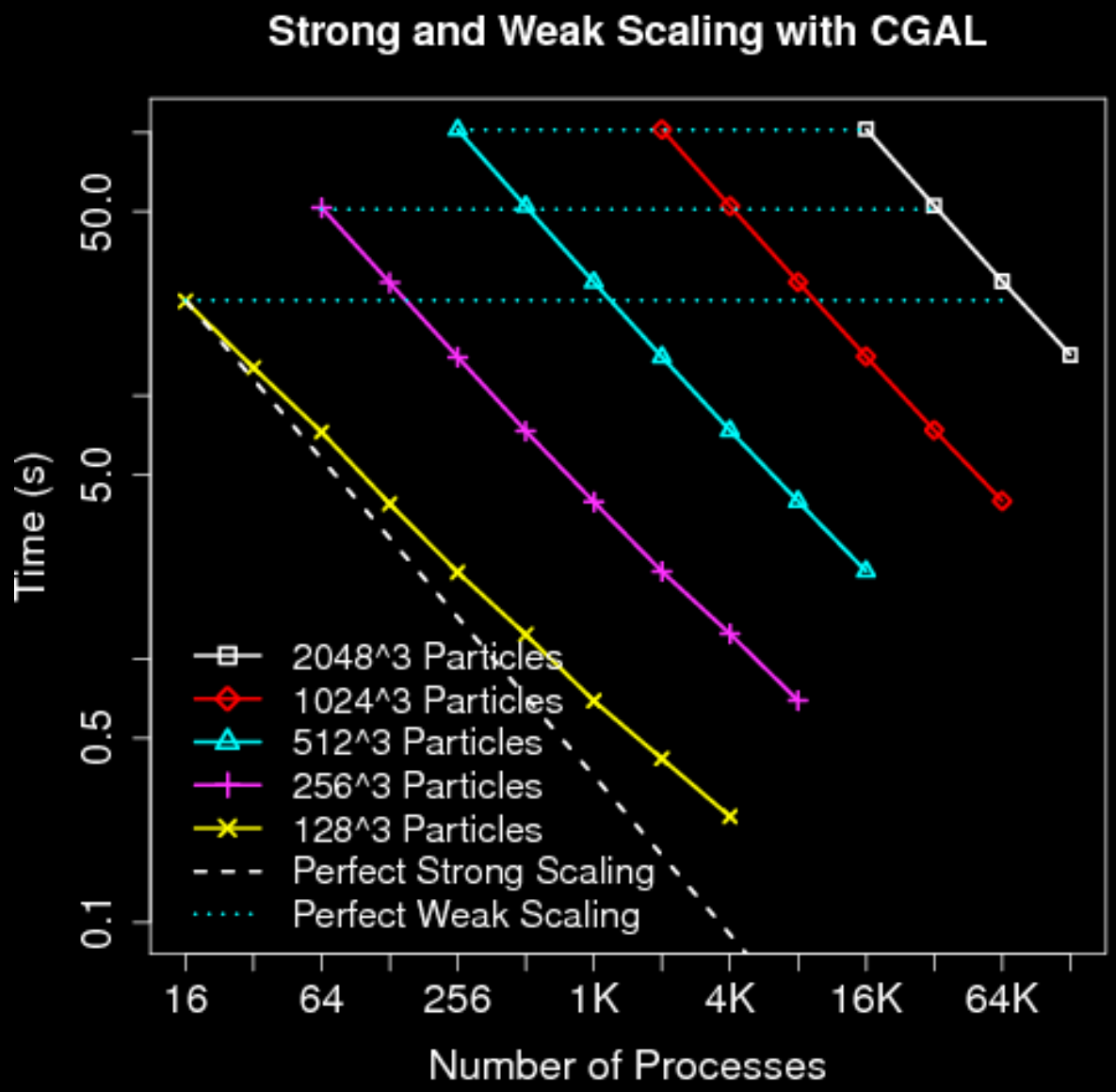
Key Ideas

- Mesh tessellations convert sparse point data into continuous dense field data.
- Tessellations are an intermediate data representation for accurate estimation of density onto a regular grid.
- The above tasks are performed at large-scale with the simulation.
- We can't keep writing new main programs for specific combinations of simulations and analyses.
- We are researching ways to couple such tasks with the advantages of tight and loose coupling (Decaf = Decoupling tightly coupled data flows).
- A separate dataflow between producer and consumer enables:
  - Aggregation, deep data permutations
  - Automatic buffering
  - Data redistribution and pipelining
  - Resilience to faults

Scalability of Voronoi Tessellation

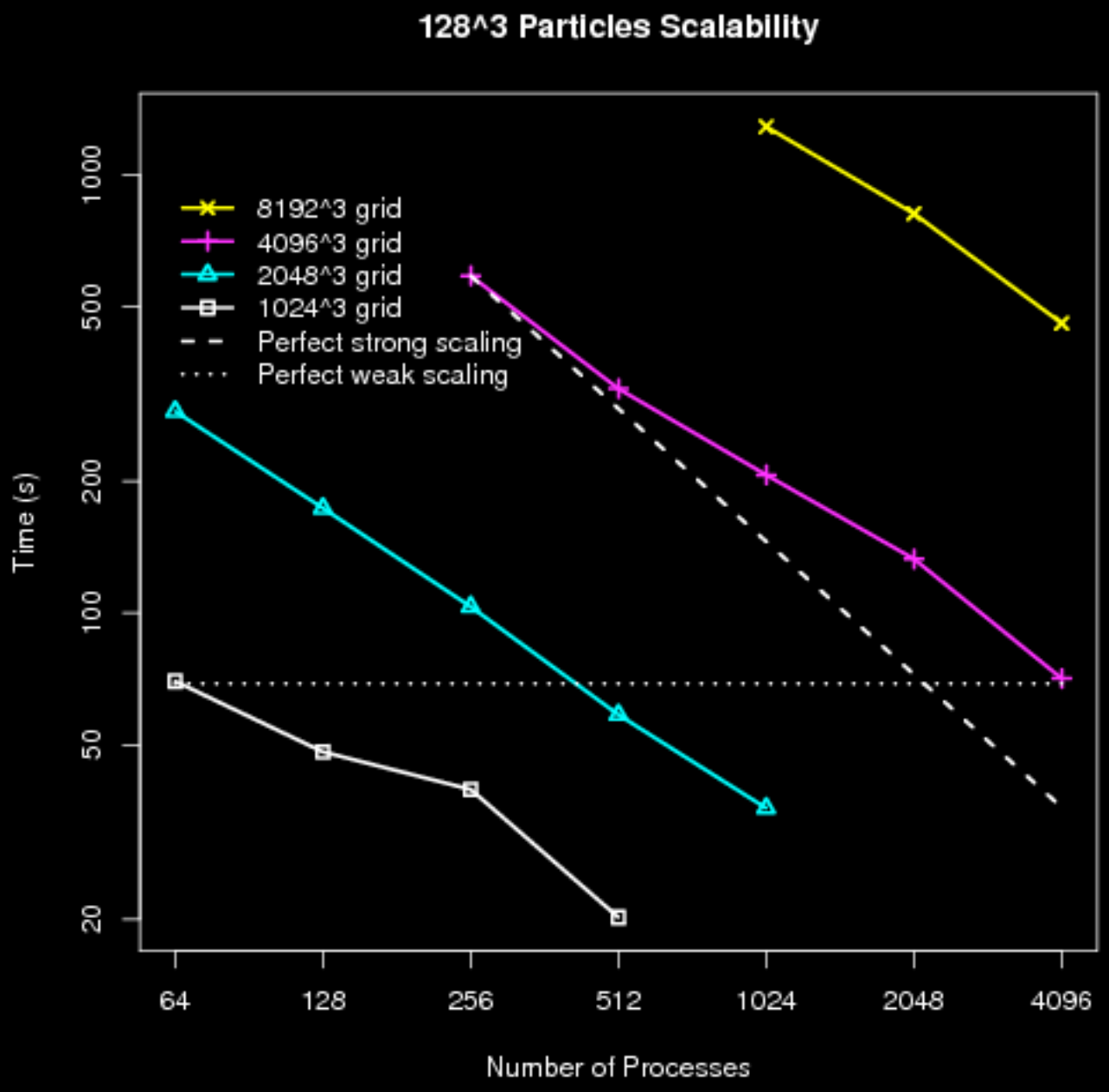


Strong and weak scaling for up to 2048^3 synthetic particles and up to 128K processes (excluding I/O) shows up to 90% strong scaling and up to 98% weak scaling.



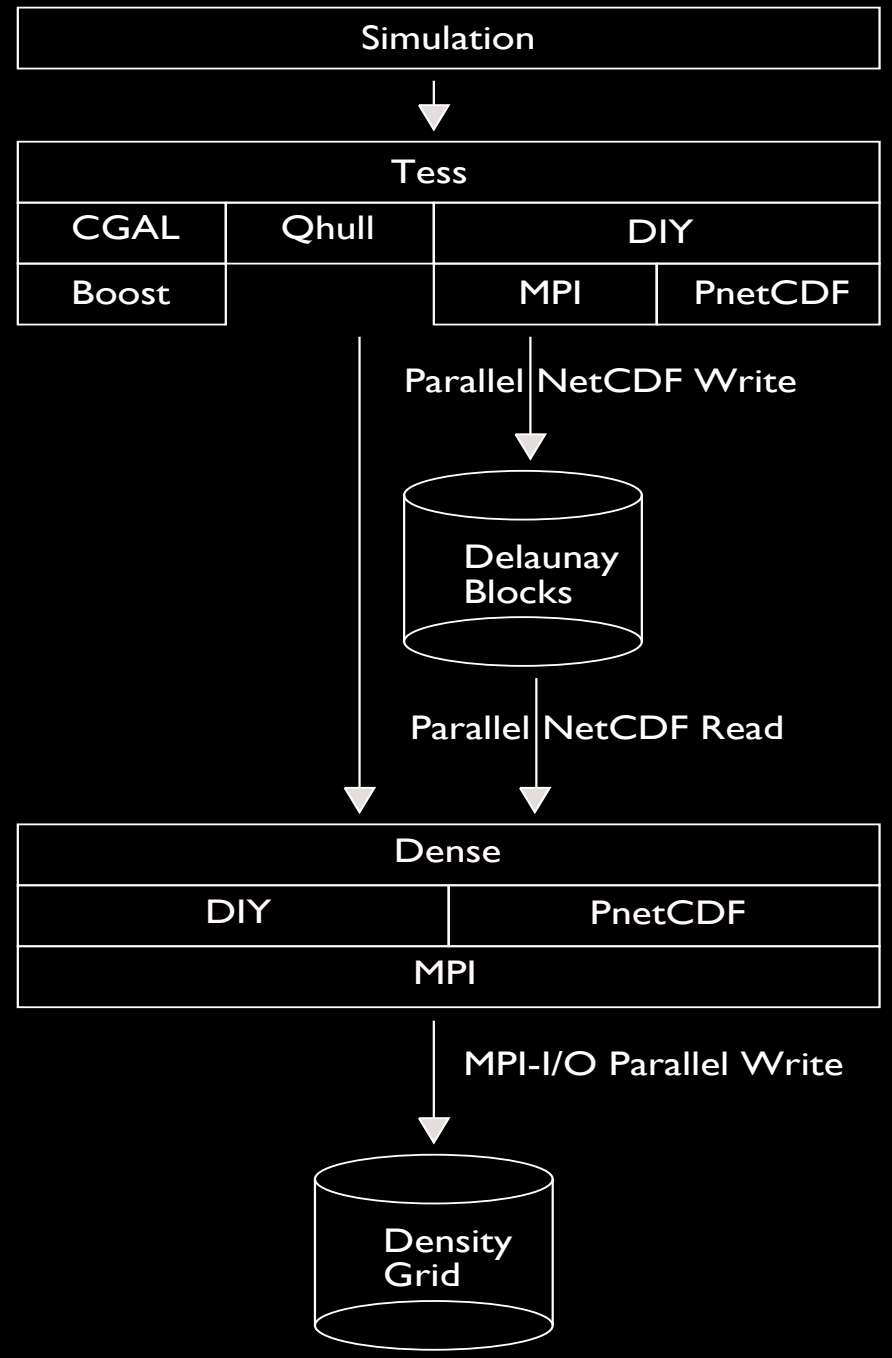
Scalability of Density Estimation

- 128^3 synthetic particles
- End-to-end time (including reading tessellation and writing image)
- 3D->2D projection
- 51% strong scaling (End-to-end) for 4096^3 grid

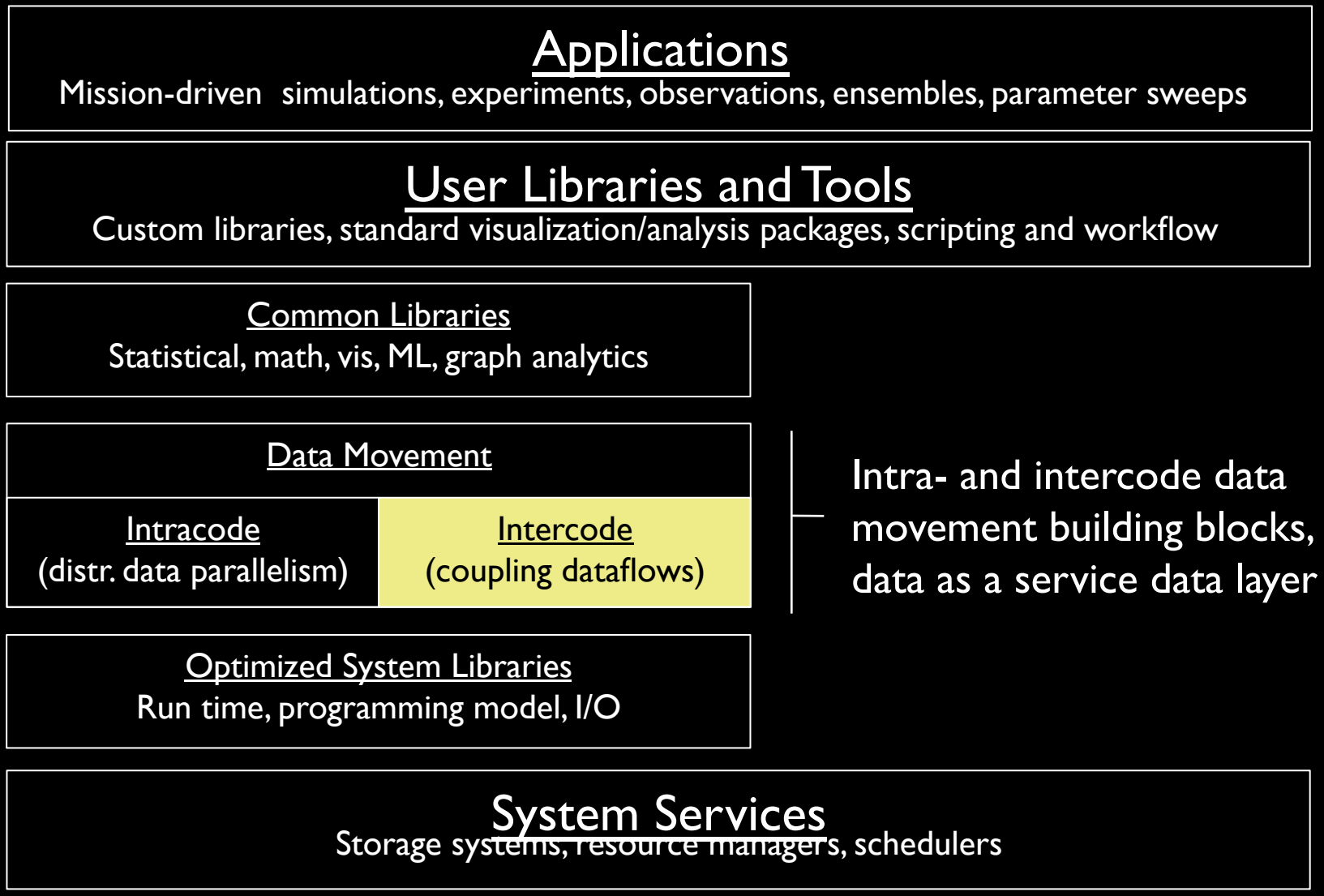


Custom Coupling of Software

- Today, we write analysis tasks as libraries with a different driver for each combination of analysis tasks.
- Writing custom one-off main programs for each combination of producer and consumer is not a scalable approach.
- Neither is tuning the producer (number of nodes, output size, etc.) to the consumer and vice versa.
- Producers and consumers ought to be written independently, and generic coupling software should manage their connection.

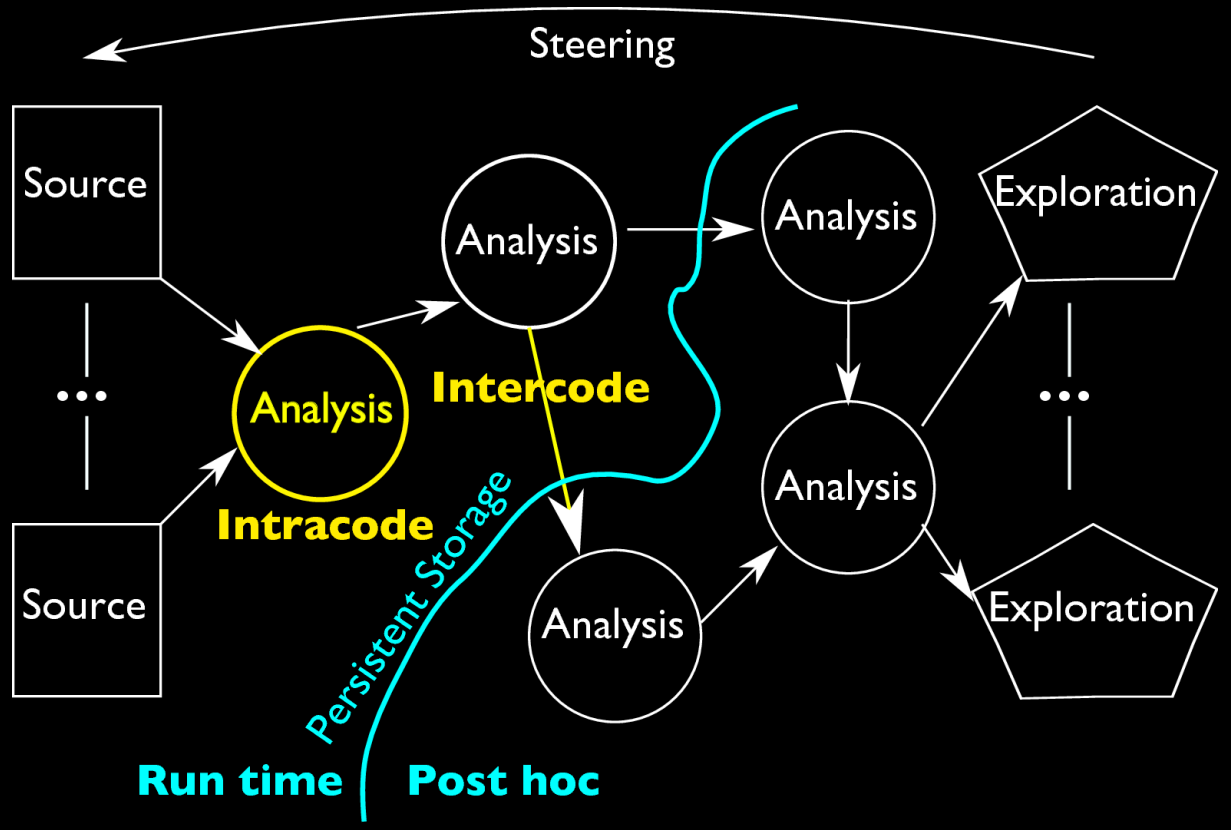


A More Generic Approach



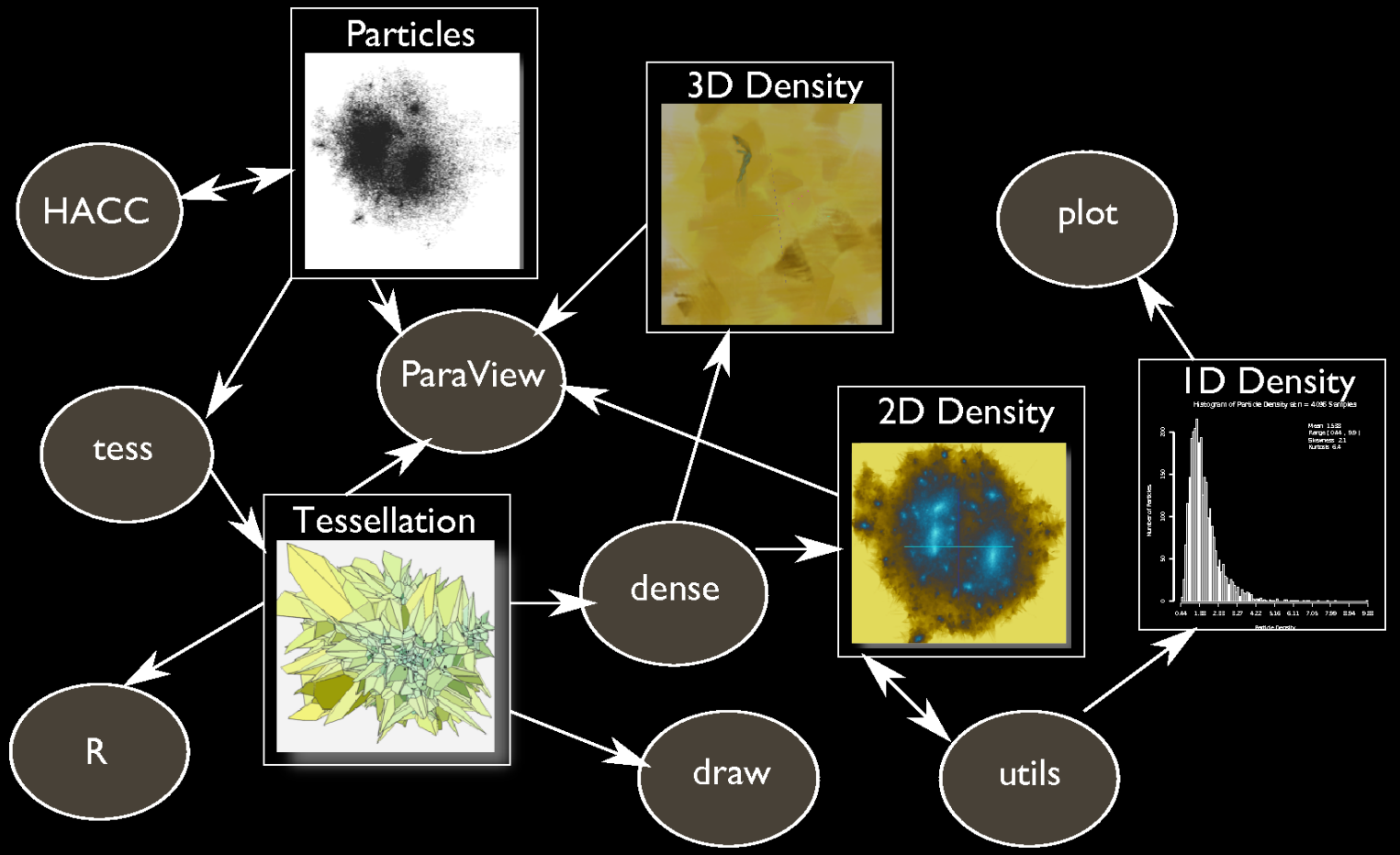
Analysis Workflow

Analysis = Any data transformation, or a network or transformations. Can be visual, analytical, statistical, or data management. Anything done to original data beyond its original generation.



Generic analysis data flow graph, primarily for simulation data, single or ensemble sources and multiple users. Results are written to persistent storage at the cyan line that partitions the graph into operations done at run time (in situ) and post hoc.

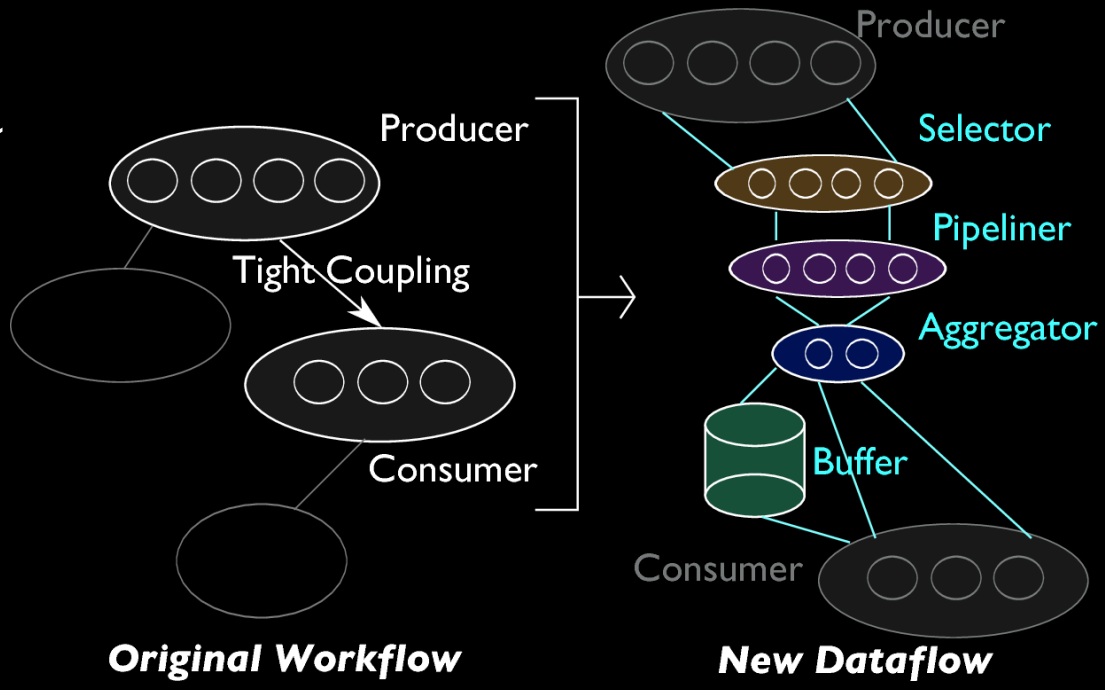
Workflow in One Application

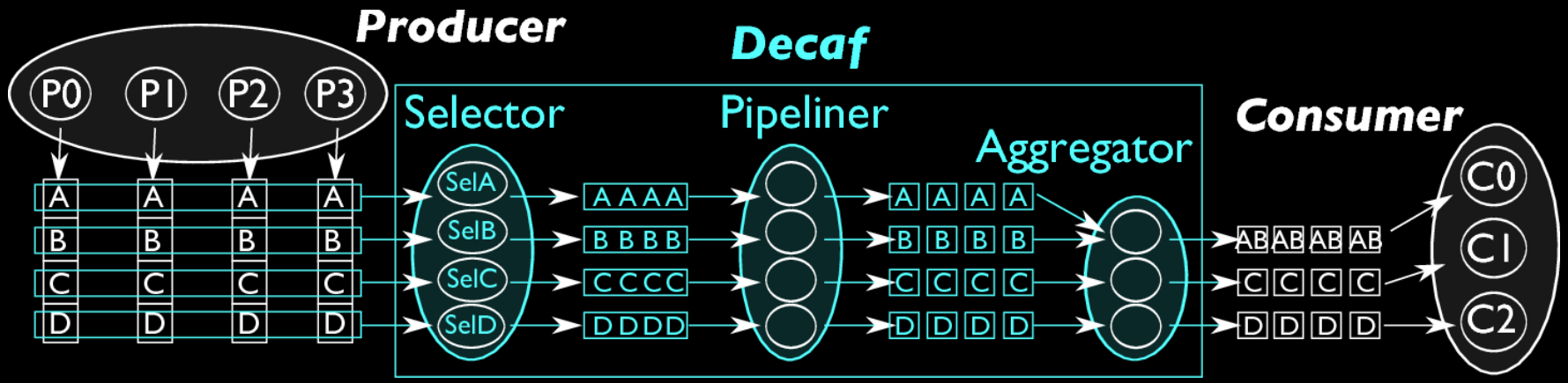


Example of a data flow network in the analysis of an N-body cosmology simulation.

Decaf: Decoupling Tightly Coupled Data Flows

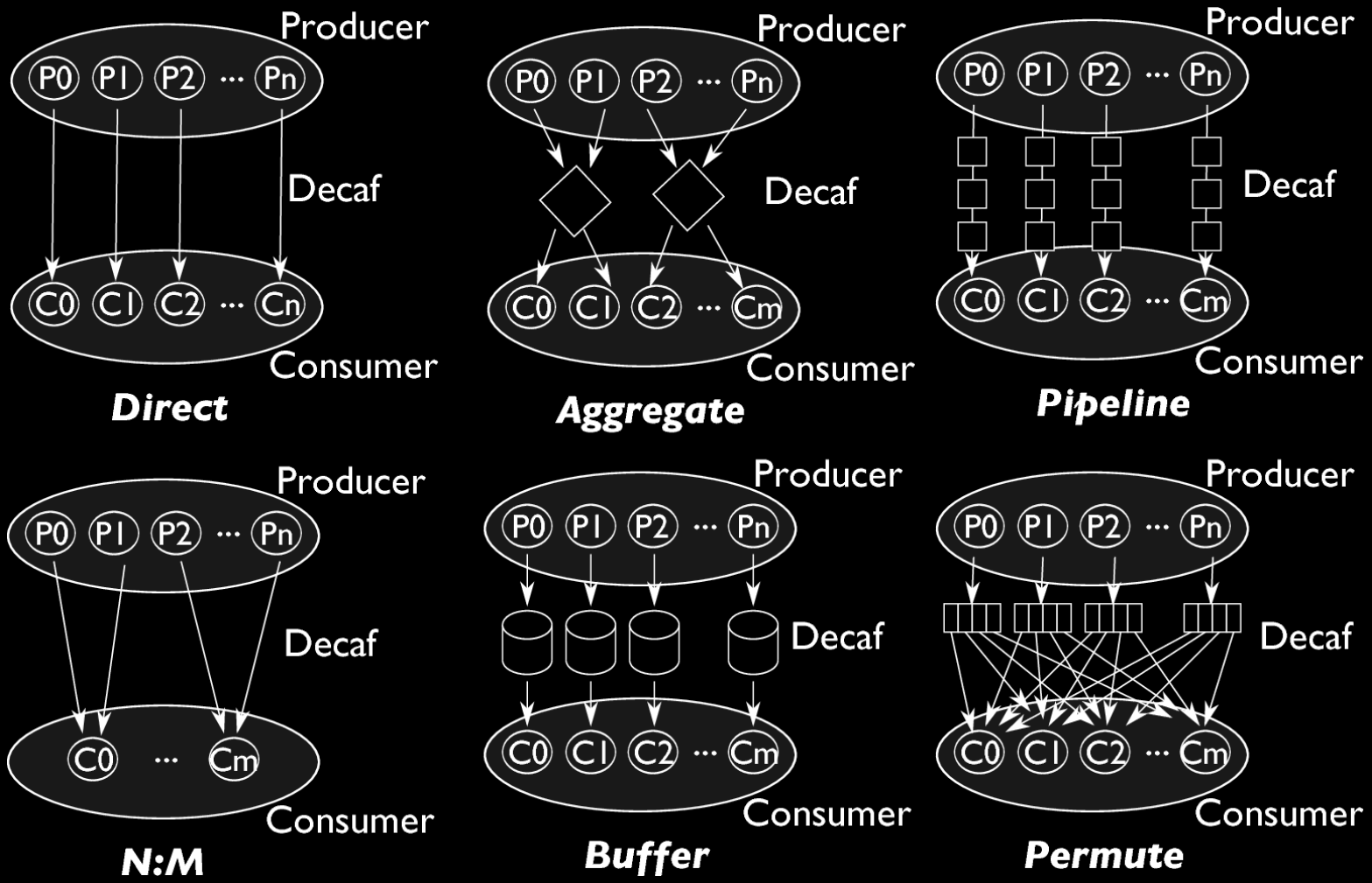
Decoupling by converting a single link into a dataflow enables new features such as fault tolerance and improved performance. We are building a generic coupling library out of 4 primitives that can be used for many purposes.





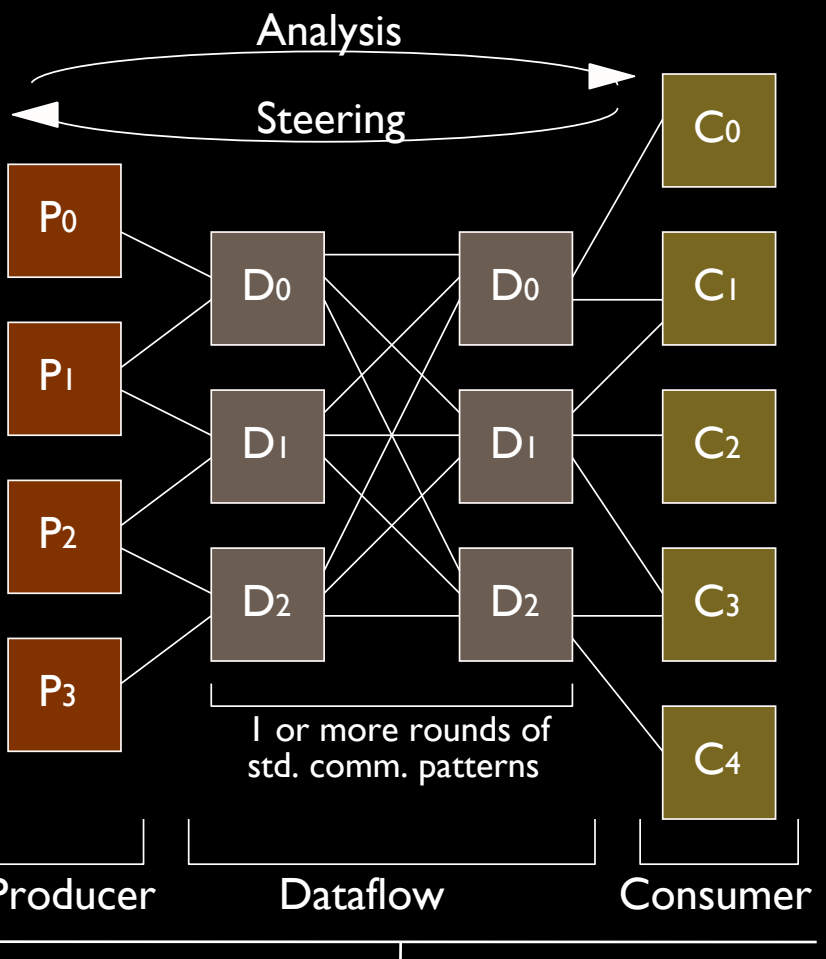
Example of selecting individual variables across nodes and combining two variables when coupling a producer (eg. CFD simulation) with a consumer (eg. visualization).

Decaf Modes

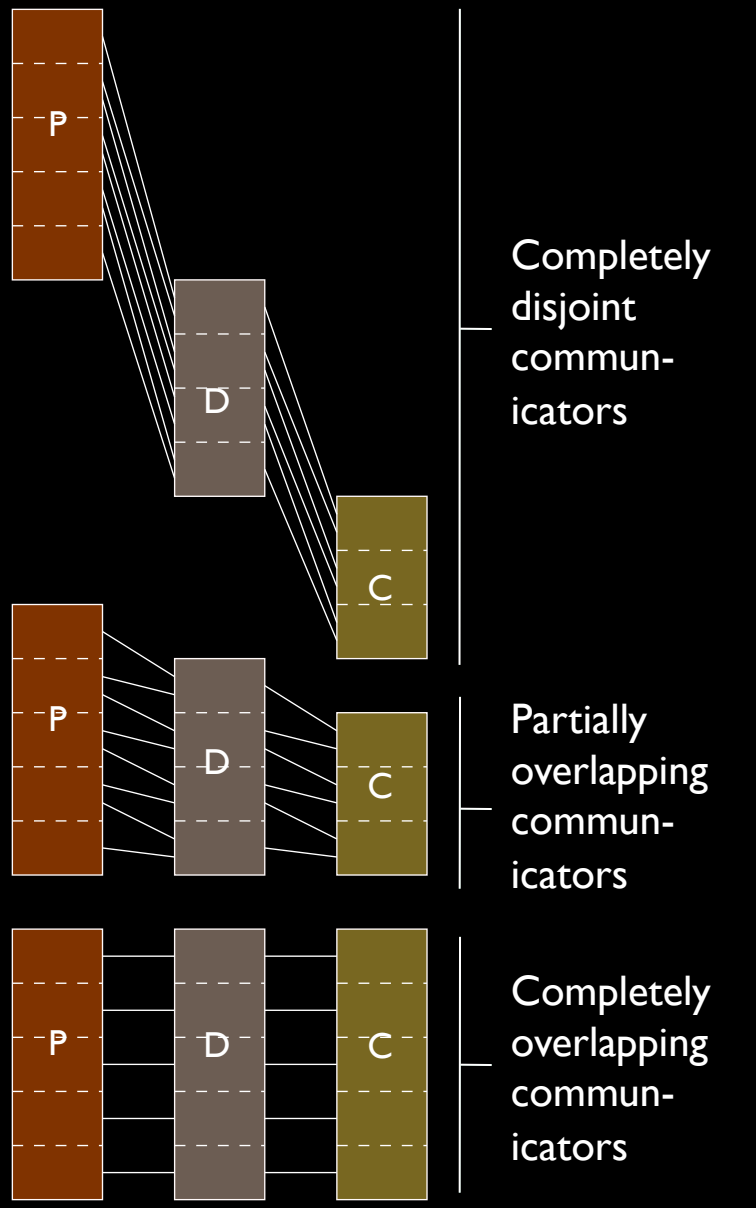


Major Decaf modes include aggregation, pipelining, and automatic buffering while potentially permuting data in an N:M and direct coupling of parallel codes.

Dataflow using Abstract Flexible Communicators

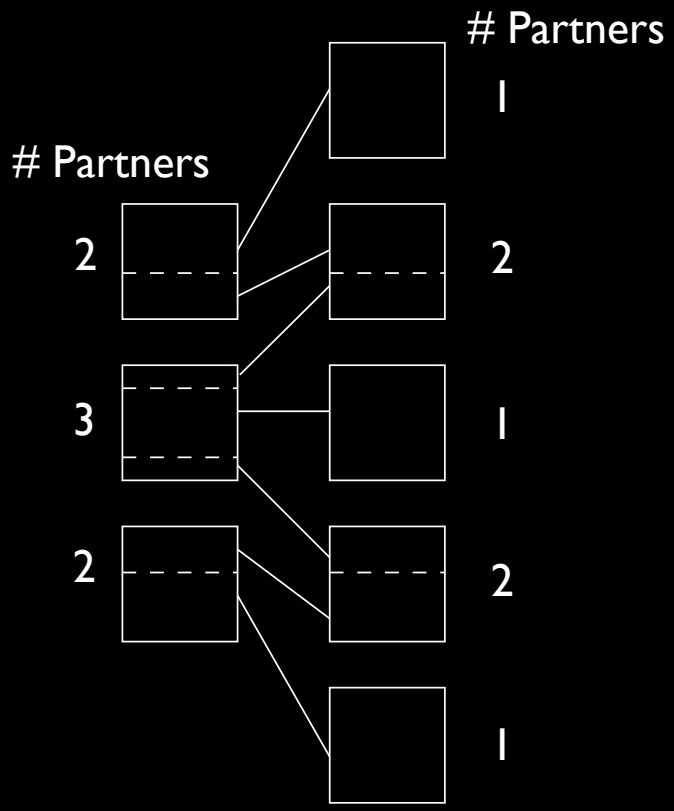


Three abstract communicators—producer, dataflow, and consumer—are used to couple producer to consumer. The dataflow can be a simple noop or a complete parallel program performing complex data transformations.

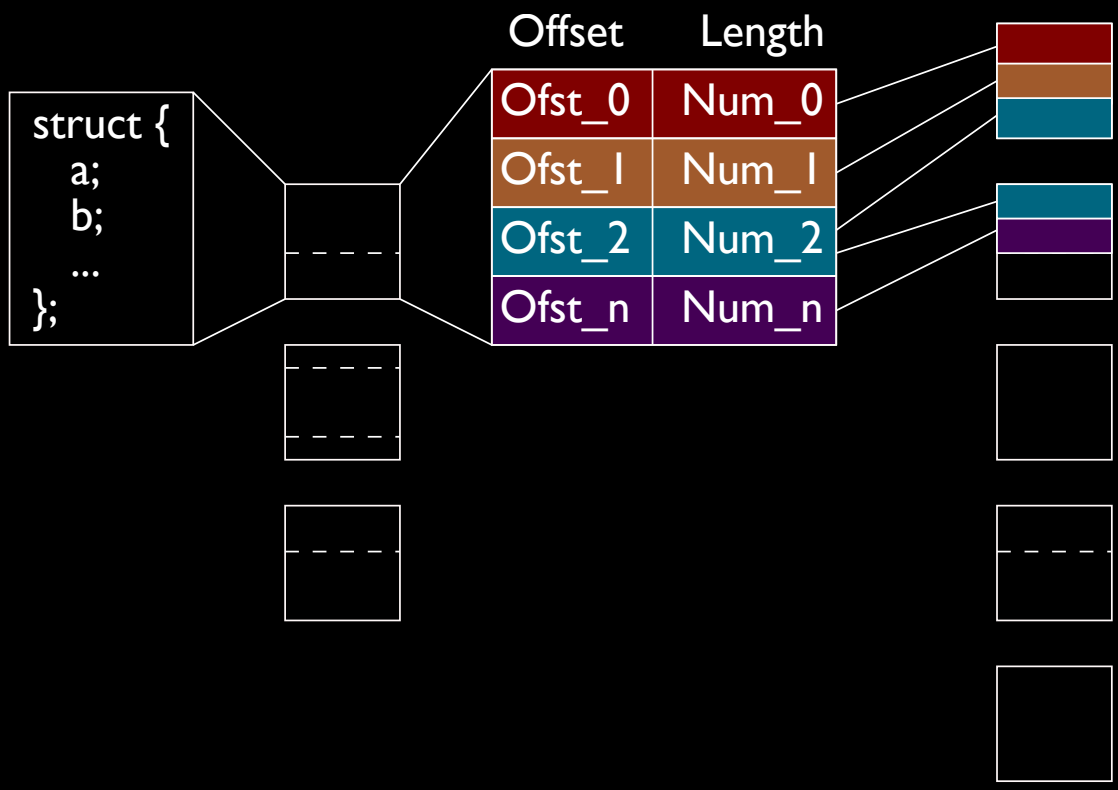


Different amounts of overlap between producer, dataflow, and consumer communicators control time and space partitioning.

Data Distribution



The number of input and output ports for each node is a function of the number of producer, dataflow, and consumer nodes.

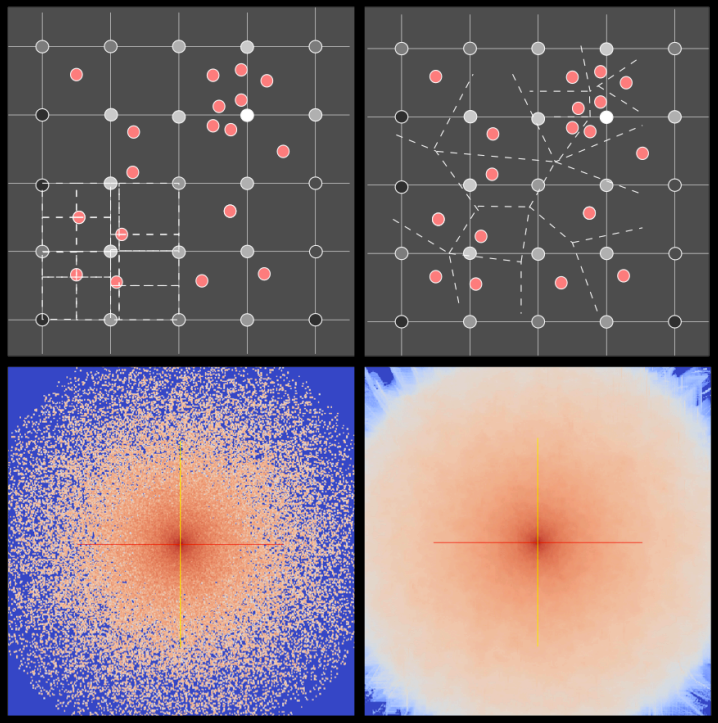


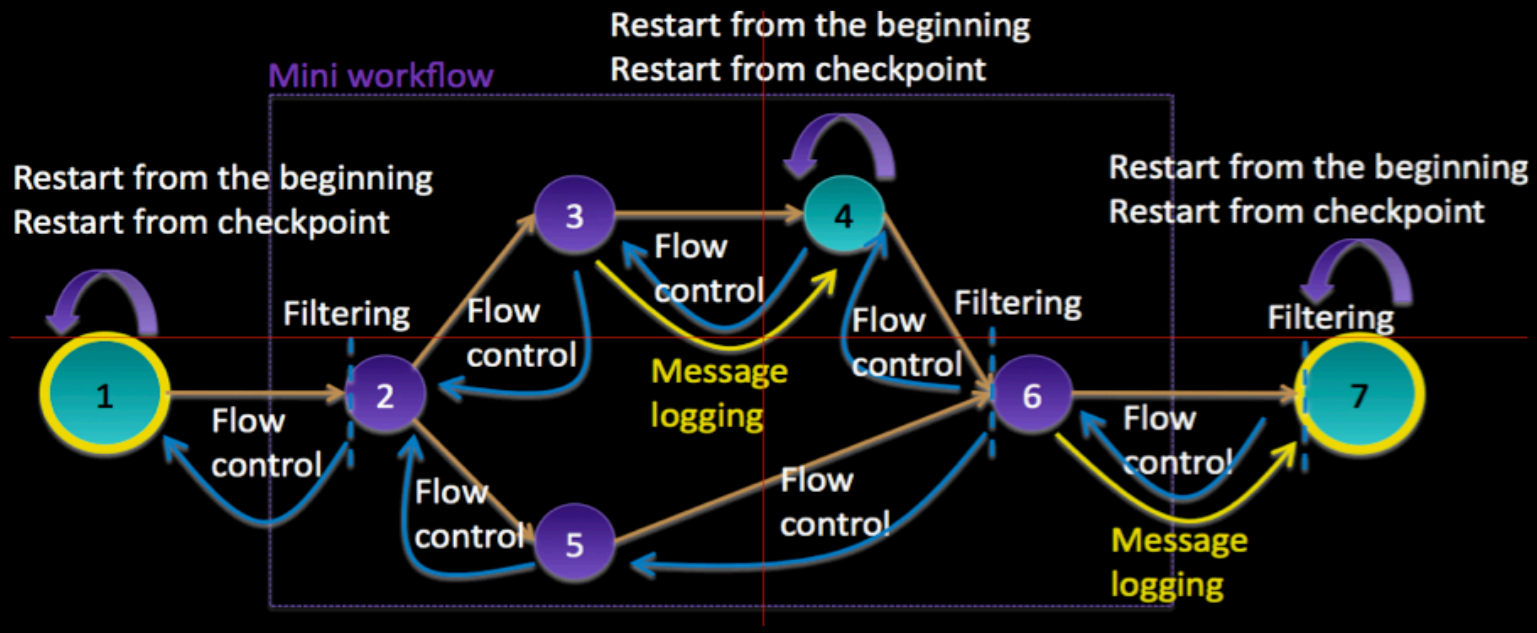
Data model subdivision is performed using the (offset, length) representation, segmenting at integer numbers of lengths such that the number of bytes approximates the fraction of the source data that will be sent to the destination.

Resilience to Faults

One of our resilience efforts attempts to detect silent data corruption by validating analysis tasks with an auxiliary method, usually less expensive and less accurate, but hopefully good enough to detect soft errors.

Another research topic is modeling the dataflow and optimally adding replication and roll back mechanisms to recover from hard (fail stop) errors and soft errors detected above.





Related Work

	Flexible Communicators	M:N redistribution	Generic Datatypes	Complex Permutations	Pipelining	Automatic Buffering	Fault Tolerance
EV Path	✔	✔	✔	✔		✔	
Damaris	✔		✔			✔	
Flow VR	✔	✔		✔			
Glean	✔	✔		✔	✔	✔	
Catalyst	✔	✔	✔				
Decaf	✔	✔	✔	✔	✔	✔	✔

The above table summarizes the state of the art by describing various tools along different dimensions with respect to the capability needed for Decaf. A dark check mark indicates that the tool has all the capability that we need in Decaf. A light check mark indicates less than complete coverage compared with our projected need.

Wrapping Up

We illustrated two specific analysis tasks and then previewed a new project to couple such tasks together with simulations.

Knowns

- Good scalability for individual data analysis tasks (intracode data movement)
- A design for an intercode data movement layer featuring:
  - A separate scalable dataflow between producer and consumer
  - Automatic buffering
  - Data redistribution and pipelining
  - Resilience to faults

Unknowns

- Synergy with existing coupling tools and transport layers
- High-level interface: workflow representation and execution
- Data model representation

U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

"The purpose of computing is insight, not numbers."

–Richard Hamming, 1962

Acknowledgments:

Facilities

Argonne Leadership Computing Facility (ALCF)

National Energy Research Scientific Computing Center (NERSC)

Funding

US DOE SDMAV2 Exascale Research

People

Jay Lofstead, Patrick Widener, Franck Cappello, Florin Isaila, Lokman Rahmani, Hadrien Croubois, Guillaume Aupy, Dmitriy Morozov, Carolyn Phillips, Adrian Pope, Hal Finkel, Katrin Heitmann, Salman Habib, Steve Rangel, Nan Lee

JLESC 14

Nov. 24, 2014

Tom Peterka

tpeterka@mcs.anl.gov