

Using Workflows to Enable Flexible VAP Development

Terence Critchlow

Matt Macduff

Justin Almquist

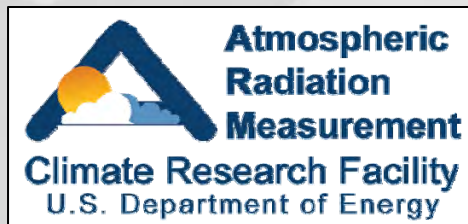
Ian Gorton

Tim Shippert

George Chin

Todd Halter

Chandrika Sivaramakrishnan



What is a workflow?

- ▶ Workflows consist of
 - Series of discrete activities (actors)
 - Flow of information between actors
 - Orchestration of interaction
- ▶ Workflows allow automation of tasks
 - Workflow engines provide a framework for defining and executing workflows
 - Relieves scientists from much of the ongoing data management overhead

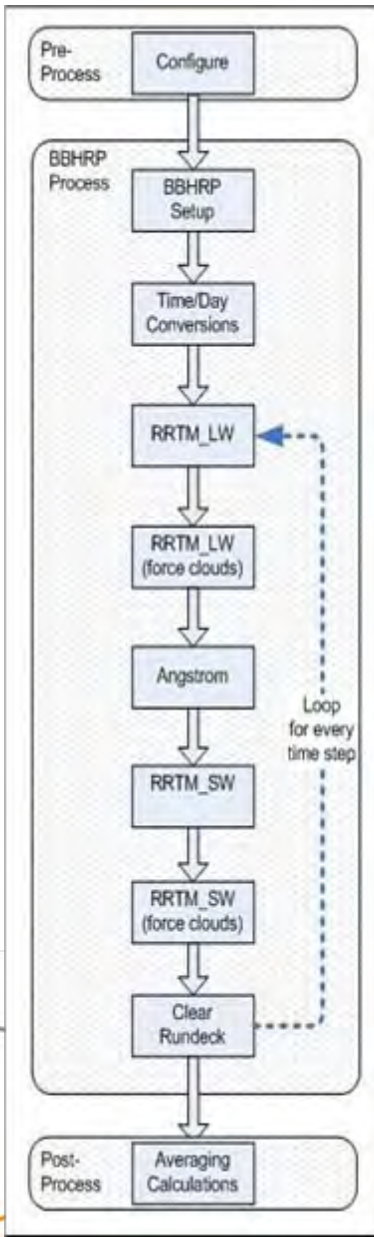
Why should I care?

Aren't scripts good enough?

- ▶ Workflow environments provide:
 - Higher level of abstraction, increasing code reuse
 - Drag-and-drop programming
 - Inherent actor-level and workflow-level parallelism
 - Provenance recording
- ▶ While scripts can do all of this, a workflow environment makes it much easier

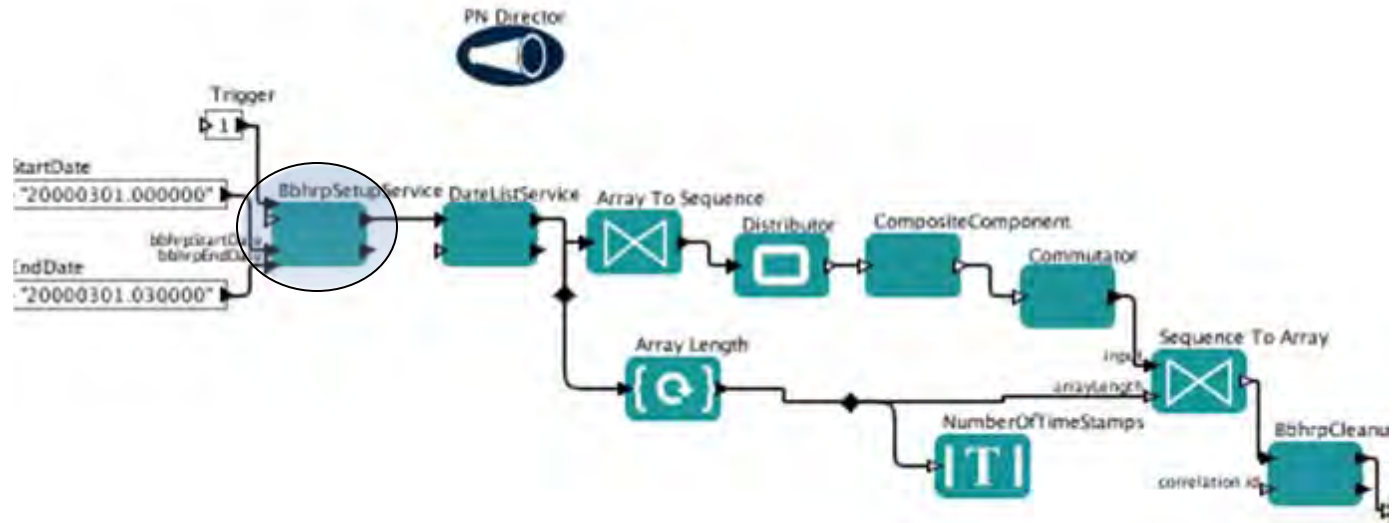
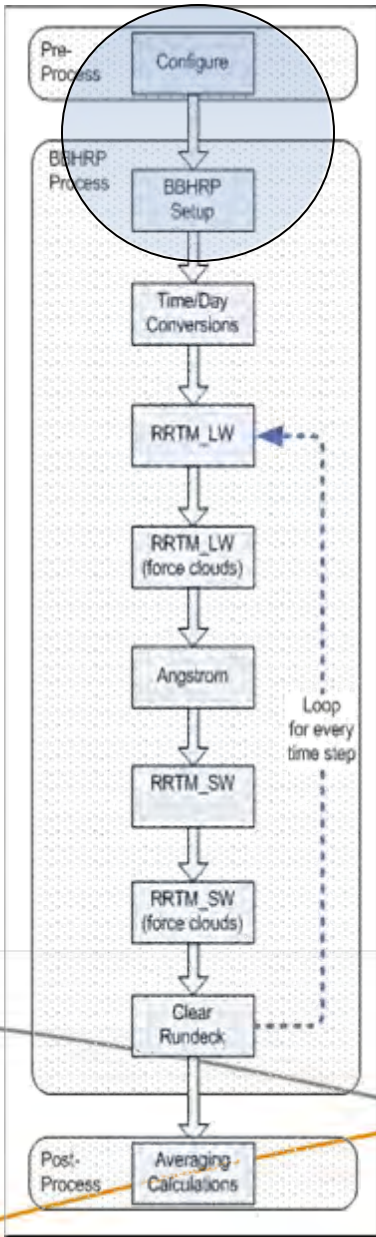
Prototyping an existing VAP using workflow technology

- ▶ Setup environment
 - Templates
 - generate input files
 - Create working directories
 - Logging
- ▶ Run core physics
 - Run various physics routines using different assumptions to gain insight into science
- ▶ Postprocessing
 - Cleanup environment
 - Perform final data transformations



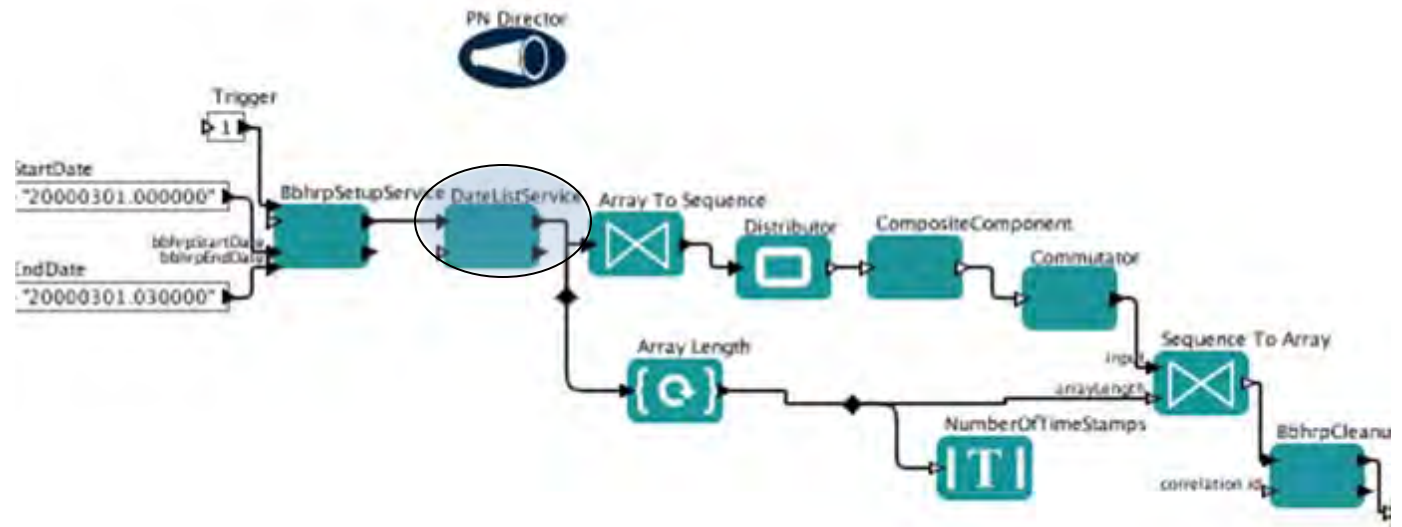
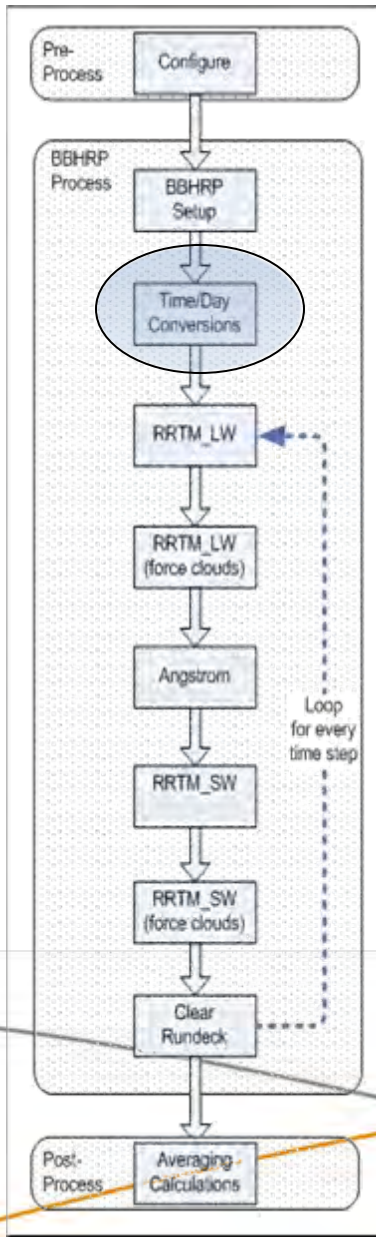
What does a workflow look like?

- There is a very close correlation between the conceptual flow and the executable workflow



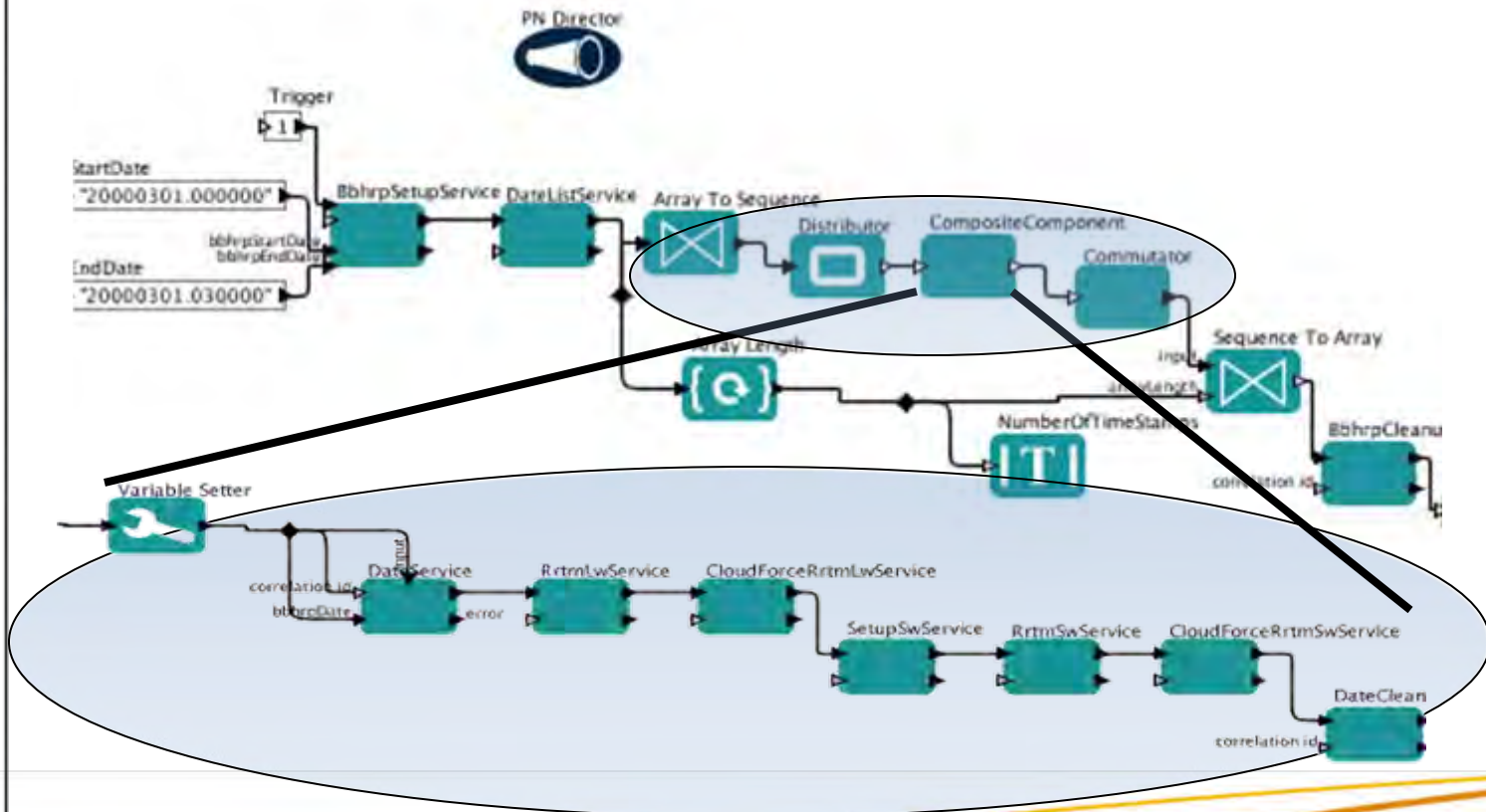
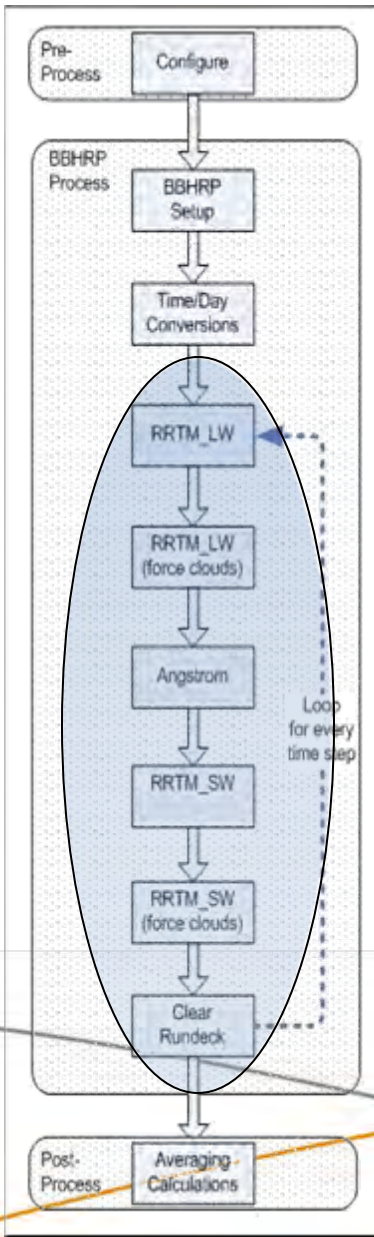
What does a workflow look like?

- There is a very close correlation between the conceptual flow and the executable workflow



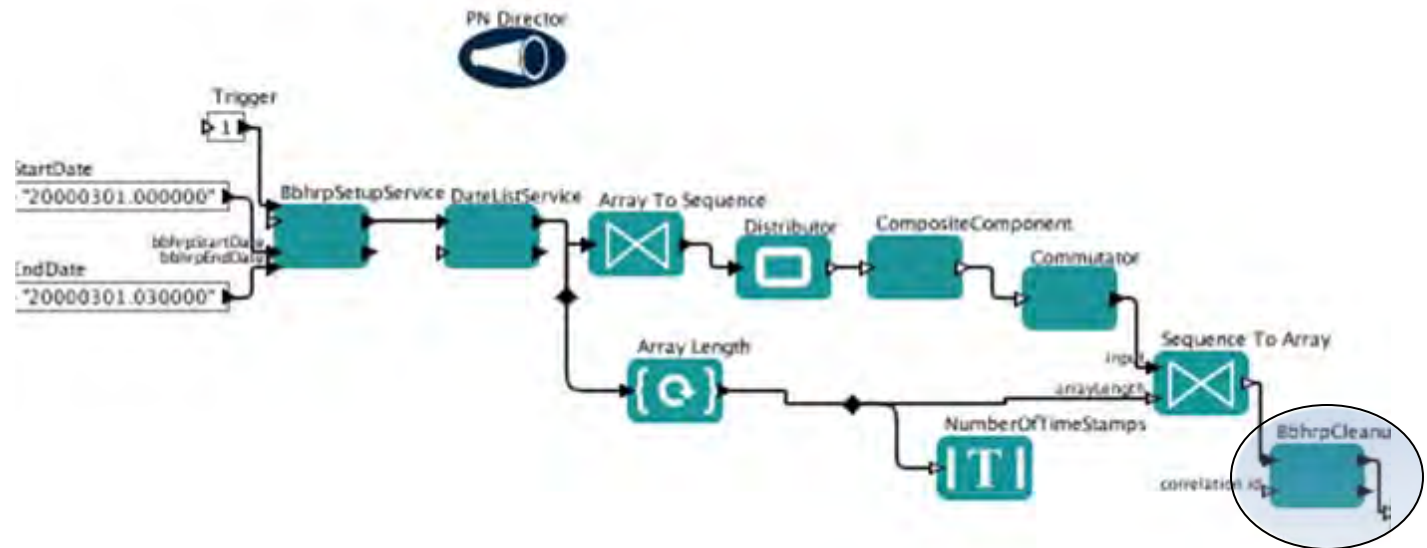
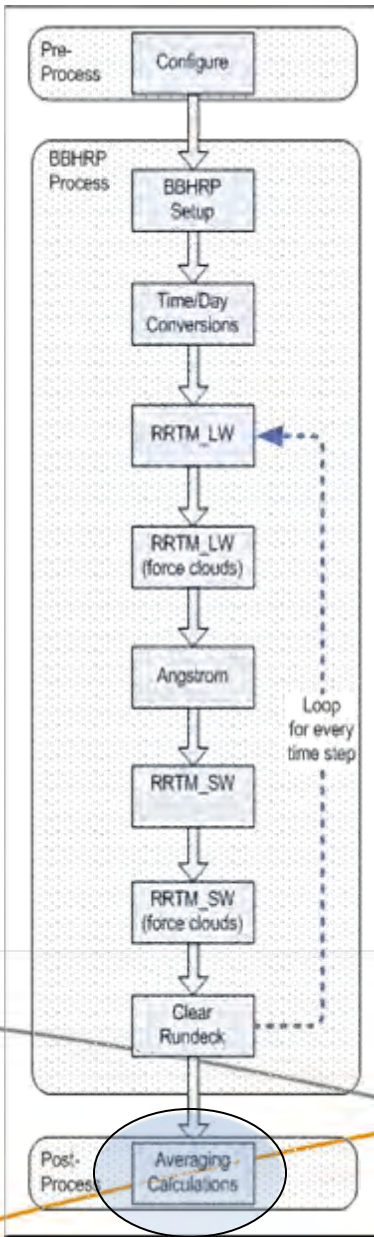
What does a workflow look like?

- There is a very close correlation between the conceptual flow and the executable workflow

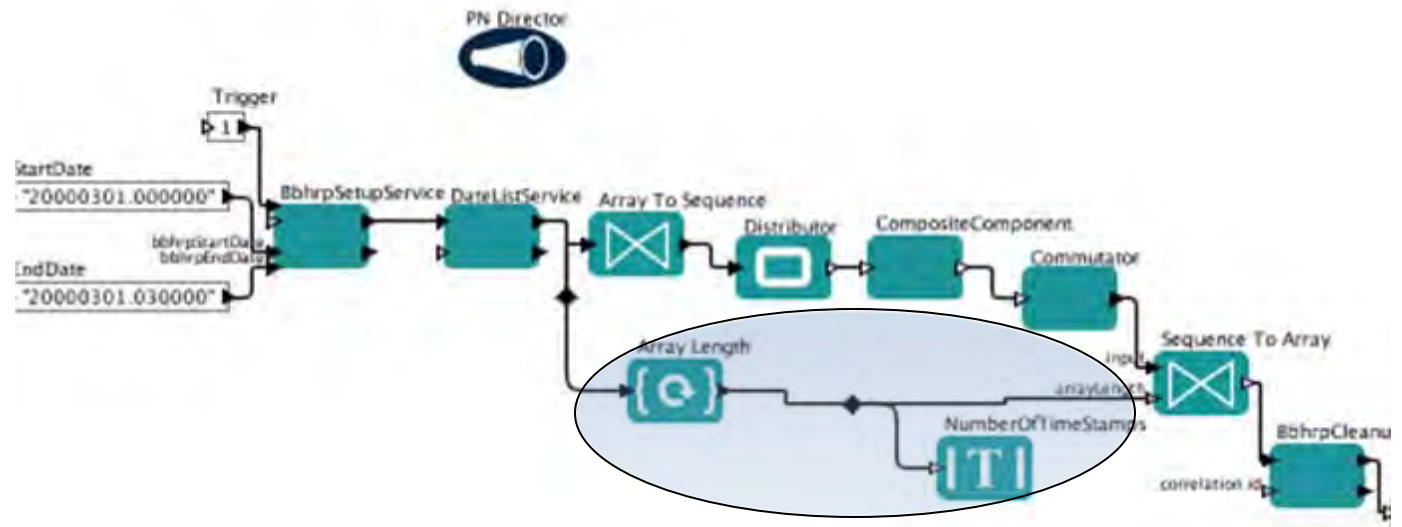
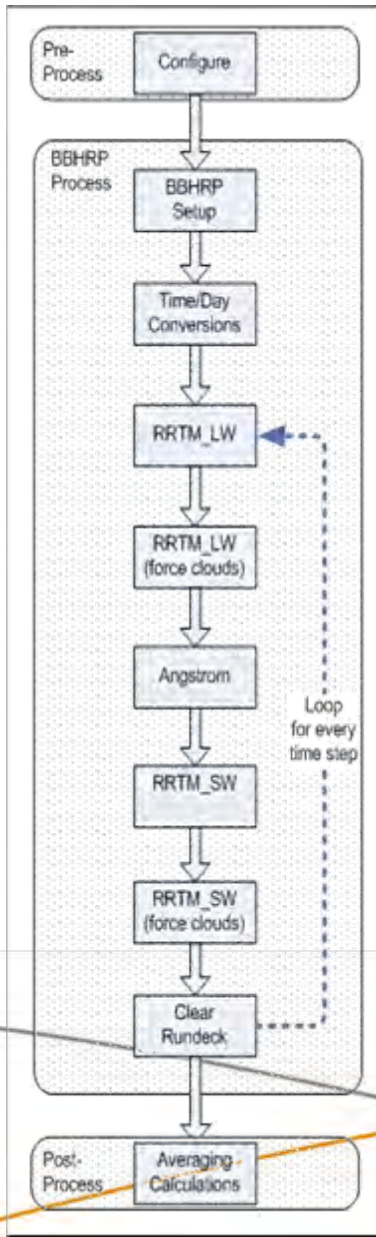


What does a workflow look like?

- There is a very close correlation between the conceptual flow and the executable workflow

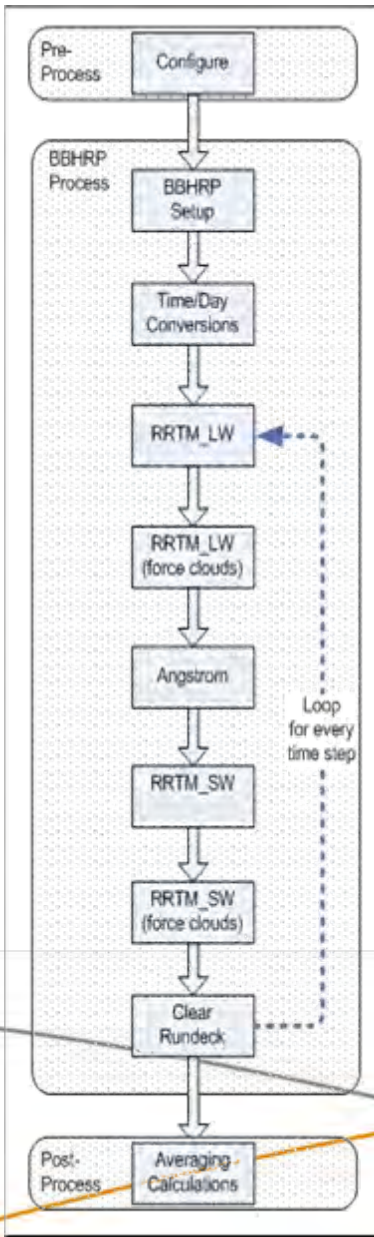


What does a workflow look like?

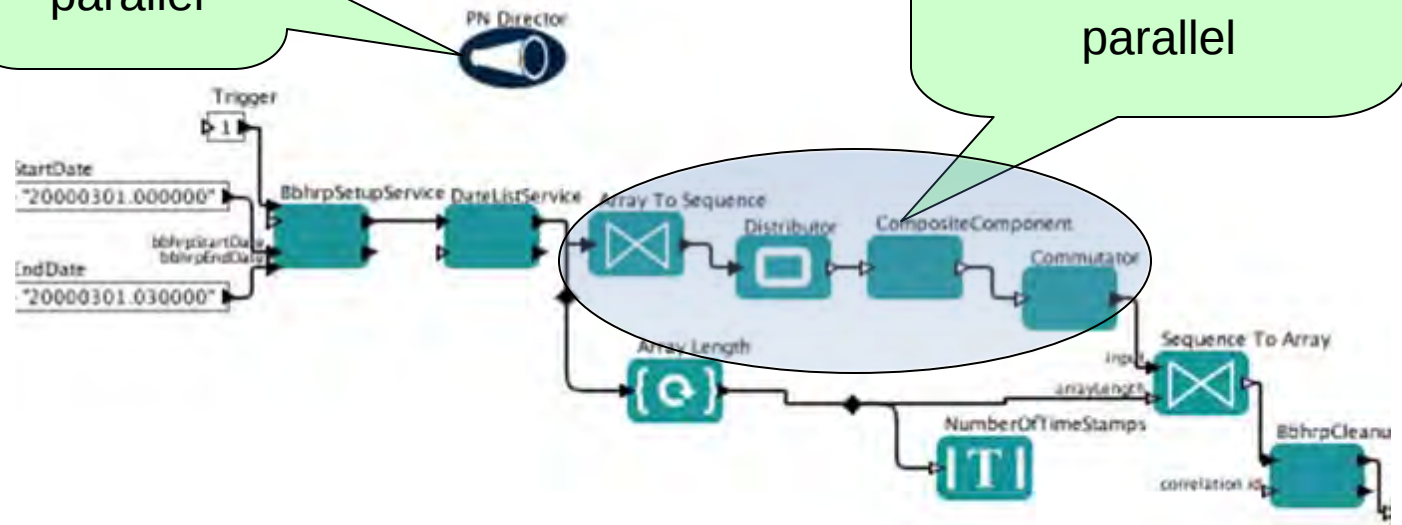


- ▶ There is a little additional bookkeeping to monitor the progress of the workflow
- ▶ A production workflow would be more fault-tolerant

What does a workflow look like?



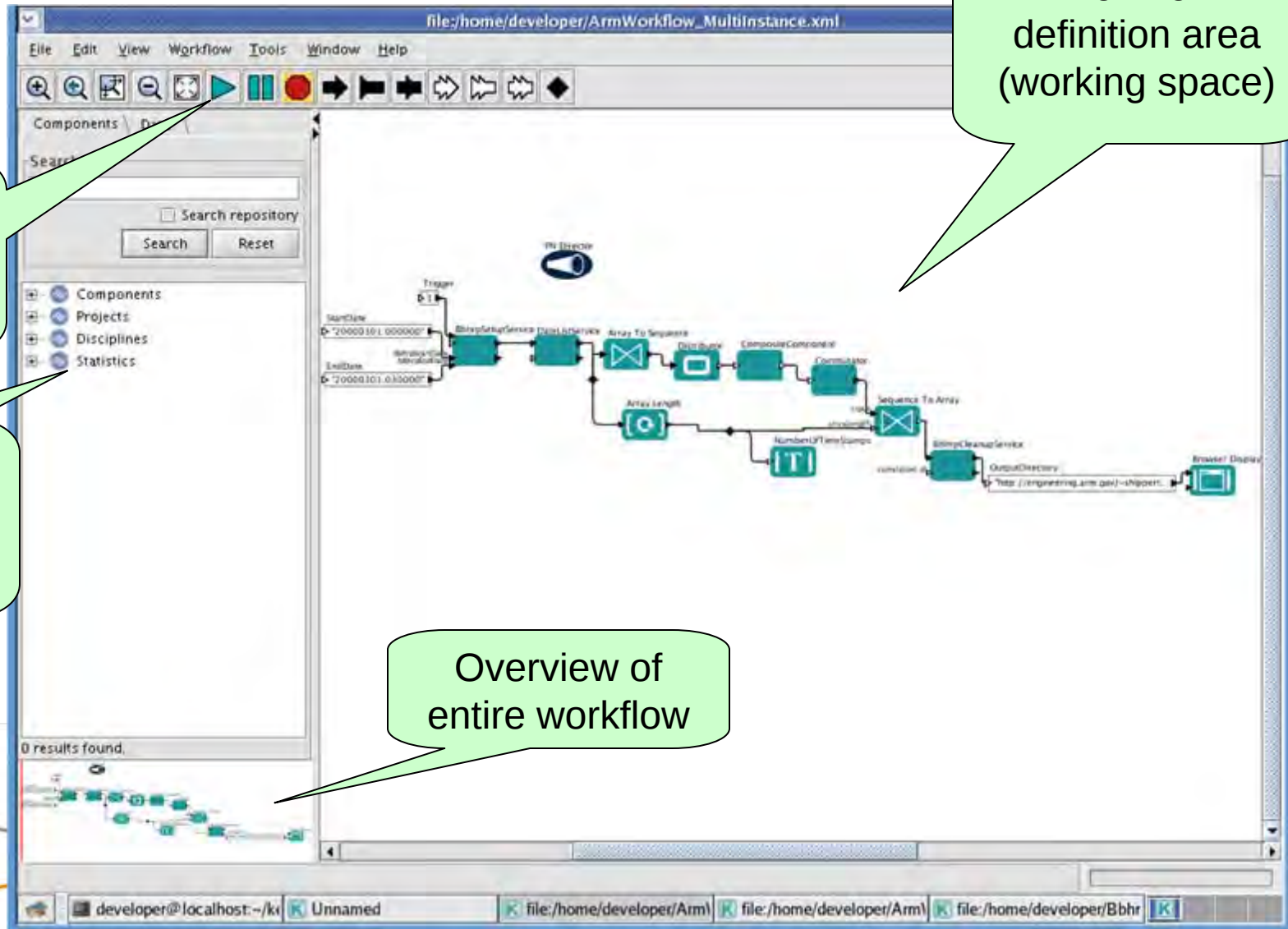
Run each actor in parallel



Executes a user-defined number of subworkflows in parallel

- ▶ Parallelism is defined during workflow construction by using existing constructs
- ▶ Threads are automatically managed at run time

Workflows are defined and executed within a workflow engine such as Kepler



Workflows are defined and executed within a workflow engine such as Kepler

The screenshot displays a workflow engine interface with two main windows. The top window, titled 'Index of /~shippert/BBHRP/medici_demo/sgprtmlwCl.c1 - Mozilla Firefox', shows a table of data. The bottom window, titled 'Index of /~shippert/BBHRP/medici_demo/sgprtmlwCl.c1 - Mozilla Firefox', shows a list of files and their sizes.

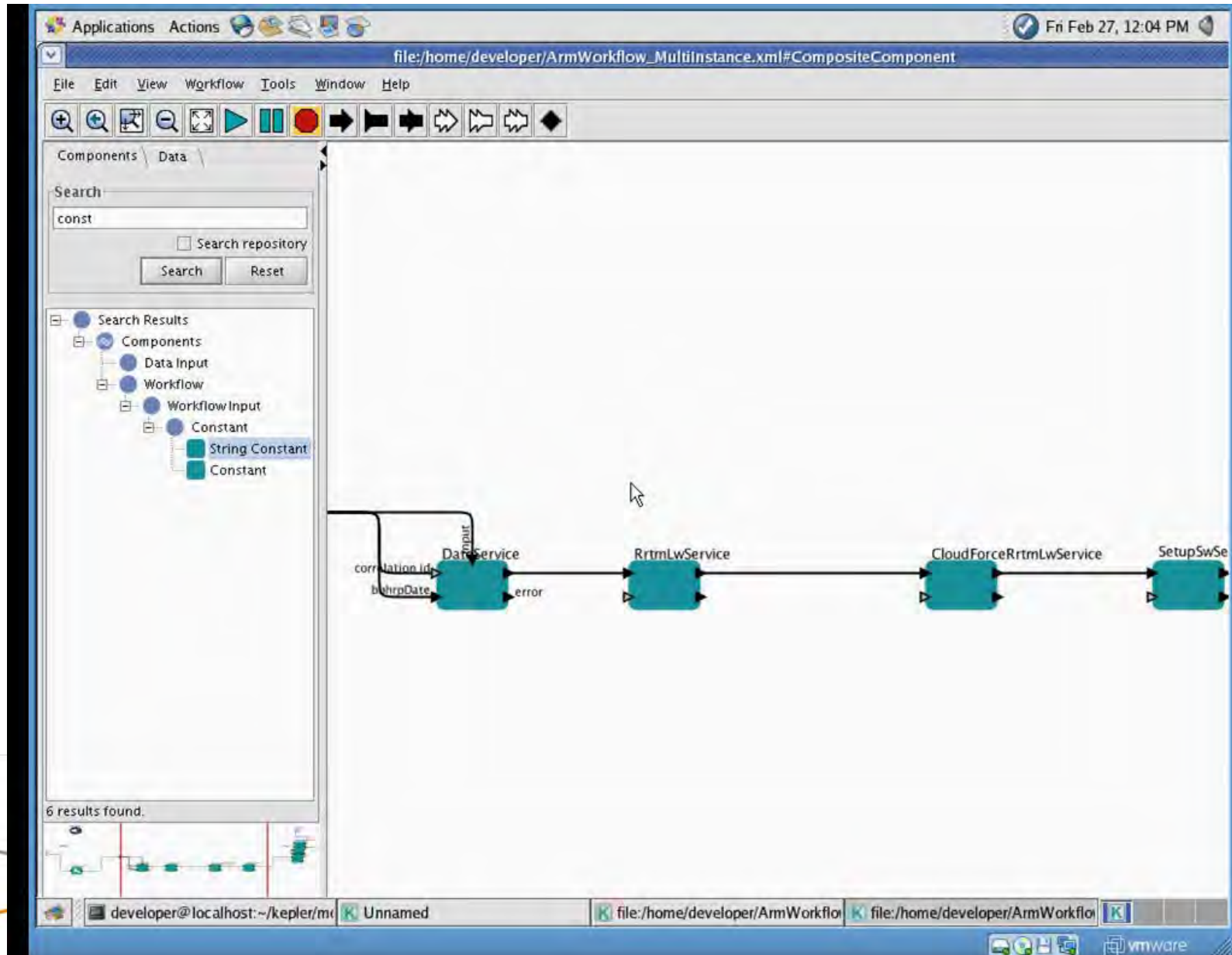
Table Data:

LEVEL	Wavenumbers: 10.0 - 3250.0 cm ⁻¹ , ATM 1	PRESSURE mb	UPWARD FLUX W/m ²	DOWNWARD FLUX W/m ²	NET FLUX W/m ²	HEATING RATE degree/day
54		.07041	269.2938	0.0000	269.2938	0.00000
53		0.1257	269.3241	0.1055	269.2185	-11.48810
52		0.2190	269.3658	0.2188	269.1470	-6.46711
51		0.2867	269.3895	0.2967	269.0928	-6.76211
50		0.3732	269.4118	0.3961	269.0157	-7.51343
49		0.4829	269.4276	0.5251	268.9025	-8.72059
48		0.6214	269.4261	0.6935	268.7326	-10.34712
47		0.7979	269.3947	0.9052	268.4895	-11.63115
46		1.024	269.3181	1.1610	268.1571	-12.39977
45		1.314	269.1902	1.4419	267.7483	-11.87663
44		1.695	269.0324	1.7330	267.2995	-9.95688
43		2.198	268.8564	2.0348	266.8217	-9.01622
42		2.871	268.6588	2.3457	266.3131	-6.37578
41		3.859	268.4397	2.6678	265.7719	-4.62460
40		5.055	268.2152	3.0320	265.1833	-4.15337
39		6.534	267.9980	3.4610	264.5370	-3.68637
38		8.658	267.8183	3.9810	263.8373	-2.77987
37		11.97	267.6931	4.6709	263.0222	-2.07720
36		15.02	267.6148	5.3299	262.2849	-2.04312
35		18.80	267.5467	6.1177	261.4290	-1.90885
34		23.61	267.4888	7.0436	260.4452	-1.72418
33		29.72	267.4402	8.1054	259.3348	-1.53483
32		37.45	267.3413	9.2938	258.0475	-1.40462

File List:

Name	Size
Parent Directory	10-Sep-
clear.sgprtmlwCl.c1.20000301.150330.out	06-Nov-
sgprtmlwCl.c1..out	26-Jan-
sgprtmlwCl.c1.1234.out	29-Oct-
sgprtmlwCl.c1.20000301.002630.out	24-Feb-2009 02:45 78k
sgprtmlwCl.c1.20000301.002730.out	24-Feb-2009 02:17 78k
sgprtmlwCl.c1.20000301.005130.out	26-Feb-2009 22:26 78k
sgprtmlwCl.c1.20000301.011530.out	26-Feb-2009 22:24 78k
sgprtmlwCl.c1.20000301.024330.out	26-Feb-2009 22:26 78k
sgprtmlwCl.c1.20000301.030730.out	07-Feb-2009 01:17 78k
sgprtmlwCl.c1.20000301.043530.out	07-Feb-2009 01:17 78k
sgprtmlwCl.c1.20000301.062730.out	07-Feb-2009 01:17 78k
sgprtmlwCl.c1.20000301.075530.out	06-Feb-2009 22:42 78k
sgprtmlwCl.c1.20000301.081930.out	06-Feb-2009 22:42 78k

What if I want to change something?



Where could ARM go with this?

- ▶ New actors could wrap
 - Data transformation capabilities
 - Science codes
 - Visualization routines
 - Data transfers (e.g. back to desktop)
- ▶ With a small number of customized actors, scientists could easily customize VAPS without significant developer involvement
- ▶ This would improve scientists' ability to directly interact with and understand the underlying data significantly advancing science

Conclusions

- ▶ Workflows provide an environment in which scientists can easily extend and adapt existing processes by sharing and reusing components
- ▶ By incorporating workflows, complex procedures such as VAPs can provide improved flexibility and usability to scientists
- ▶ The resulting ability to customize analysis based on individual goals can significantly improve scientists' ability to gain insight from data

Questions

- ▶ Poster session today
 - A Prototype for VAP Development Using Workflow Technology (Matt MacDuff)
- ▶ Contact
 - Terence Critchlow
 - Terence.Critchlow@pnl.gov
 - 509-375-2134
- ▶ Web sites:
 - <http://dicomputing.pnl.gov/>
 - <http://kepler-project.org/>



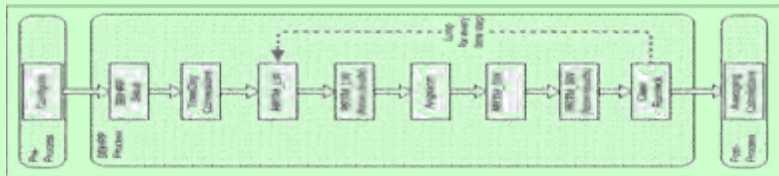
There is no magic here

Kepler Workflow Engine



Medici Infrastructure Framework

Science Codes



- ▶ Contains existing science codes, data transformation scripts
- ▶ Codes will continue to be used as they currently are
- ▶ Large codes may be broken into smaller components for easier reuse
- ▶ If code does not already exist to do everything, then some will need to be written

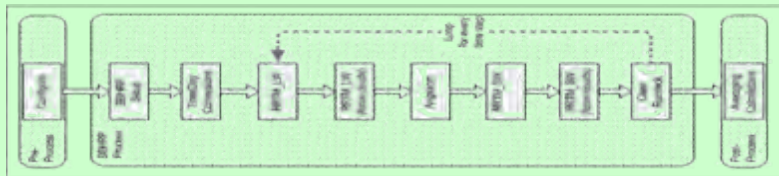
There is no magic here

Kepler Workflow Engine



Medici Infrastructure Framework

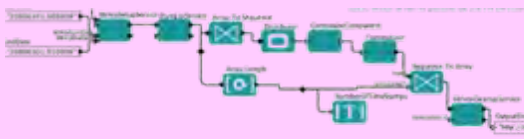
Science Codes



- ▶ Wraps underlying components into *Web Services*
- ▶ Presents a web-based view of these components and allows them to be executed from external machines
- ▶ Hides details like machine name and login information from scientists

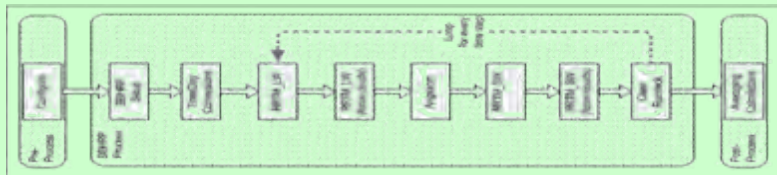
There is no magic here

Kepler Workflow Engine



Medici Infrastructure Framework

Science Codes



- ▶ Defines interaction between actors
- ▶ Coordinates actor execution and data flow
- ▶ Hides many underlying infrastructure details
- ▶ Provides access to both ARM-centric and externally developed capabilities