
Parallelization techniques: Applying Map, Reduce and Cross concepts using bioActors



Ilkay ALTINTAS, Ph.D.

Deputy Coordinator for Research, San Diego Supercomputer Center, UCSD

Lab Director, Scientific Workflow Automation Technologies

altintas@sdsc.edu

SDSC



bioKepler.org

What is Parallelization?

SDSC


UC San Diego



bioKepler - September, 2012

bioKepler.org

What is Parallelization?

SDSC


UC San Diego



bioKepler - September, 2012

bioKepler.org

Distributed Computing Environments

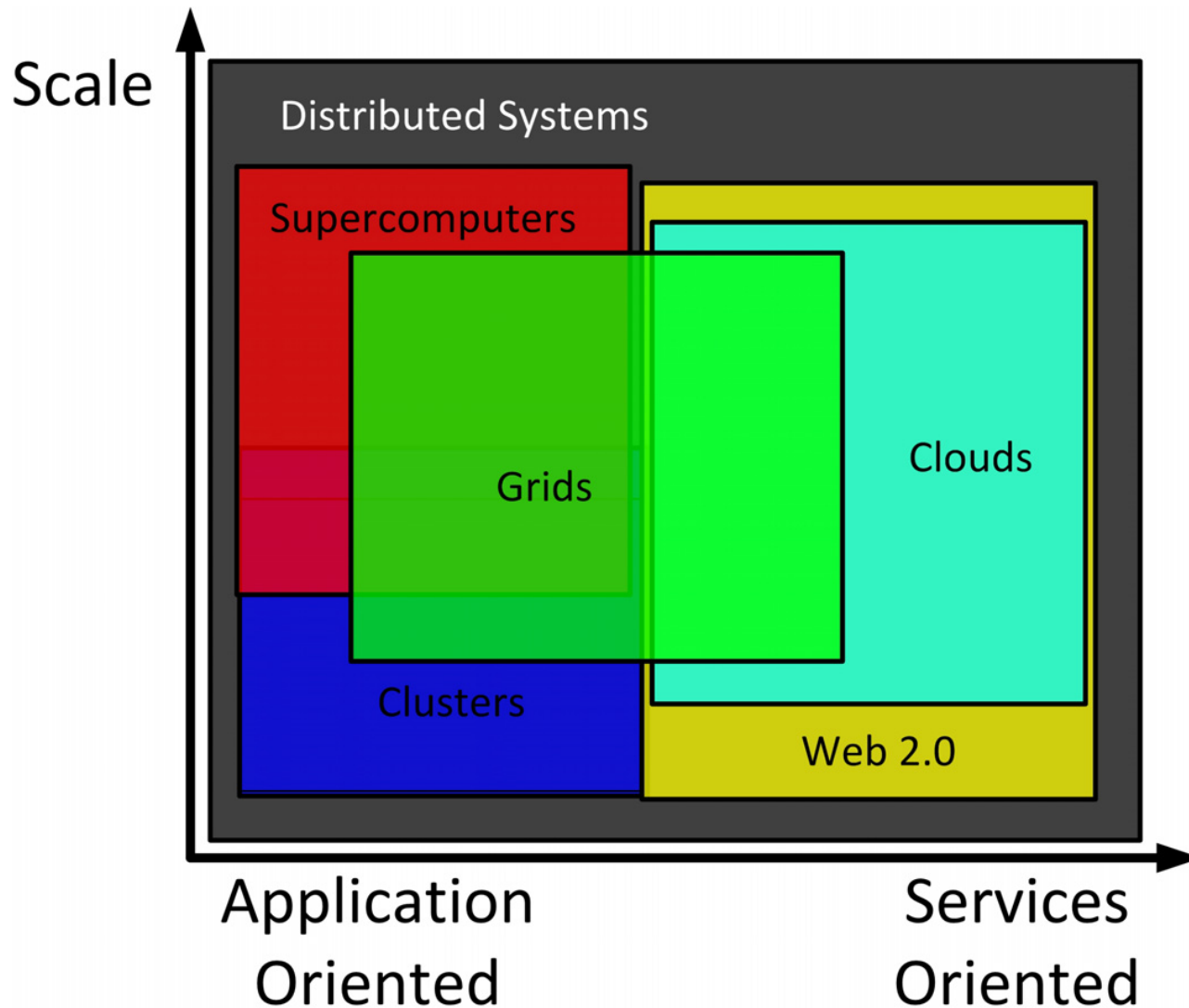


Figure 1 FROM:
"Cloud Computing and Grid Computing 360-Degree Compared",
Ian Foster, Yong Zhao, Ioan Raicu,
Shiyong Lu. Grid Computing
Environments Workshop (GCE), 2008.

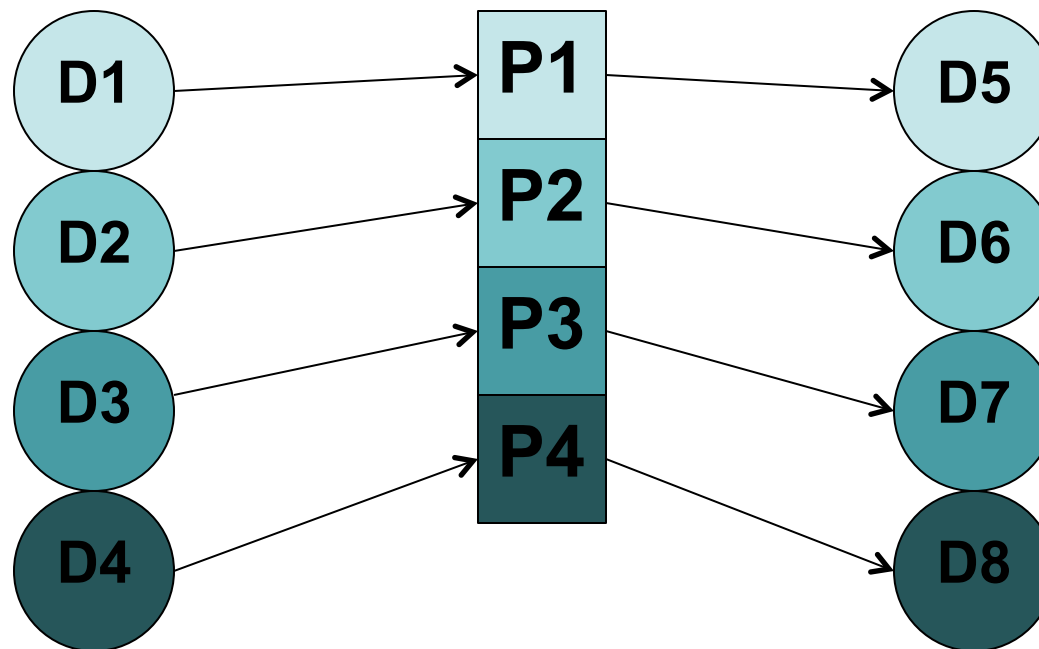
Figure 1: Grids and Clouds Overview

Parallelization Solutions in Distributed Environments

- *Traditional parallel programming interfaces*
 - Examples: MPI and OpenMP
 - Hard to implement
 - Original sequential tools cannot be reused
- *Parallel job execution*
 - Examples: SGE and Condor
 - Original sequential tools can be reused
 - Create small jobs by splitting data or tasks
 - Hard to achieve data locality for each job
- *Data parallel job execution*
 - Examples: Hadoop and Stratosphere
 - Original sequential tools can be reused
 - Support customized and automatic data partition and distribution
 - Support data locality for each job through special distributed file system, HDFS

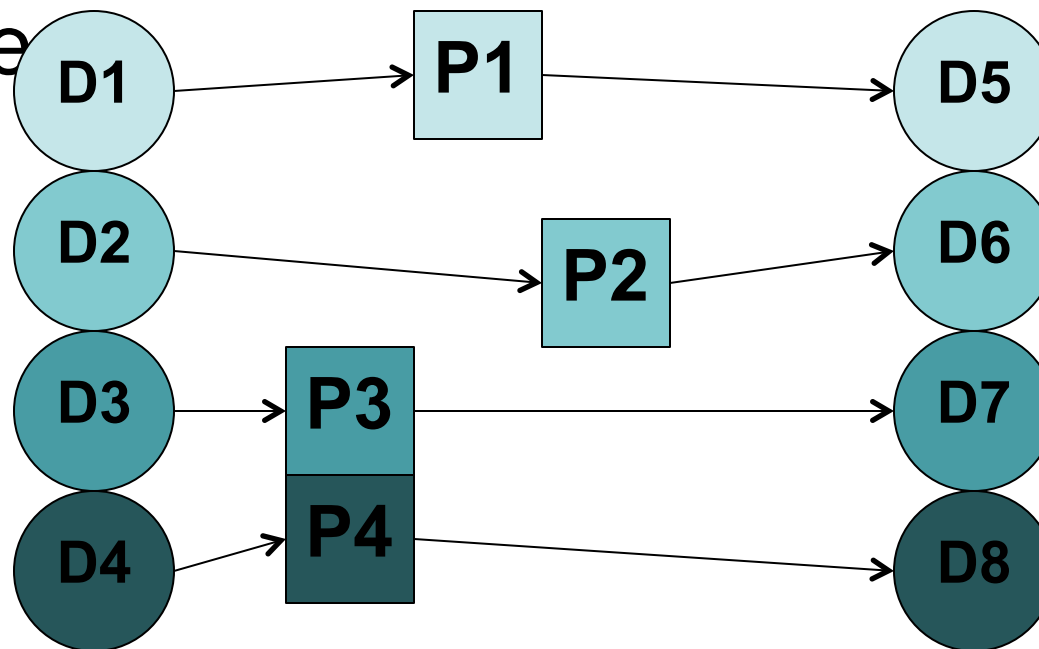
Data Parallel Task Execution

- Static executables run as processes
- Independent data items are assigned to processes



Distributed Data Parallel (DDP) Task Execution

- Static executables run as processes on distributed environments
- Independent data items are assigned to processes



MapReduce:

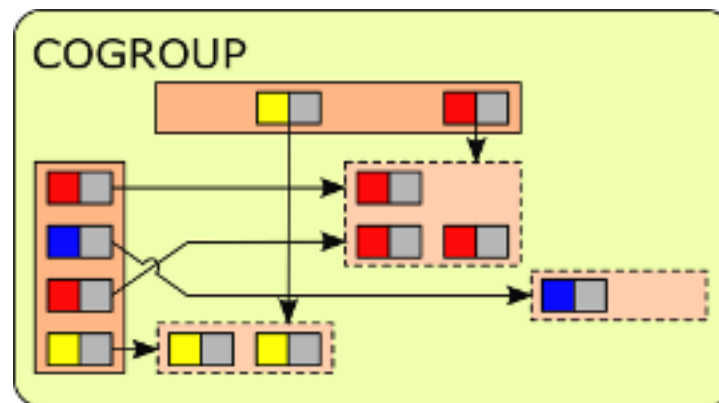
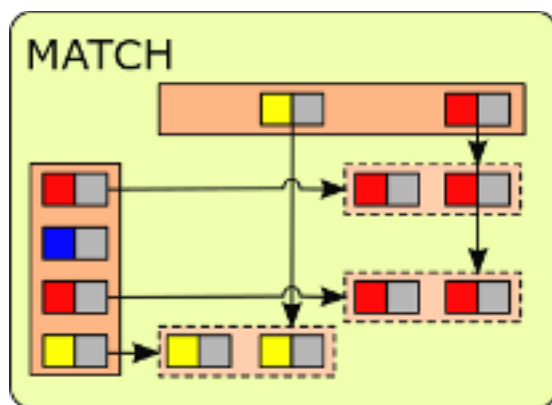
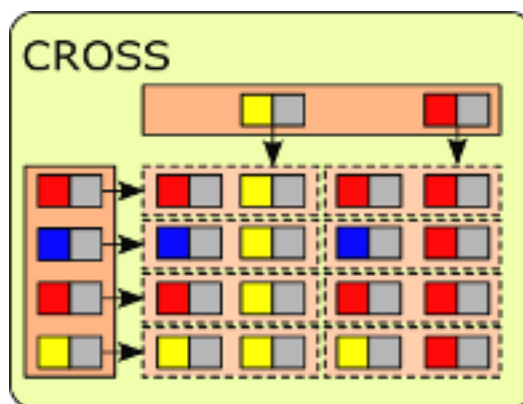
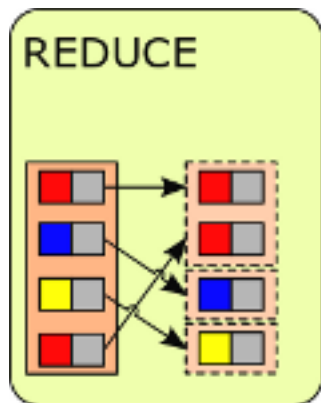
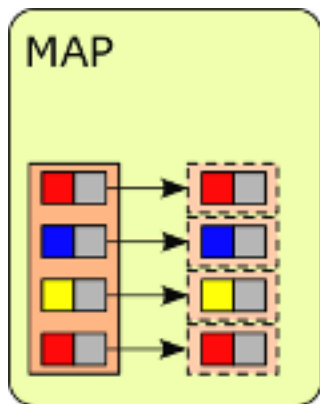
A Typical DDP Execution Pattern

- Chop the data based on a feature of interest
(**value**) (key)
- Iterate a function on each value
- Order the intermediate data products'
(**intermediate value**)
- Stitch the intermediate values

- Can execute using a specialized engine
Examples: Hadoop and Nephela

Many Other DDP Patterns

Images taken from:
<http://www.stratosphere.eu>



Distributed Data-Parallel bioActors

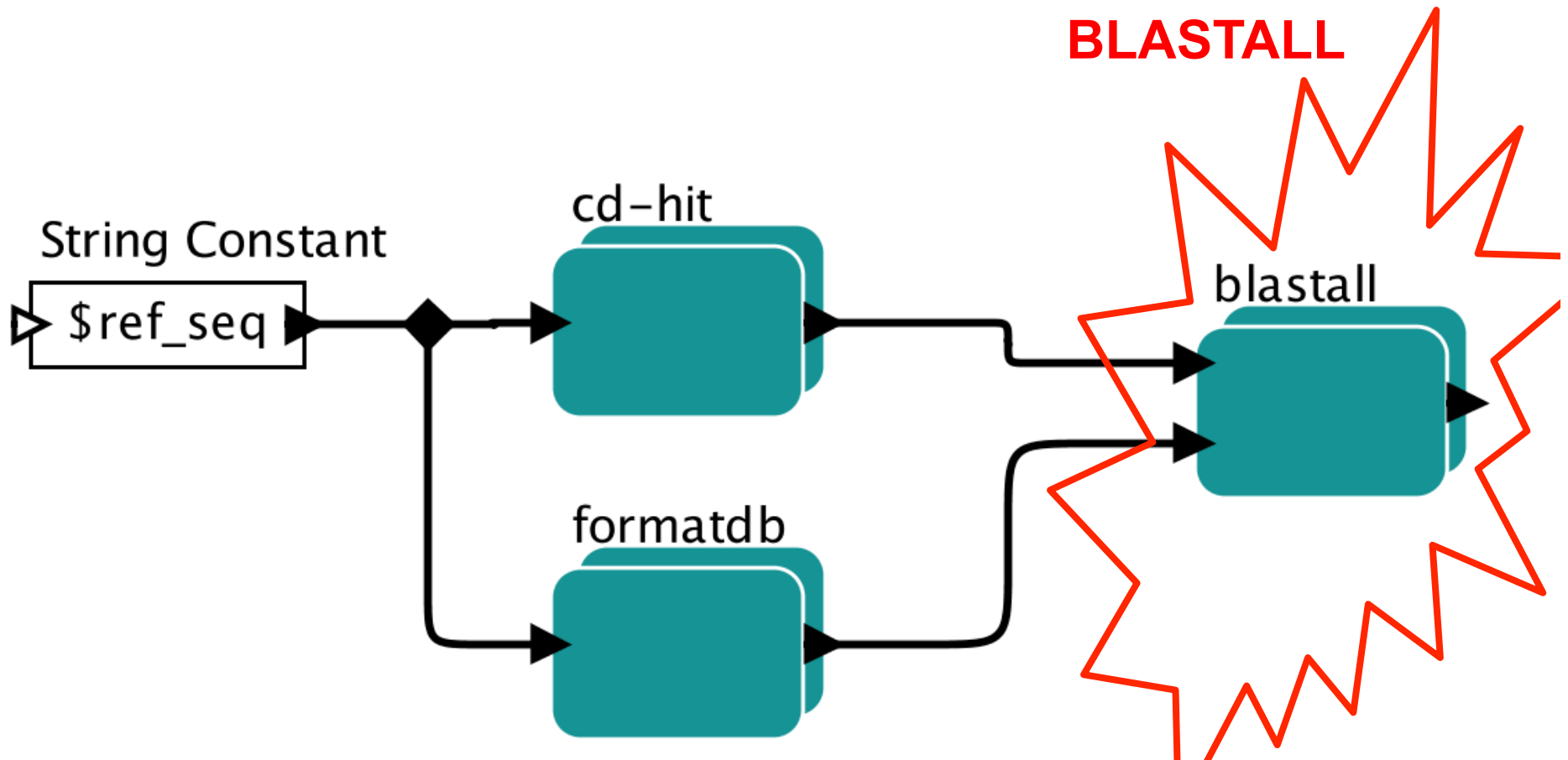
- Set of steps to execute a bioinformatics tool in DDP environment
- Customized from the [ExecutionChoice](#) actor
- Includes:
 - Data-parallel patterns, e.g., Map, Reduce, Cross, All-Pairs, etc., to specify data grouping
 - I/O to interface with storage
 - Data format specifying how to split and join

A Workflow with Three bioActors

SDF Director



● ref_seq: small.faa



Configuring the *BLASTALL* bioActor

.BLAST-bioActor-with-SGE-MapOnly-MapReduce-CrossReduce.blastall

Kepler

Local Execution Sun Grid Engine MapOnly MapReduce CrossReduce

blastall

program: blastall

programOptions: -p blastn -d \$ref_seq -i \$input -b1 -v1 -m8 -e \$max_evalue

Inputs

input: /Users/jianwu/Projects/bioinfo/testQueryFile/testQuery.fa

Outputs

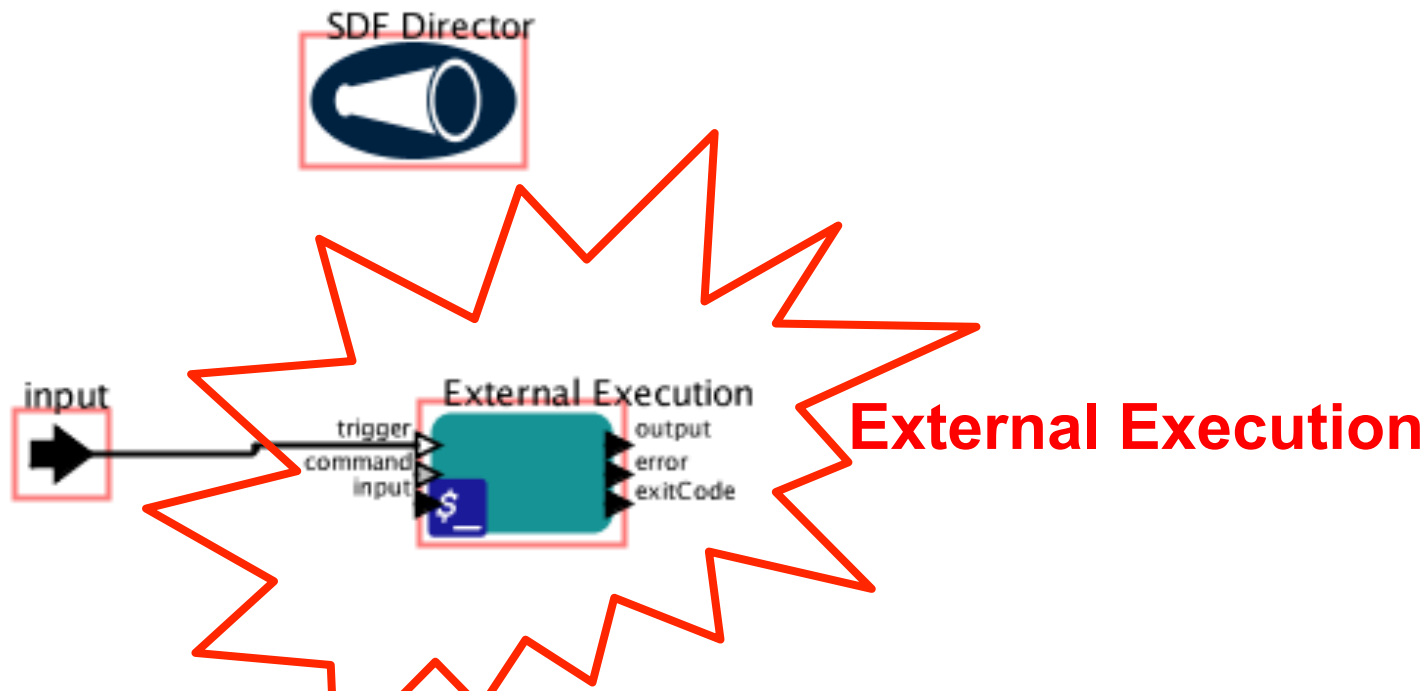
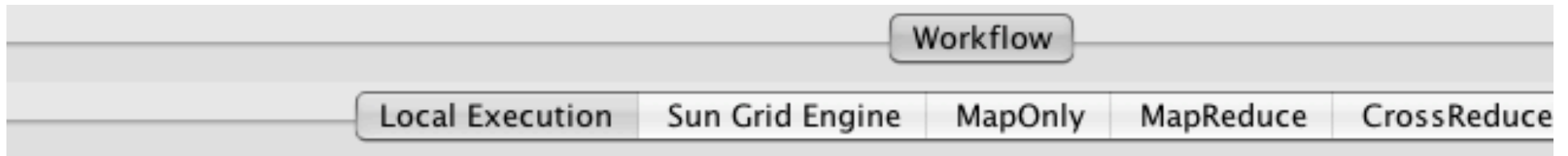
outputOption: -o

output: \$output_dir/test.out

Execution Choice: CrossReduce

Add Remove Cancel OK

Inside the LocalExecution Tab




Inside the MapReduce Tab

Workflow

Local Execution Sun Grid Engine MapOnly MapReduce CrossReduce

- ReferenceFile: \$DataDir/Ocean_Alaska.fa
- LocalRefFile: \$DataDir/Ocean_Alaska.fa
- DBFormatCmd: formatdb
- DBFormatOptions: -p F -o F
- DBPath: /tmp/
- DataDir: /Users/jianwu/Projects/bioinfo/testQueryFile/
- BLASTDir: /Users/jianwu/Projects/CAMERA/cvs/workflows/ca
- StratosphereConfPath:
- ParallelNumber: 2
- BLASTCmd: blastall
- BLASTOptions: -p blastn -i stdin -m 8 -e 1e-5
- QueryFile: \$DataDir/testQuery.fa
- RefSeqLength: 3311966
- AlignmentLimit: 250
- OutputFile: \$DataDir/blast-map-reduce.out

SDF Director



input trigger ReferenceFile2 \$ReferenceFile

Stratosphere Blast

OutputFile2 \$OutputFile

Stratosphere Blast

4

The diagram illustrates a workflow configuration for a Stratosphere Blast job. It features a central 'Stratosphere Blast' task, represented by a teal rounded rectangle, which is highlighted with a large, jagged red starburst. To the left, an 'SDF Director' icon (a blue megaphone) is shown. Below it, a flow diagram shows an 'input' arrow pointing to a 'ReferenceFile2' task, which contains the variable '\$ReferenceFile'. A 'trigger' arrow points to this task. From the 'ReferenceFile2' task, an arrow points to the 'Stratosphere Blast' task. Below the 'Stratosphere Blast' task, another arrow points to an 'OutputFile2' task, which contains the variable '\$OutputFile'. A 'trigger' arrow points to this task. The entire workflow is enclosed in a window titled 'Workflow' with tabs for 'Local Execution', 'Sun Grid Engine', 'MapOnly', 'MapReduce', and 'CrossReduce'. The 'MapReduce' tab is selected. The configuration parameters are listed in two columns, with red and blue bullet points. The 'Stratosphere Blast' task is highlighted with a large, jagged red starburst.


Inside the MapReduce Tab

Workflow

Local Execution Sun Grid Engine MapOnly MapReduce CrossReduce

- ReferenceFile: \$DataDir/Ocean_Alaska.fa
- LocalRefFile: \$DataDir/Ocean_Alaska.fa
- DBFormatCmd: formatdb
- DBFormatOptions: -p F -o F
- DBPath: /tmp/
- BLASTCmd: blastall
- BLASTOptions: -p blastn -i stdin -m 8 -e 1e-5
- QueryFile: \$DataDir/testQuery.fa
- OutputFile: \$DataDir/blast-map-reduce.out
- DataDir: /Users/jianwu/Projects/bioinfo/testQueryFile/
- BLASTDir: /Users/jianwu/Projects/CAMERA/cvs/workflows/ca
- StratosphereConfPath:
- ParallelNumber: 2
- RefSeqLength: 3311966
- AlignmentLimit: 250

SDF Director



input

trigger

ReferenceFile2

\$ReferenceFile

Stratosphere Blast

OutputFile2

\$OutputFile

trigger

The diagram illustrates a workflow within the SDF Director. It starts with an 'input' arrow pointing to a 'ReferenceFile2' box containing '\$ReferenceFile'. A 'trigger' arrow points to the 'ReferenceFile2' box. From the 'ReferenceFile2' box, an arrow points to a 'Stratosphere Blast' box. Below the 'ReferenceFile2' box is an 'OutputFile2' box containing '\$OutputFile'. A 'trigger' arrow points to the 'OutputFile2' box. An arrow points from the 'OutputFile2' box to the 'Stratosphere Blast' box. A diamond-shaped connector is located on the arrow between the 'OutputFile2' box and the 'Stratosphere Blast' box.

BLASTALL with MapReduce

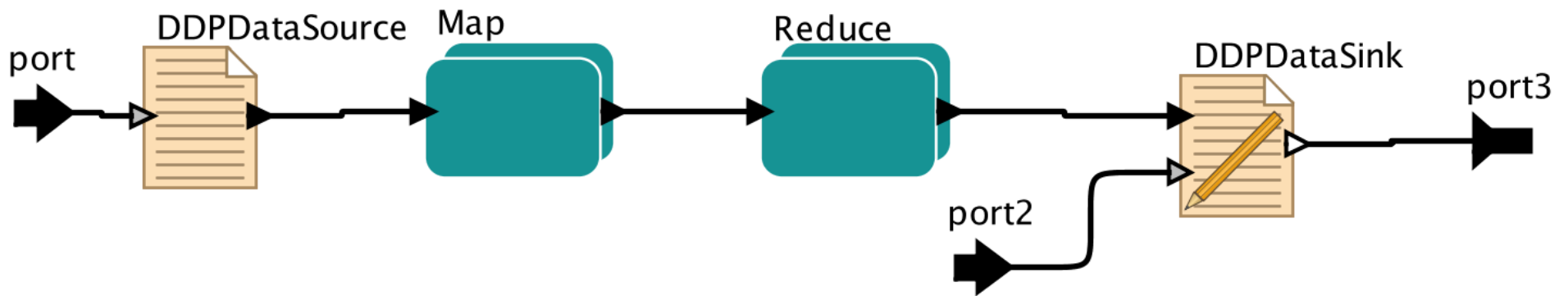
The screenshot shows the Kepler GUI window titled ".handson.blastall". It features two tabs: "Local Execution" and "MapReduce", with "MapReduce" selected. The interface contains a list of configuration parameters, each with a text input field and a "Configure" button to its right. The parameters and their values are as follows:

Parameter	Value
DBFormatCmd:	<code>\$BLASTDir/formatdb</code>
BLASTCmd:	<code>\$BLASTDir/blastall</code>
BLASTOptions:	<code>-p blastn -i stdin -m 8 -e 1e-5</code>
QueryFile:	<code>/usr/local/bioinf/sampledatablast/testQuery.fa</code>
RefSeqLength:	3311966
DBFormatOptions:	<code>-p F -o F</code>
DBPath:	<code>/tmp/</code>
AlignmentLimit:	250
ParallelNumber:	2
StratosphereConfPath:	
LocalRefFile:	<code>/usr/local/bioinf/sampledatablast/ocean_alaska_seqs/Ocean_Ala</code>
BLASTDir:	<code>/usr/bin</code>
ReferenceFile:	<code>/usr/local/bioinf/sampledatablast/ocean_alaska_seqs/Ocean_Ala</code>
OutputFile:	<code>property("user.home")+ "/blast-map-reduce.out"</code>

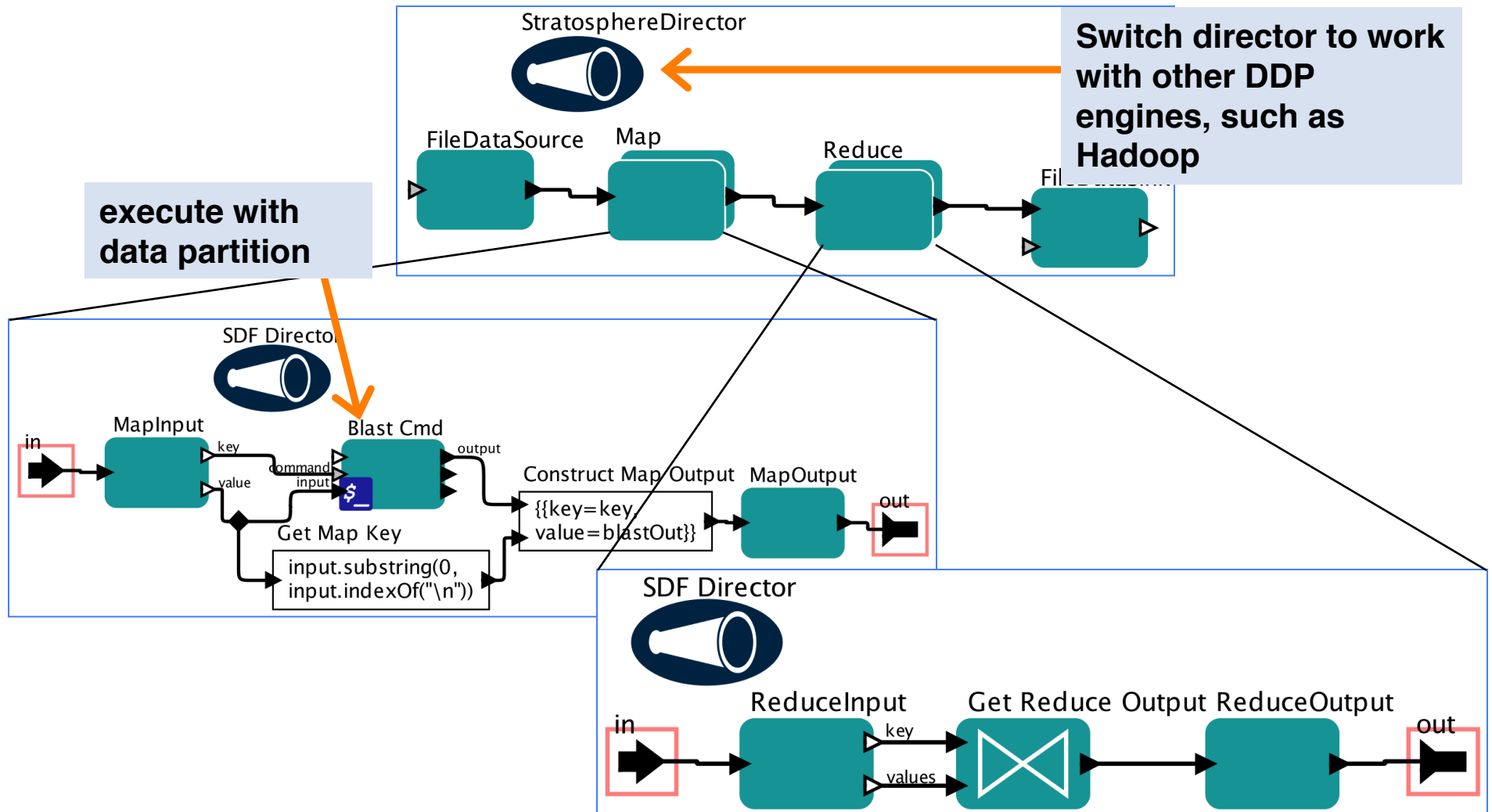
At the bottom of the window, there are three buttons: "Add", "Cancel", and "OK".

Inside the Stratosphere Blast

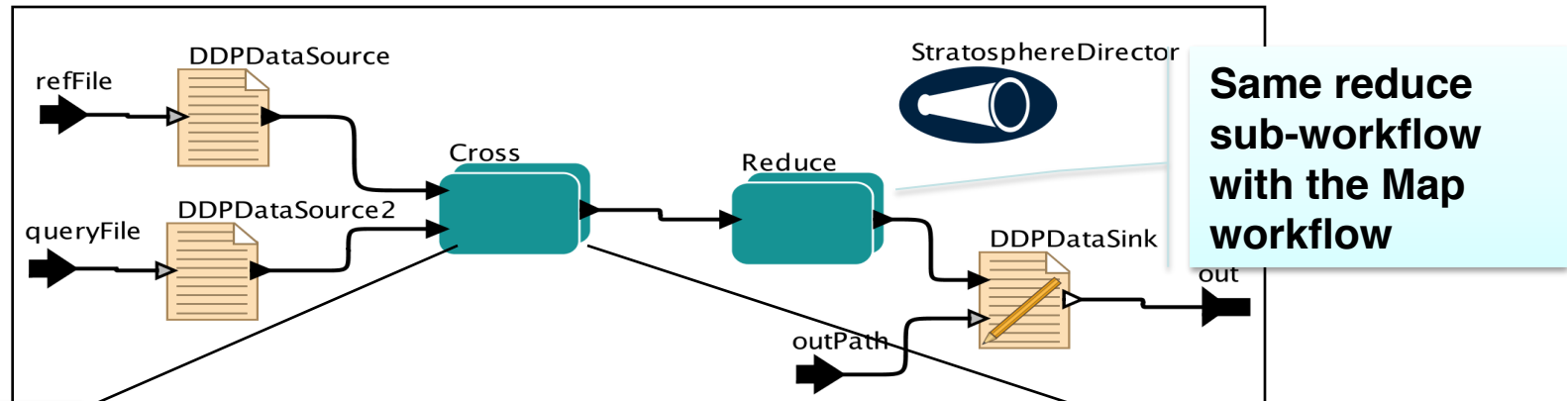
StratosphereDirector



DDP BLAST Workflow via Splitting Query Sequences



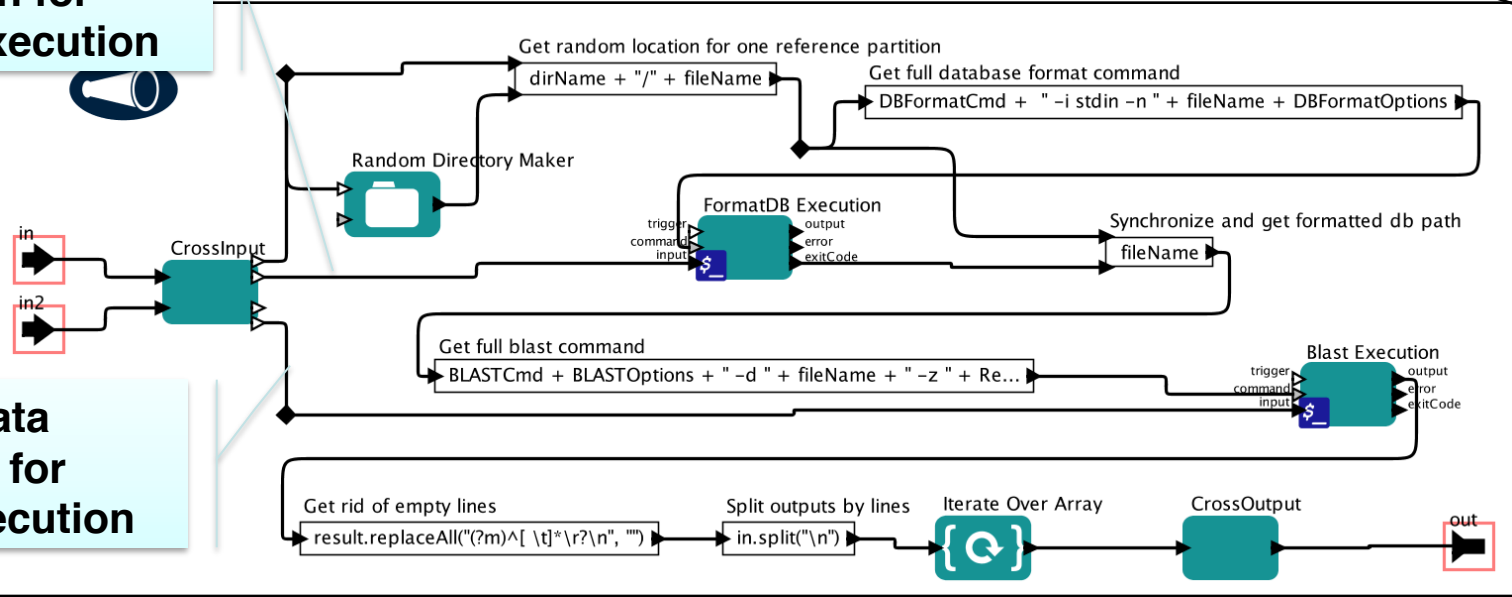
DDP BLAST Workflow using Cross and Reduce



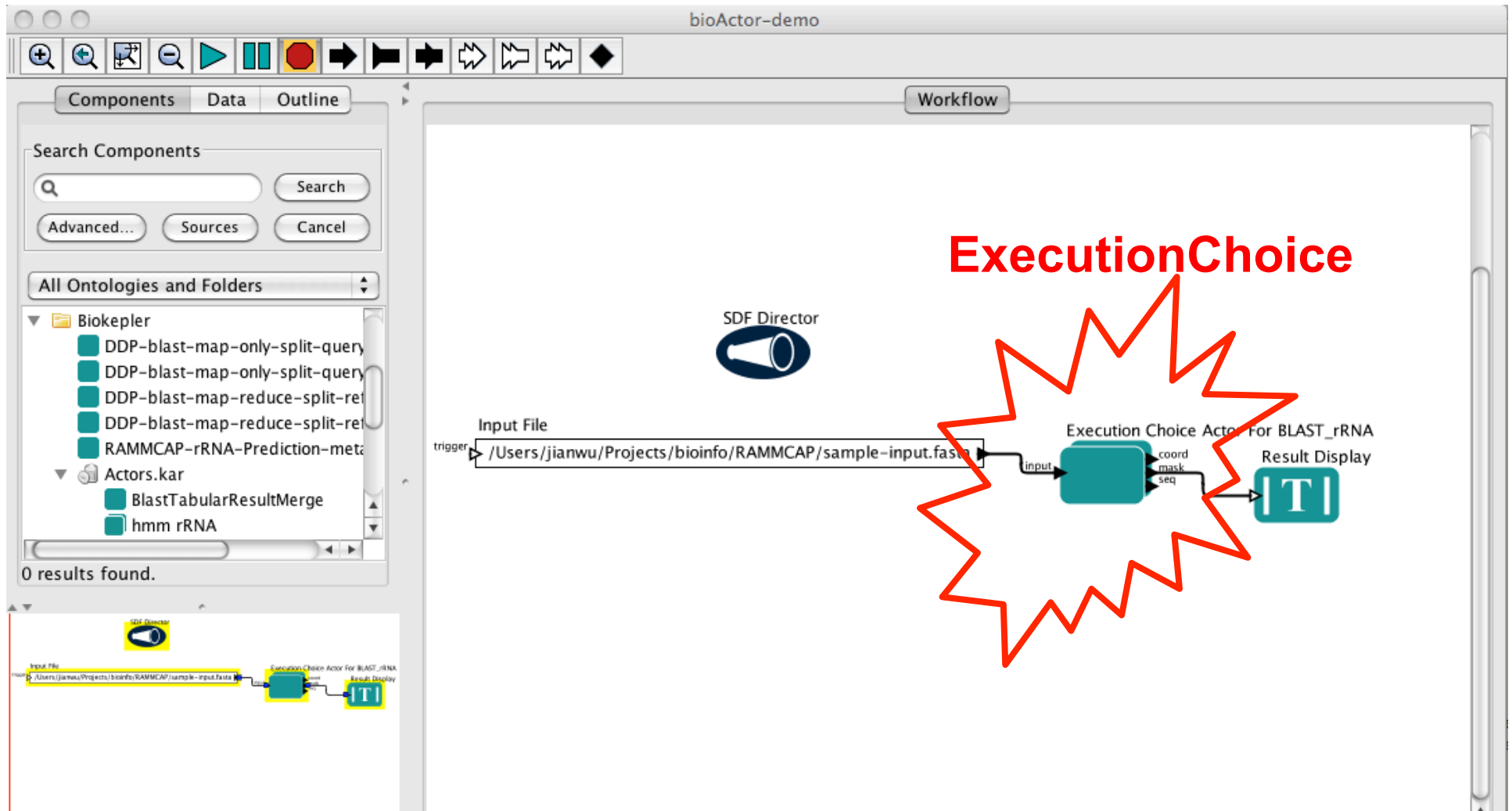
Same reduce sub-workflow with the Map workflow

Reference data partition for each execution

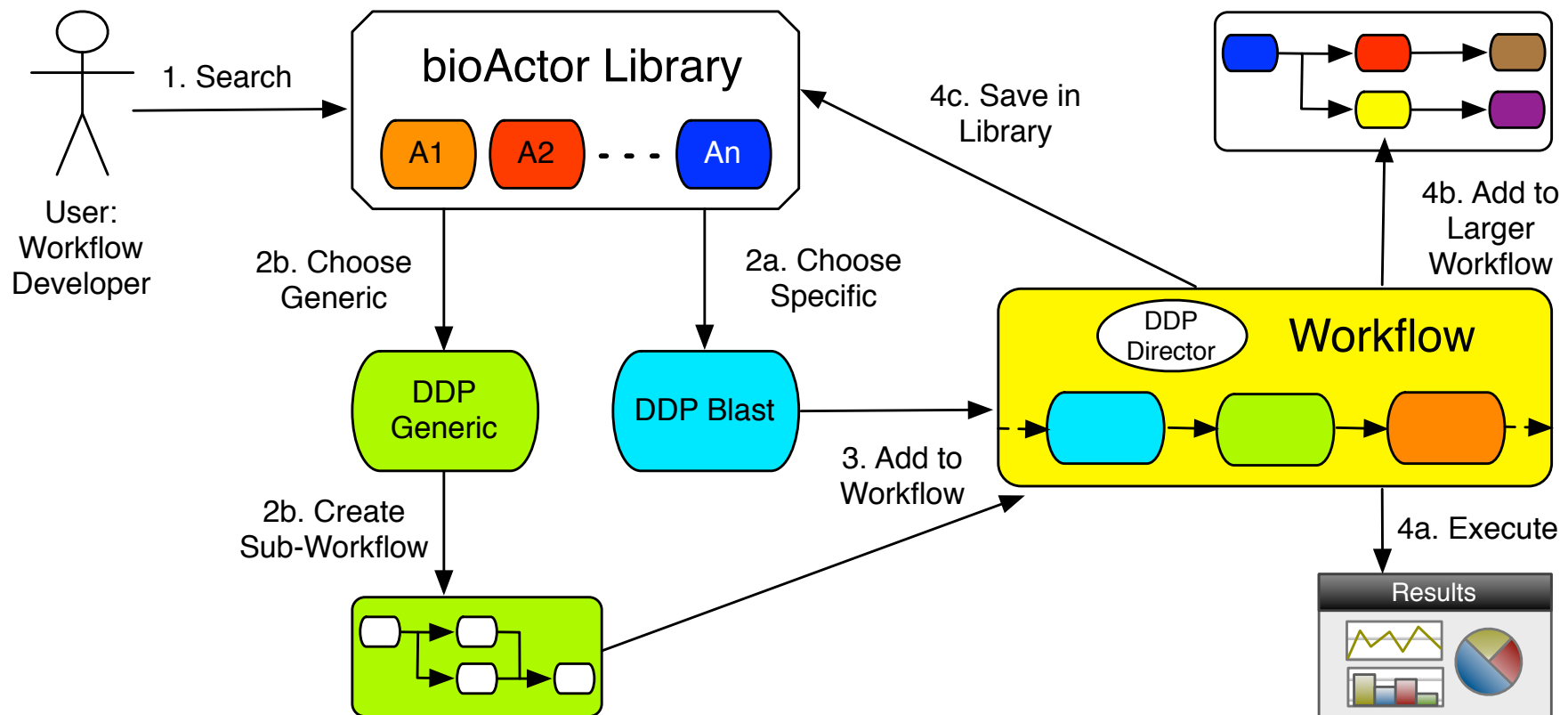
Query data partition for each execution



What if the bioActor I need is not available?



DDP bioActor Usage Model



***NEXT:
Kepler Interface and Introductory Examples on
Using Kepler***

Daniel Crawl

1st Workshop on bioKepler Tools and Its Applications

SDSC


UC San Diego



bioKepler - September, 2012

bioKepler.org