bioKepler: A Comprehensive Bioinformatics Scientific Workflow Module for Distributed Analysis of Large-Scale Biological Data

Ilkay Altintas[1], Daniel Crawl[1], Weizhong Li[2], Shulei Sun[2], Jianwu Wang[1], Sitao Wu[2]
http://www.biokepler.org/

bioKepler Project

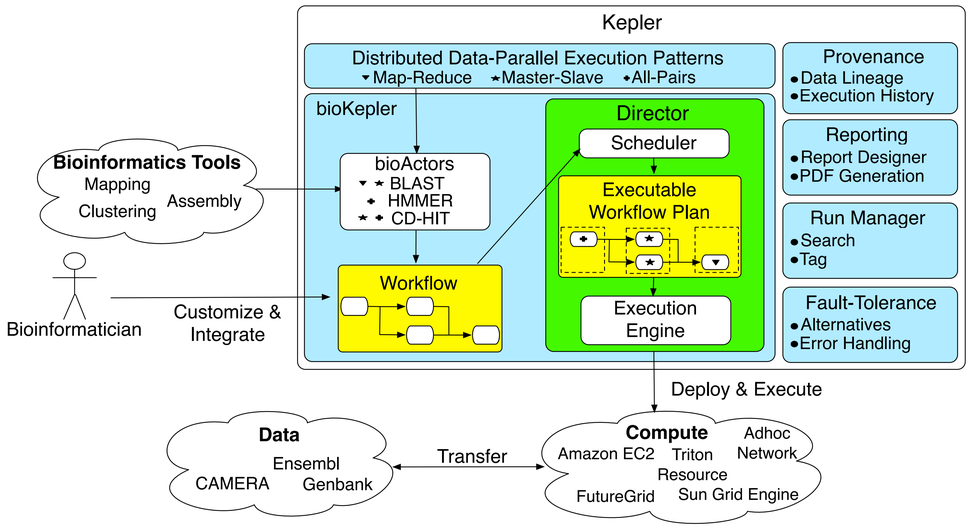
Challenges

- How can large-scale sequencing data be analyzed systematically in a way that incorporates and enables reuse of best practices by the scientific community?
- How can such analyses be easily configured or programmed by end users with various skill levels to formulate actual bioinformatics workflows?
- How can such workflows be executed on available computing resources in an efficient and intuitive manner?

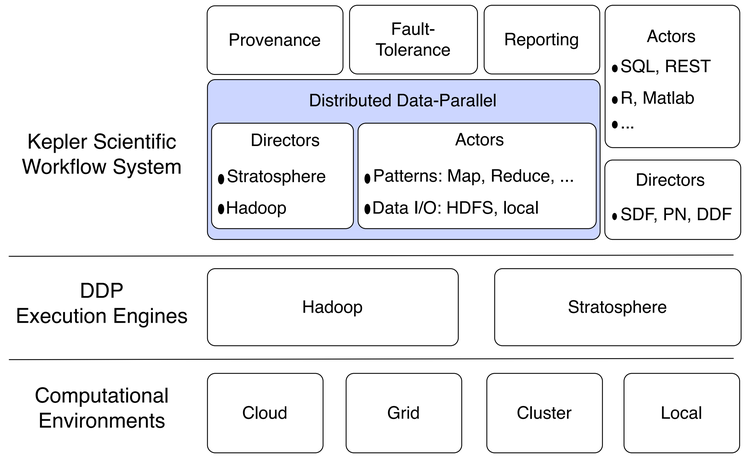
Approaches

- Use Distributed Data-Parallel (DDP) frameworks, e.g., MapReduce, to accelerate the execution of bioinformatics tools.
- Create configurable, reusable and executable DDP bioinformatics components in a Scientific Workflow System.
- Support different execution engines and computational environments and optimize workflow execution.

Conceptual Framework



Software Architecture

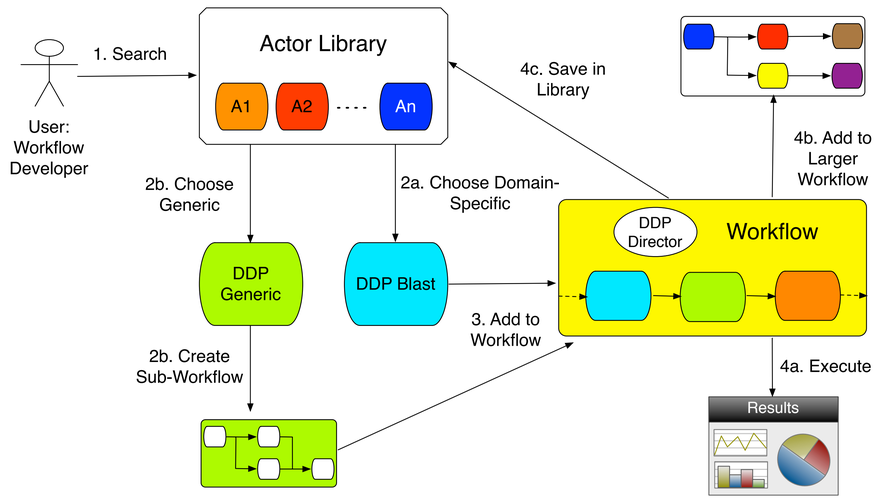


Sample bioActors

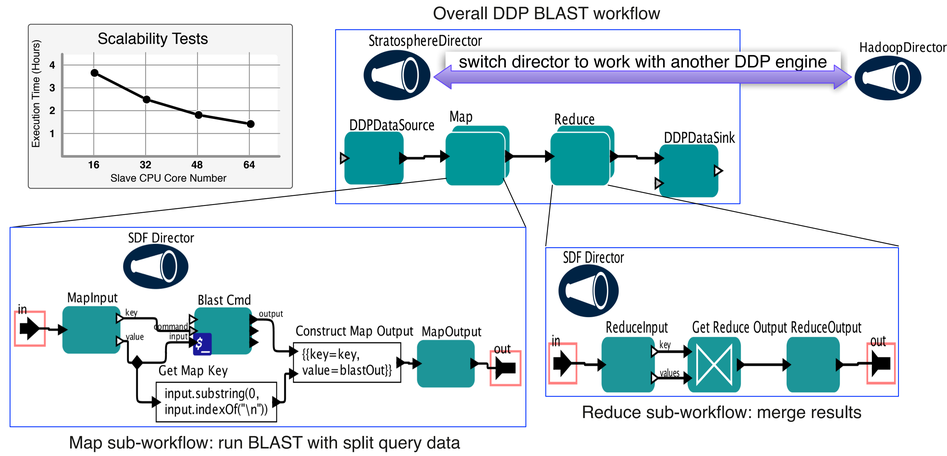
- **Alignment**: BLAST, BLAT
- **Hidden Markov Model**: HMMER
- **Functional Annotation**: COG, KOG, PRK, Pfam, KEGG
- **Profile-Sequence Alignment**: PSI-BLAST
- **Mapping**: Bowtie, BWA, Samtools
- **Assembly**: Velvet, SOAPdenovo, Abyss, Trinity
- **Multiple Alignment**: ClustalW, Muscle
- **Clustering**: CD-HIT, Blastclust
- **Gene Prediction**: Glimmer, Genescan, Fraggenescan
- **tRNA prediction**: tRNA-scan, Meta-RNA
- **Phylogeny**: FastTree, RAxML

(blue color tools have been built into bioActors, others are being built)

Usage of bioActor



Distributed Data-Parallel BLAST Workflow



- FileDataSource actor reads query sequence data, splits it, and generates key value pairs for parallel execution of the downstream Map sub-workflow.
- Each map/reduce sub-workflow execution only processes the partial data from its MapInput/ReduceInput actor.

bioKepler Workshop

http://www.biokepler.org/workshops

Time: September 5th and 6th, 2012
Venue: San Diego Supercomputer Center
Focus:

- Evaluation of bioinformatics and computational tools for bioActor development
- Generation of bioinformatics workflows based on conceptual workflows presented by workshop attendees

[1]San Diego Supercomputer Center
[2]Center for Research in Biological Systems
University of California, San Diego