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

*Project Website: http://www.biokepler.org*

Ilkay Altintas[1], Daniel Crawl[1], Weizhong Li[2], Shulei Sun[2], **Jianwu Wang**[1], Sitao Wu[2]

[1]*San Diego Supercomputer Center, UCSD*

[2]*Center for Research in Biological Systems, UCSD*

UCSD

Kepler

# Kepler: a Scientific Workflow System


www.kepler-project.org

- **A cross-project collaboration**
  initiated August 2003
  download times > 40,000
- **2.3 released on 20 Jan 2012**
- **Builds upon the open-source Ptolemy II framework**

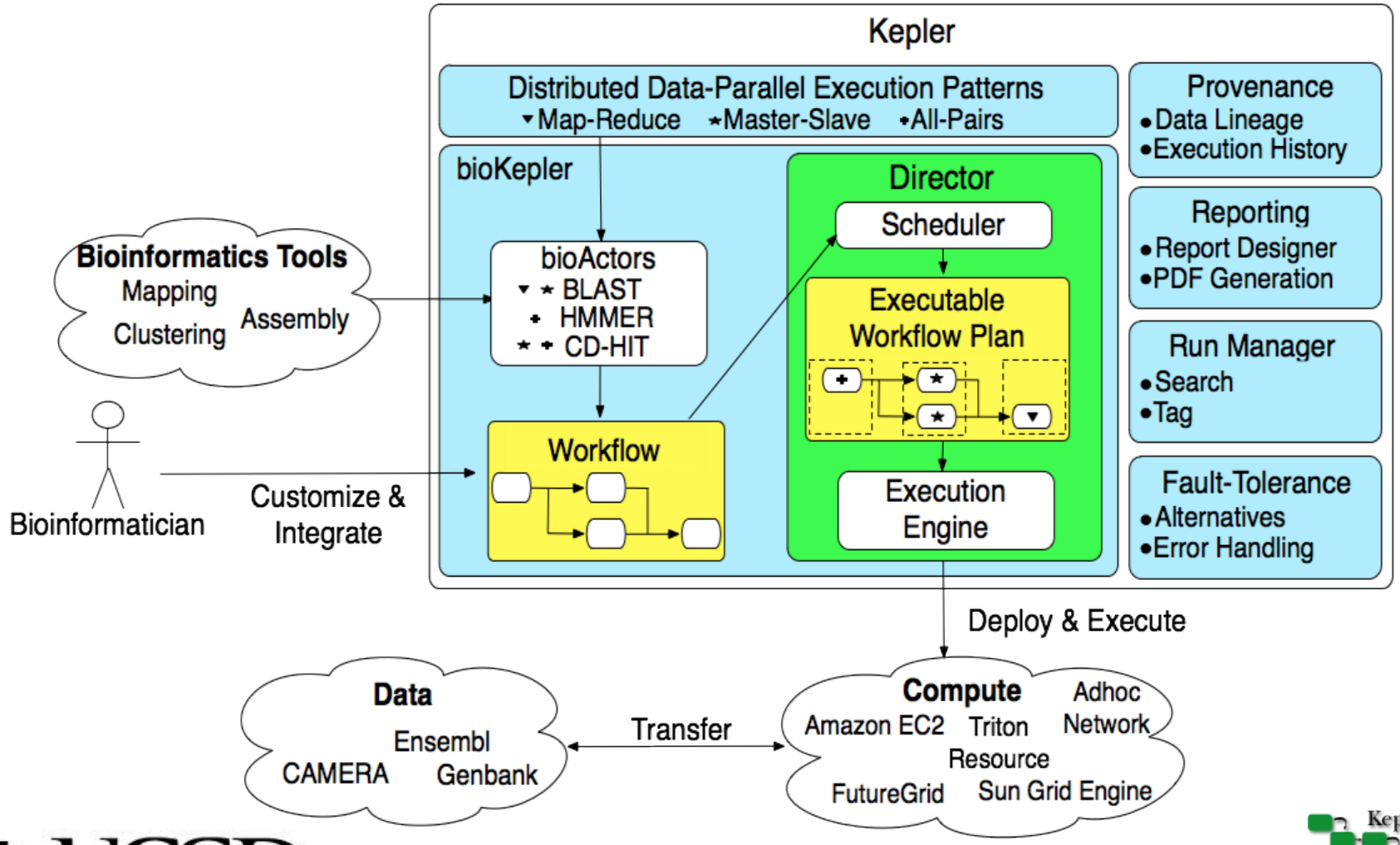
**Ptolemy II: A laboratory for investigating design**
**KEPLER: A problem-solving environment for Scientific Workflow**

**KEPLER = "Ptolemy II + X" for Scientific Workflows**

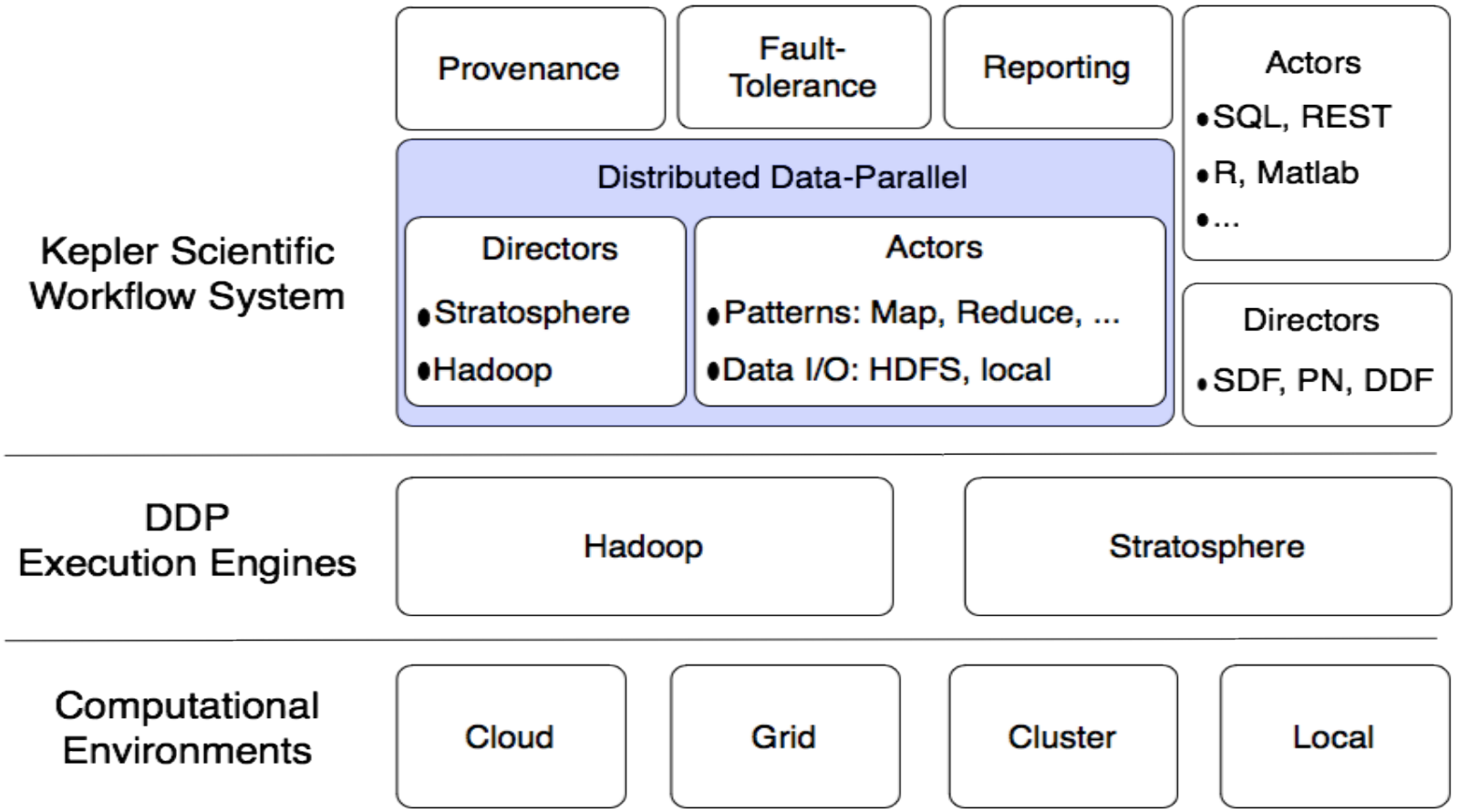
# bioKepler: a Module Being Built in Kepler

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

# Conceptual Framework



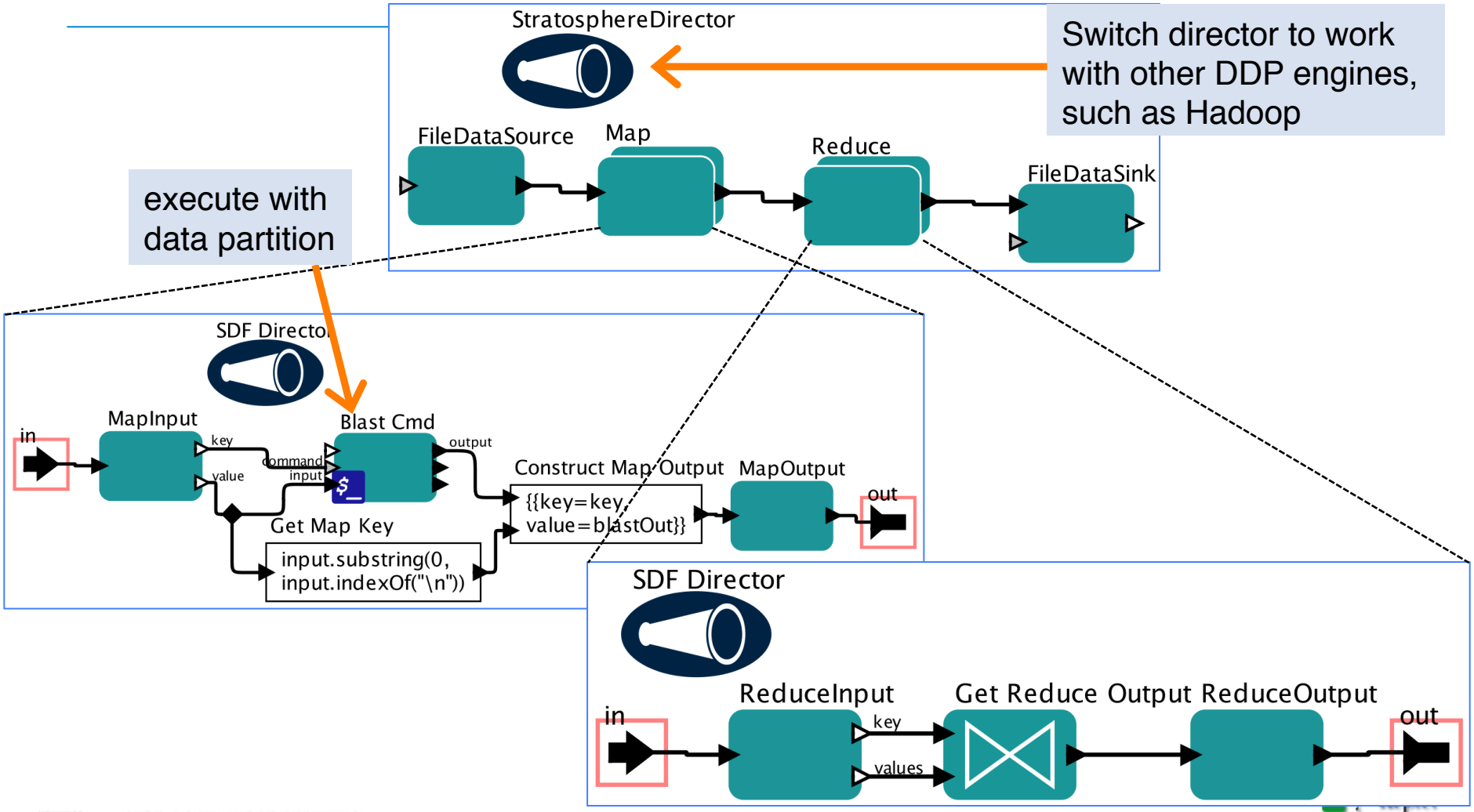
# Software Architecture



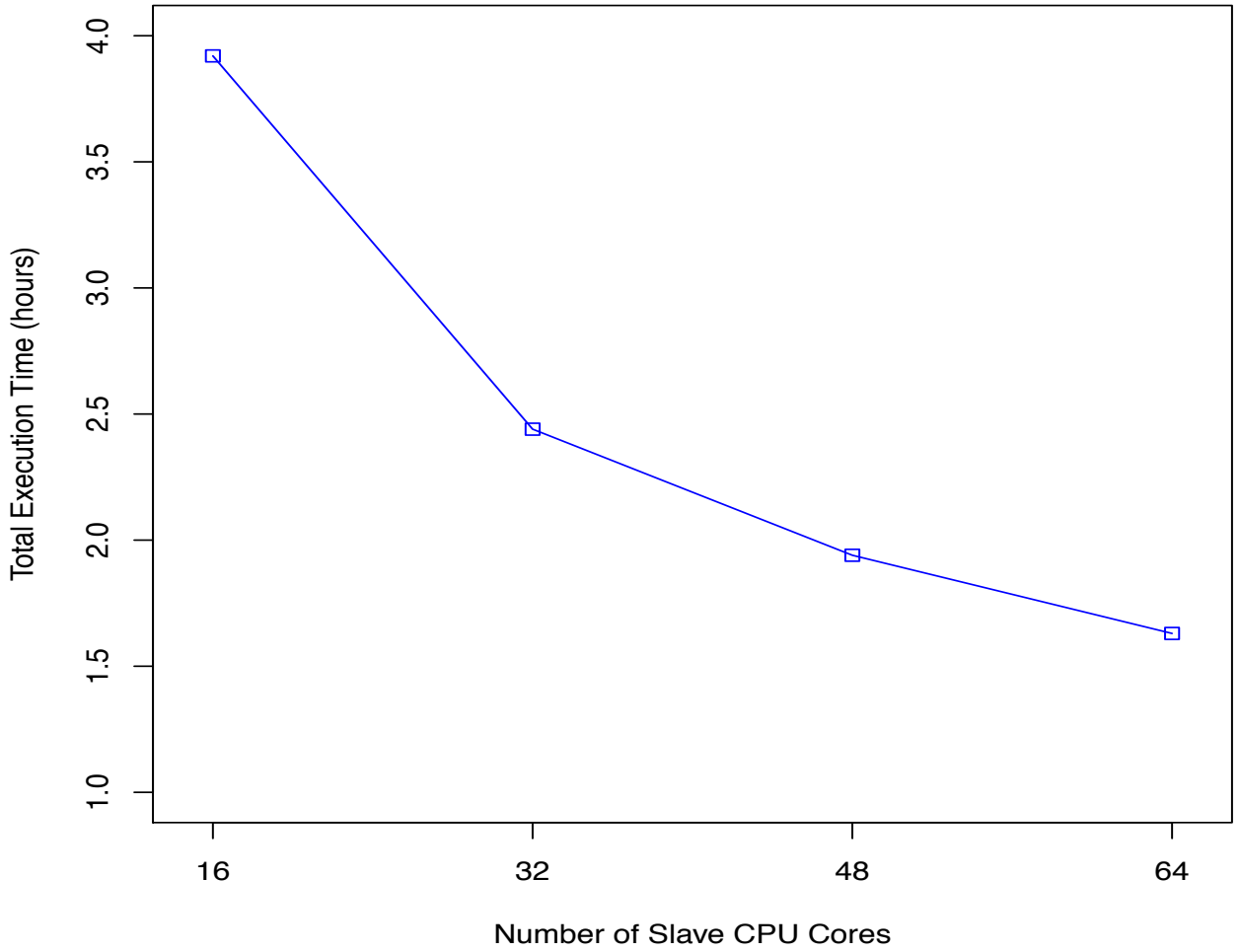
# Sample bioActors

- **Alignment:** BLAST, BLAT
- **Profile-Sequence Alignment:** PSI-BLAST
- **Hidden Markov Model:** HMMER
- **Mapping: Bowtie,** BWA, Samtools
- **Multiple Alignment:** ClustalW, Muscle
- **Clustering:** CD-HIT, Blastclust
- **Gene Prediction:** Glimmer, Genescan, Fraggenescan
- **tRNA prediction:** tRNA-scan, Meta-RNA
- **Phylogeny:** FastTree, RAxML

# DDP BLAST Workflow via Splitting Query Sequences



# DDP BLAST Workflow Experiments



# Questions?

- **More Information**

*jianwu@sdsc.edu*

*http://www.biokepler.org*

*http://www.kepler-project.org*

- **Acknowledgements**