On Optimization of Scientific Workflows to Support Streaming Applications in Distributed Network Environments

Qishi Wu, Yi Gu, Xukang Lu

University of Memphis
{qishiwu,yigu,xlv}@memphis.edu

Mengxia Zhu, Patrick Brown

Southern Illinois University
{mzhu, patiek}@cs.siu.edu

Wuyin Lin, Yangang Liu

Brookhaven National Laboratory
{wlin,lyg}@bnl.gov

The 5th Workshop on Workflows in Support of Large-Scale Science
(WORKS10)

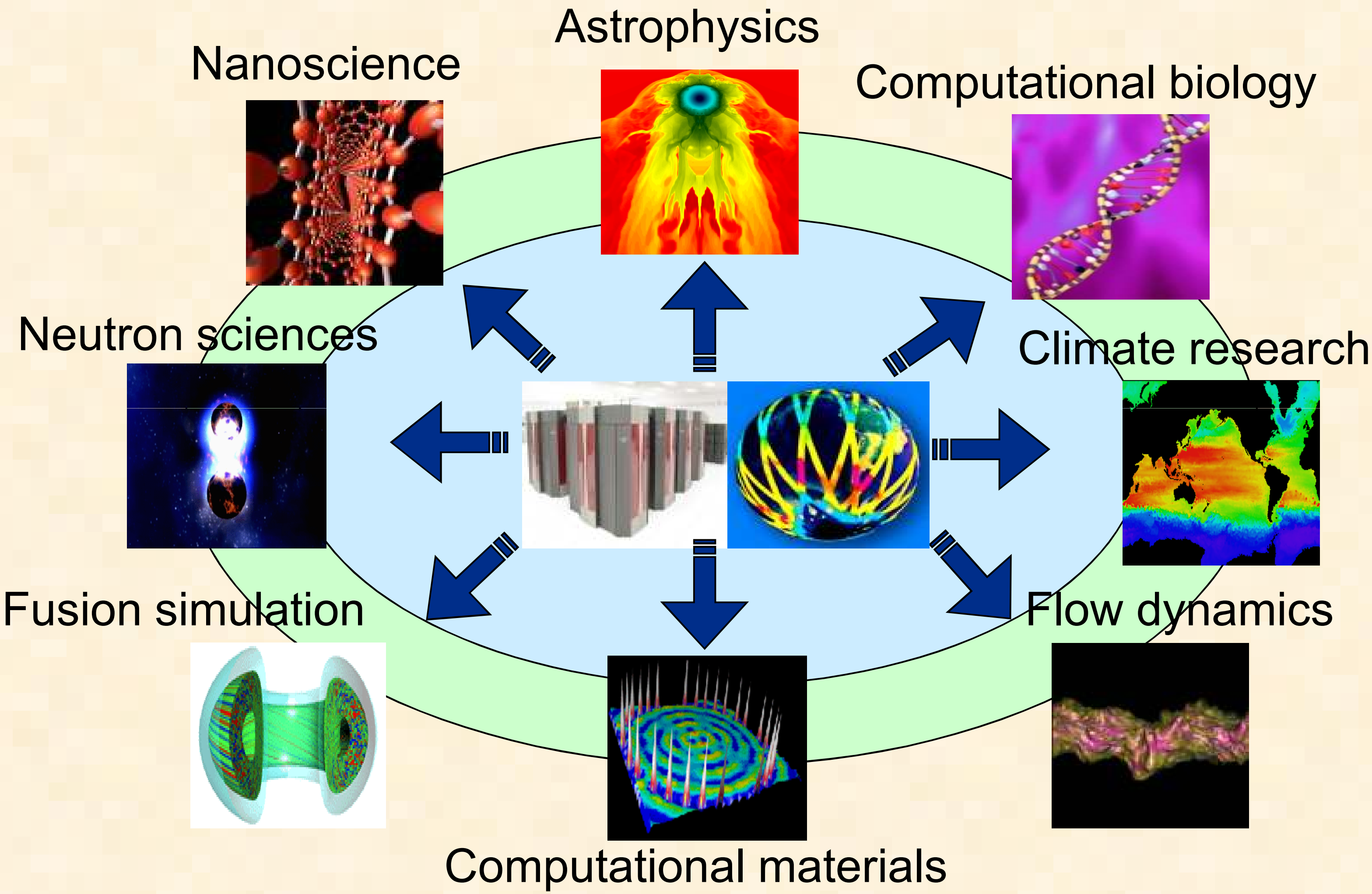
In conjunction with SC 2010, New Orleans, November 14, 2010

Sponsored by Department of Energy

Outline

- Background
- Mathematical Models and Problem Formulation
- Algorithm Design for MFR
- System Development – SWAMP
- Performance Evaluation
- Conclusion and Future Work

Background



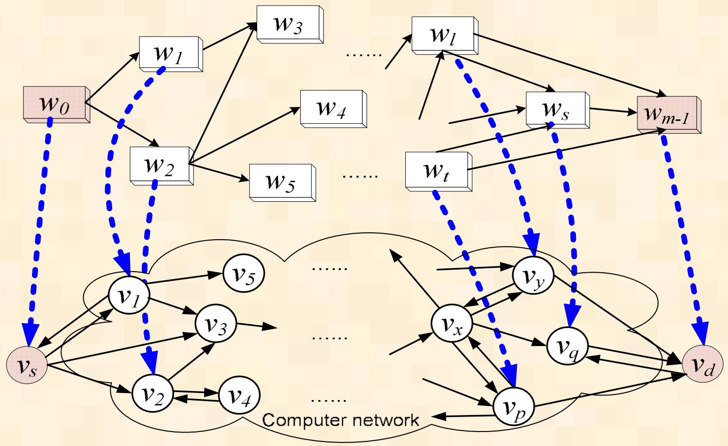
THE UNIVERSITY OF MEMPHIS

Southern Illinois University Carbondale

BROOKHAVEN NATIONAL LABORATORY

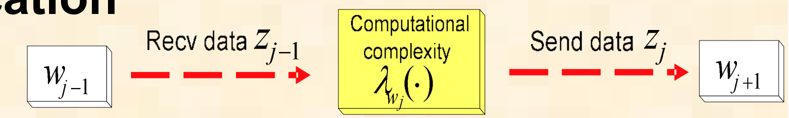
- **Applications structured as computing workflows**
  - **Simple case ⟶ Linear pipeline (a special case of DAG)**
  - **Complex case ⟶ DAG-structured graph**
- **Resources of different types**
  - **Deployed at various research institutes and laboratories**
  - **Accessed through wide-area network connections**
- **Challenges**
  - **Optimize application performance**
    - **Frame rate (throughput), end-to-end delay, reliability, etc.**
  - **Meet multifarious user requirements**
    - **Remote visualization, online computational monitoring and steering**
  - **Fully utilize distributed system and network resources**
  - **Automate computing process**

Mathematical Models and Problem Formulation

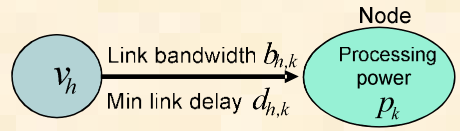


General DAG-structured workflow mapping

- **Cost models**
  - **A distributed computing workflow application**



    - **DAG-structured** $G_w = (V_w, E_w) \mid |V_w| = m$
    - **Vertices represent computing modules** $w_0, w_1, \cdots, w_{m-1}$
    - **Directed edges represent computing dependency**
  - **A heterogeneous computer network**



    - **An arbitrary weighted graph** $G_c = (V_c, E_c)$ $|V_c| = n$
    - **Nodes** $v_0, v_1, \cdots, v_{n-1}$ **interconnected by directed communication links**
  - **Goal: map modules to computer nodes to achieve *Maximum Frame Rate (MFR)* for streaming applications**
    - ***Frame Rate (FR)* or throughput, i.e. the inverse of the global Bottleneck (BN) of the workflow**

THE UNIVERSITY OF MEMPHIS | Southern Illinois University Carbondale | BROOKHAVEN NATIONAL LABORATORY

Module execution time:

$$T_{\text{exec}}(w,v)=\sum_{t=t_w^s}^{t_w^f}\frac{\alpha(t)\cdot\delta_w(t)}{p}, \text{ where}$$

$$\alpha(t)=\sum_{w\in V_w:(t_w^f-t)(t-t_w^s)\ge 0} x_{wv},\quad \lambda_w(z_w)=\sum_{t=t_w^s}^{t_w^f}\delta_w(t)$$

Data transfer time:

$$T_{\text{tran}}(e,l)=\sum_{t=t_e^s}^{t_e^f}\frac{\beta(t)\cdot\delta_e(t)}{b}+d, \text{ where}$$

$$\beta(t)=\sum_{e\in E_w:(t_e^f-t)(t-t_e^s)\ge 0} y_{el},\quad z_e=\sum_{t=t_e^s}^{t_e^f}\delta_e(t)$$

- **Objective function**

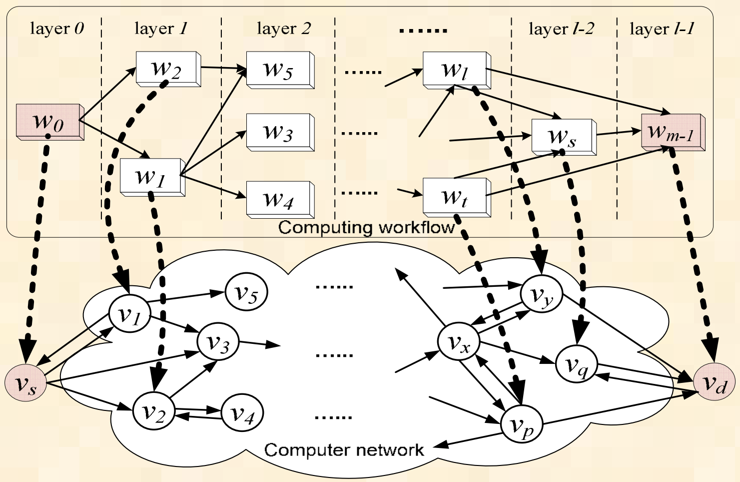
  Maximize frame rate determined by the global Bottleneck Time (BT) to achieve smooth dataflow for streaming applications

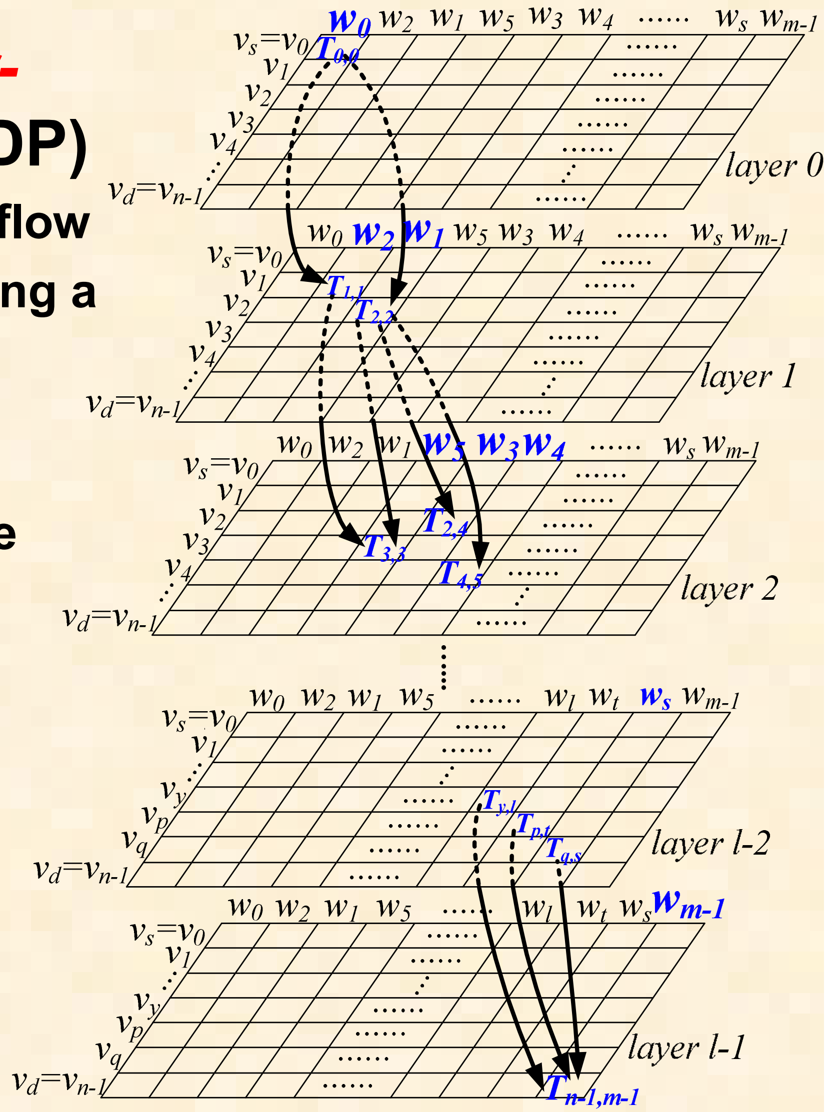
$$T_{\text{BT}}(G_w \text{ mapped to } G_c)=\max_{\substack{w_i\in V_w, e_{i,j}\in E_w\\ v_{i'}\in V_c, l_{j',k'}\in E_c}}\begin{pmatrix}T_{\text{exec}}(w_i,v_{i'})\\ T_{\text{tran}}(e_{j,k},l_{j',k'})\end{pmatrix}=\max_{\substack{w_i\in V_w, e_{i,j}\in E_w\\ v_{i'}\in V_c, l_{j',k'}\in E_c}}\begin{pmatrix}\sum_{t=t_{w_i}^s}^{t_{w_i}^f}\frac{\alpha_i(t)\cdot\delta_{w_i}(t)}{p_{i'}}\\ \sum_{t=t_{e_{j,k}}^s}^{t_{e_{j,k}}^f}\frac{\beta_{j,k}(t)\cdot\delta_{e_{j,k}}(t)}{b_{j',k'}}+d_{j',k'}\end{pmatrix}$$

$$\max_{\text{all possible mappings}}\left(\frac{1}{T_{\text{BT}}}\right)$$

Algorithm Design

- **MFR mapping scheme ---- *Layer-oriented DP-based algorithm* (LDP)**
  - **Topologically sort the computing workflow**
  - **Map it to the network layer-by-layer using a DP-based procedure**
  - **Select the best mapping in the current column which is used to decide the mapping for succeeding modules in the later columns**



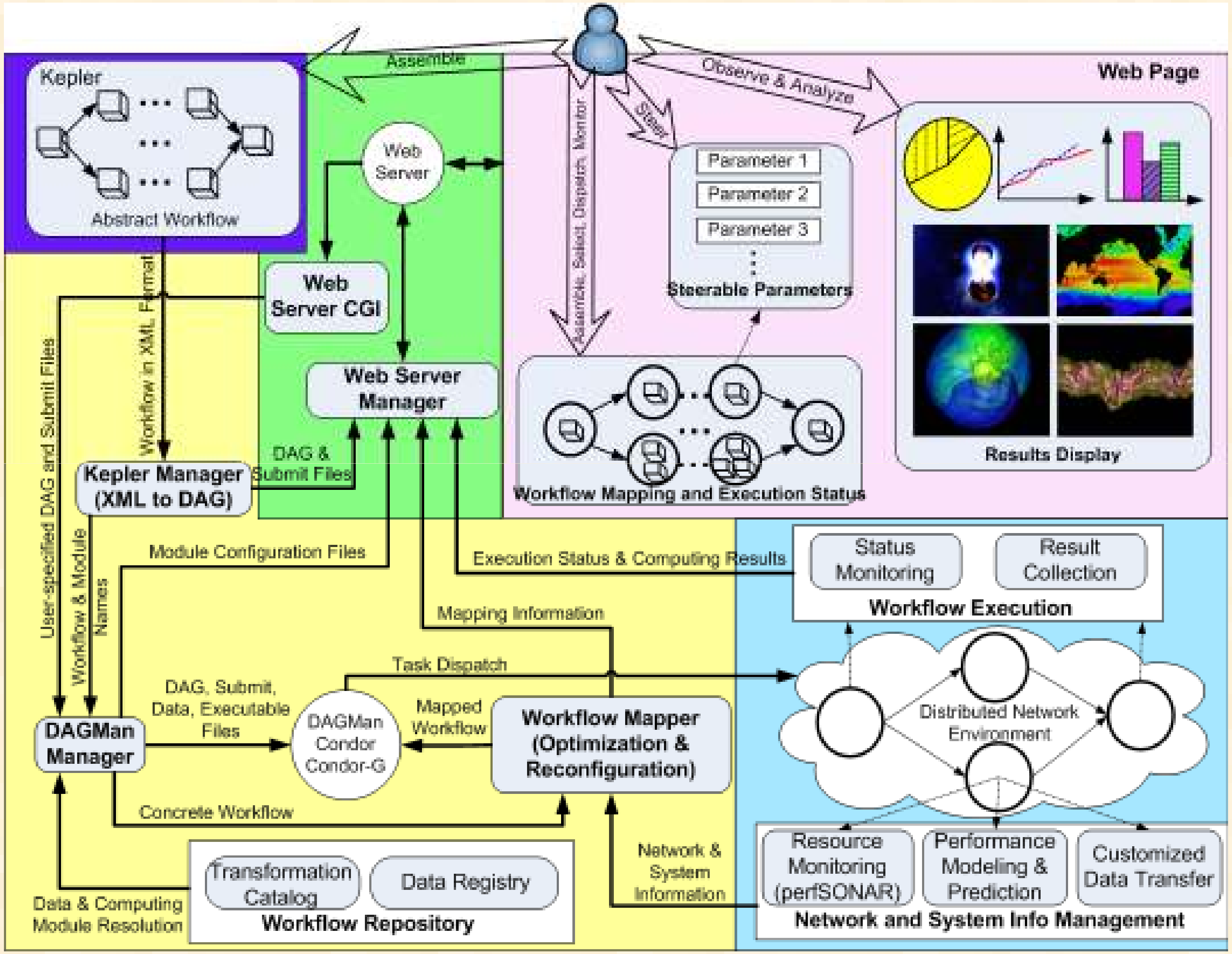


Scientific Workflow Automation and Management Platform (SWAMP)

- **Provide a unified web-based user interface to automate and manage workflow executions**
- **Support visual construction of abstract workflows**
  - A generic web-based graphical toolkit
  - The GUI of Kepler
- **Feature a special network-aware workflow mapper**
  - Automatically map abstract workflows to various networks such as ESG/OSG
    - Condor, PBS, and LSF
  - Achieve optimal end-to-end performance (MFR) based on real-time network status measurements
- **Interact with a variety of data movement services**
  - GridFTP (already incorporated)
  - SRM, TeraPaths, RFT, SRB, OSCARS, etc. (in plan)

THE UNIVERSITY OF MEMPHIS Southern Illinois University Carbondale BROOKHAVEN NATIONAL LABORATORY

SWAMP Framework



THE UNIVERSITY OF MEMPHIS   Southern Illinois University Carbondale   BROOKHAVEN NATIONAL LABORATORY

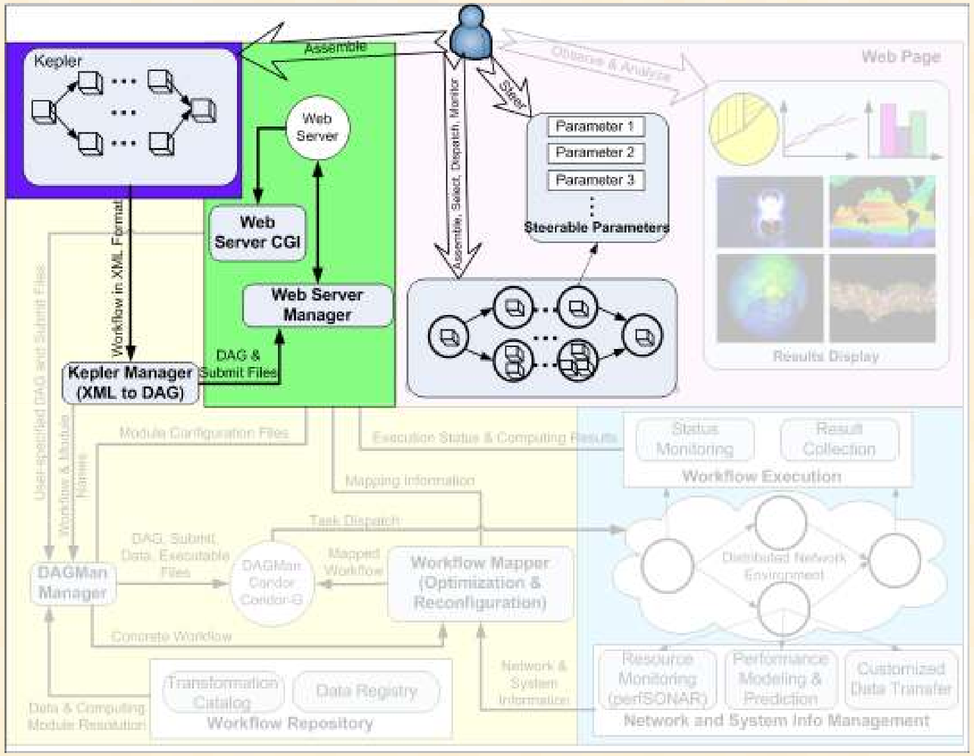
Web Interface

- Users interact exclusively with SWAMP via the web interface
  - Workflow Generation
  - Workflow Selection and Dispatch
  - Workflow Monitor and Result Display
- Web Interface Framework Design
  - A Model-View-Controller (MVC) architecture
    - HTML
    - CSS
    - JavaScript
    - PHP using Zend Framework 1.9 and PHP 5

THE UNIVERSITY OF MEMPHIS Southern Illinois University Carbondale BROOKHAVEN NATIONAL LABORATORY

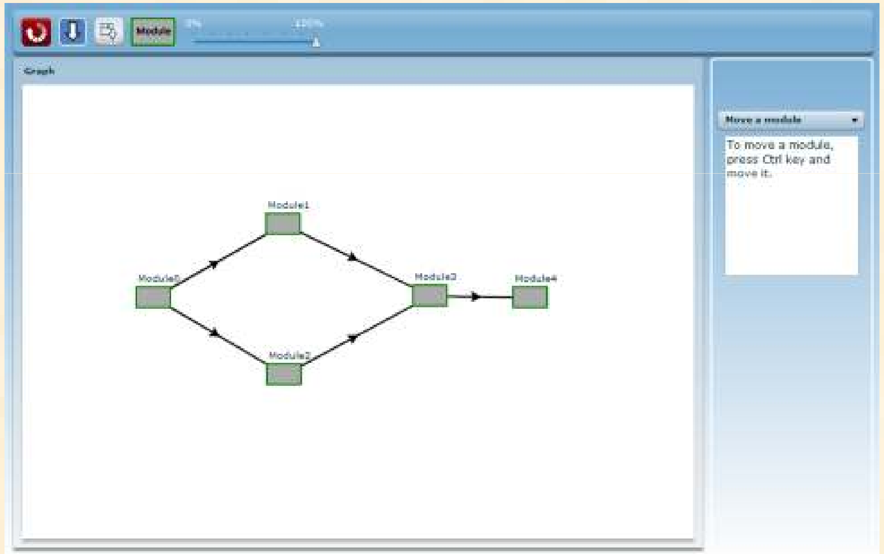
Workflow Generation

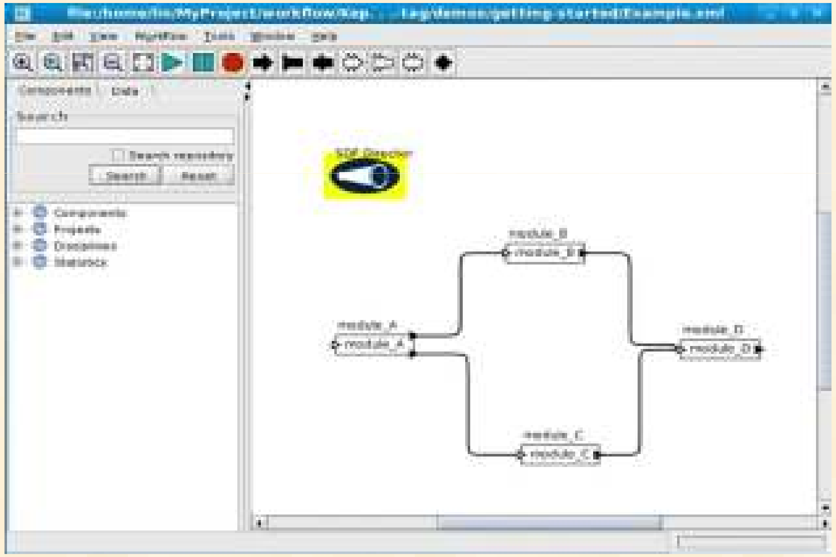
- **Involved components**
  - Kepler Manager
  - Web Interface
  - Web Server Manager
  - Web Server CGI



- **Workflow composition**
  - An intuitive way
- **Abstract workflows**
  - Modules are independent of underlying resources

- **Interactive composition of small & simple workflows**
  - **A generic web-based graphical toolkit**
  - **The GUI of Kepler**





THE UNIVERSITY OF MEMPHIS

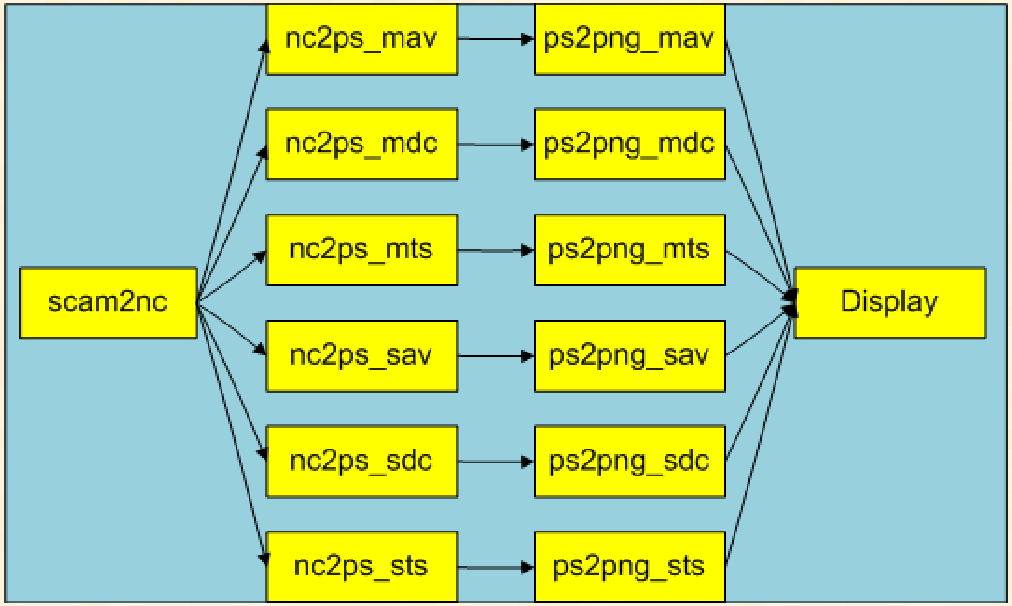
Southern Illinois University Carbondale

BROOKHAVEN NATIONAL LABORATORY

- **Automatic generation of large & complex workflows**
  - **Programs/Web Services**
    - **Combine parameters in a way specified by scientists**
    - **Inputs to programs/Web Services are part of workflow descriptions**
      - **Input files**
      - **Executables and their parameters**
      - **Output files produced by executables**
  - **Scale well with different applications**
  - **Two use cases**
    - **Climate Modeling (community)**
    - **SNS**

A use case: Climate Modeling at BNL

- **NCAR single column CAM (SCAM)**
  - Facilitate the development and evaluation of climate model parameterizations for the NCAR Community Atmosphere Model (CAM)
  - For each combination of available physics packages, data to be used, and experiment controls, we treat it as a dispatch of one SCAM workflow
- **SCAM Workflow**
  - Perform an SCAM model experiment (the scam2nc module) to generate simulated data
  - Both model simulated data and observed data are used for post-processing and visualization



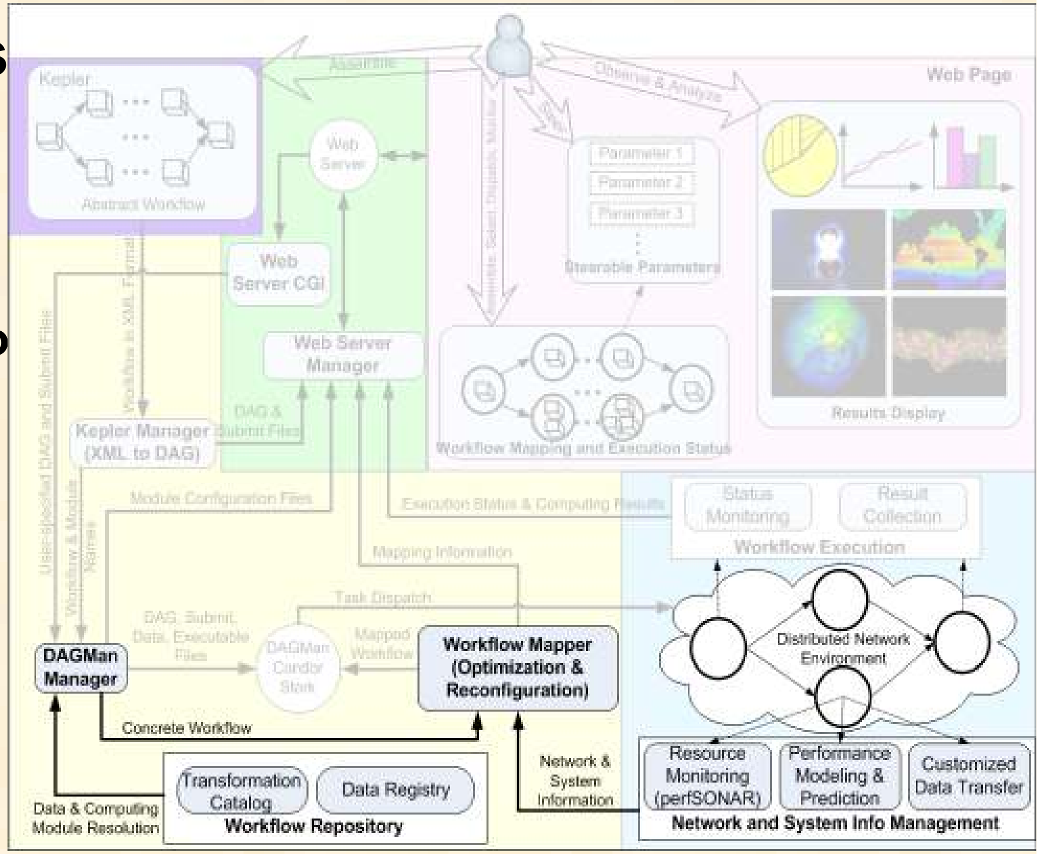
THE UNIVERSITY OF MEMPHIS

Southern Illinois University Carbondale

BROOKHAVEN NATIONAL LABORATORY

Workflow Mapping

- **Involved components**
  - DAGMan Manager
  - Workflow Mapper
  - Workflow Repository
  - Network and System Info Management



- **Need to discover**
  - Available networking/ computing resources, data and executables

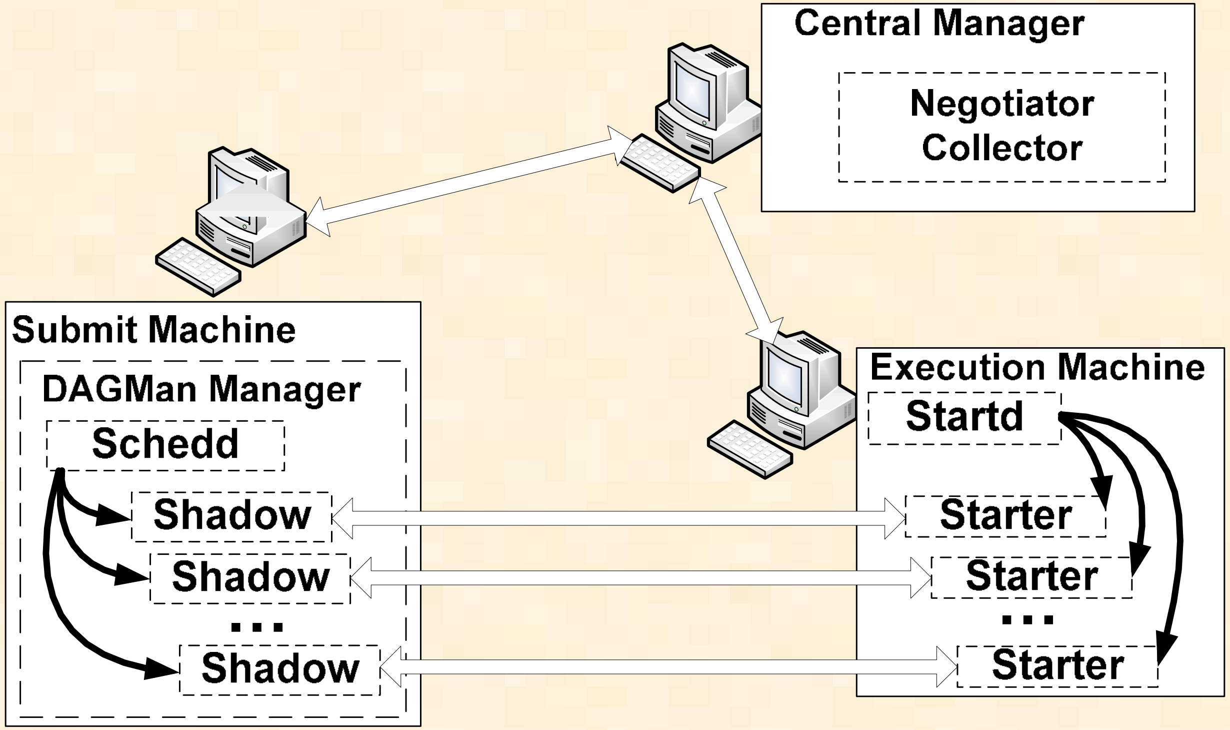
- **Executable Workflows**
  - Workflow modules are bound to specific computing resources
  - Data movements are specified

THE UNIVERSITY OF MEMPHIS
Southern Illinois University Carbondale
BROOKHAVEN NATIONAL LABORATORY

- **Discovery of available data movement services**
  - **GridFTP (move huge global climate modeling data from A to B)**
  - **SRM, TeraPaths, RFT, SRB, OSCARS, etc.**
- **Discovery of available networking/computing resources**
  - **Query the Network and System Info Management component**
    - **Interact with the Network Weather Service (NWS), One-Way Active Measurement Protocol (OWAMP) and Bandwidth Control (BWCTL) to find out link delay and available link bandwidth network and predictions**
  - **Query information systems**
    - **Interact with Globus Monitoring and Discovery Service (MDS), OSG RSV, and others to find out the number of CPUs, CPU frequency, queuing length, available disk space, etc.**
- **Employs network-aware mapping scheme**
  - **Layer-oriented Dynamic Programming (LDP)**

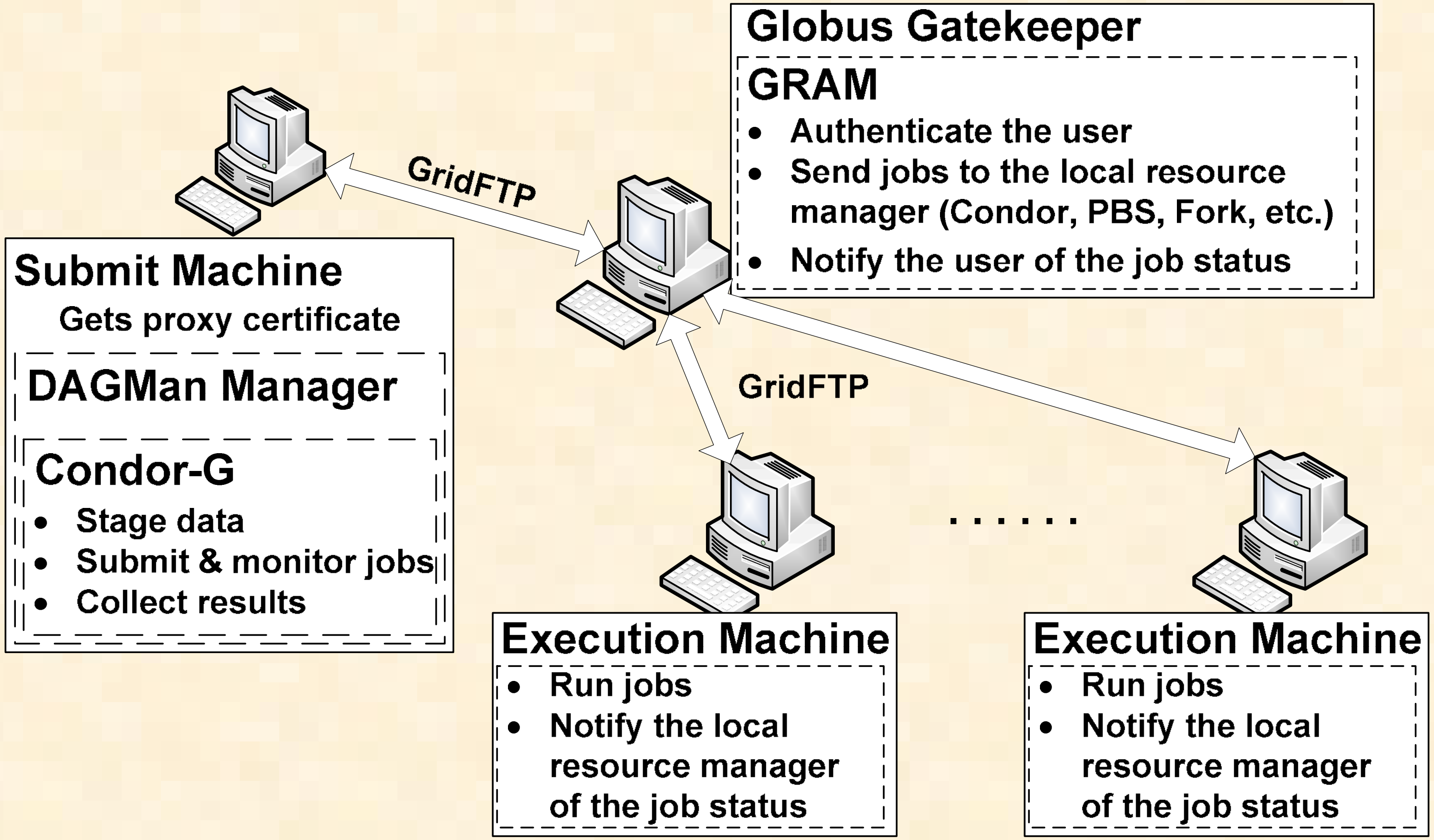
Workflow Execution

- Execution of executable workflows
  - Condor or Condor-G
  - Both use DAGMan to manage dependencies



Execution procedure of SWAMP in Condor Pool

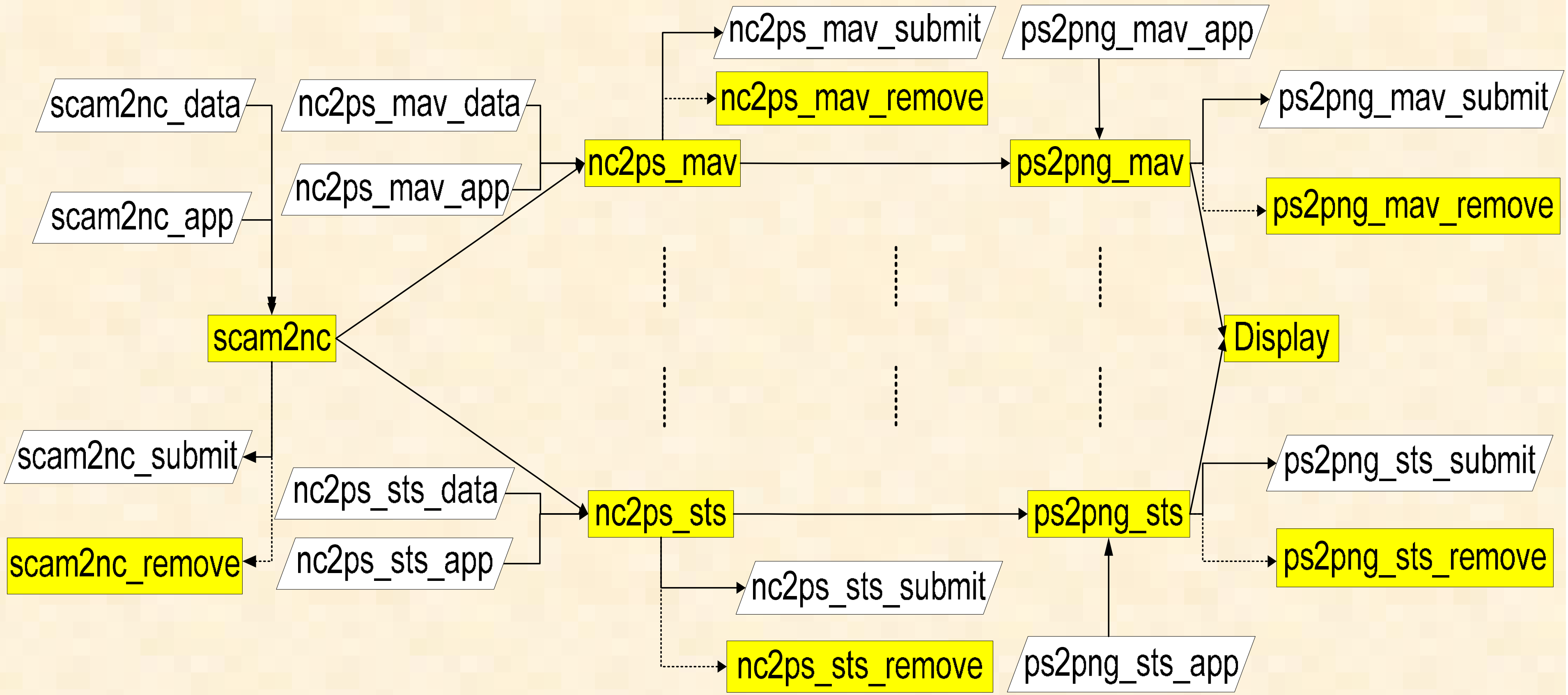
- **Execution of executable workflows by Condor-G**



Execution procedure of SWAMP in grid environments

Workflow Execution

- **Executable SCAM workflow**



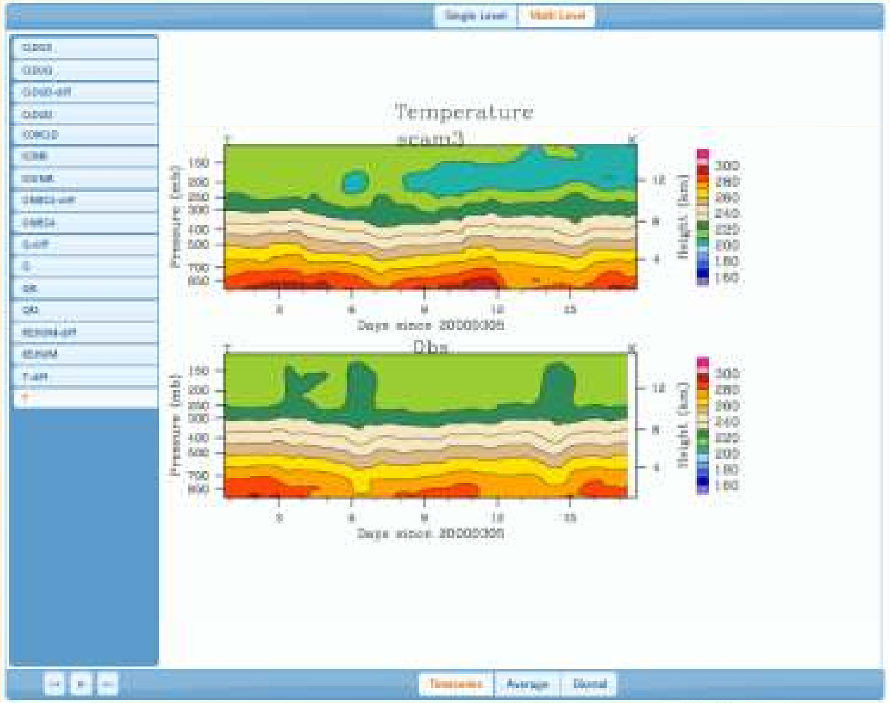
THE UNIVERSITY OF MEMPHIS

Southern Illinois University Carbondale

BROOKHAVEN NATIONAL LABORATORY

Workflow Display

- **Workflow status display**
  - Read Condor's logs to track status
  - Show real-time status through web interface
- **Immediate visualization of results**
- **Different types of display**
  - Graph
  - Gallery
  - Video
  - ……



THE UNIVERSITY OF MEMPHIS
Southern Illinois University Carbondale
BROOKHAVEN NATIONAL LABORATORY

Data Provenance Tracking

- **Data management is growing in complexity**
- **Data provenance**
  - Manage information about how data were produced starting from its original history
  - Verify the correctness of simulation data
  - Debug and reproduce simulations
  - Track workflows and simulations
- **SWAMP stores provenance information**
  - Where computing modules were executed
  - How long the execution took place
  - Which files were used and generated
  - Other task-level information

THE UNIVERSITY OF MEMPHIS
Southern Illinois University Carbondale
BROOKHAVEN NATIONAL LABORATORY

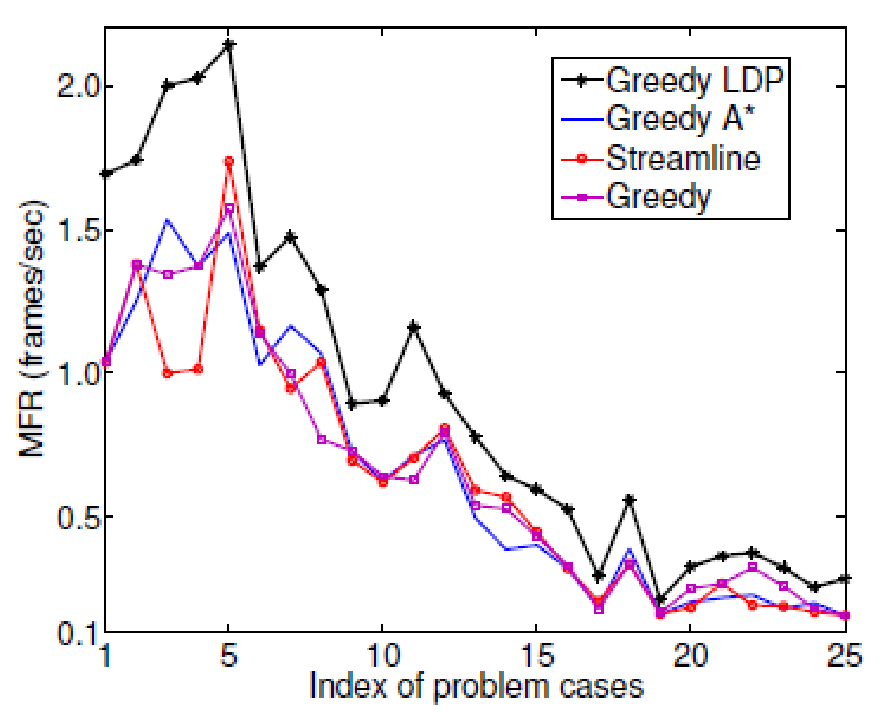
Performance Evaluation

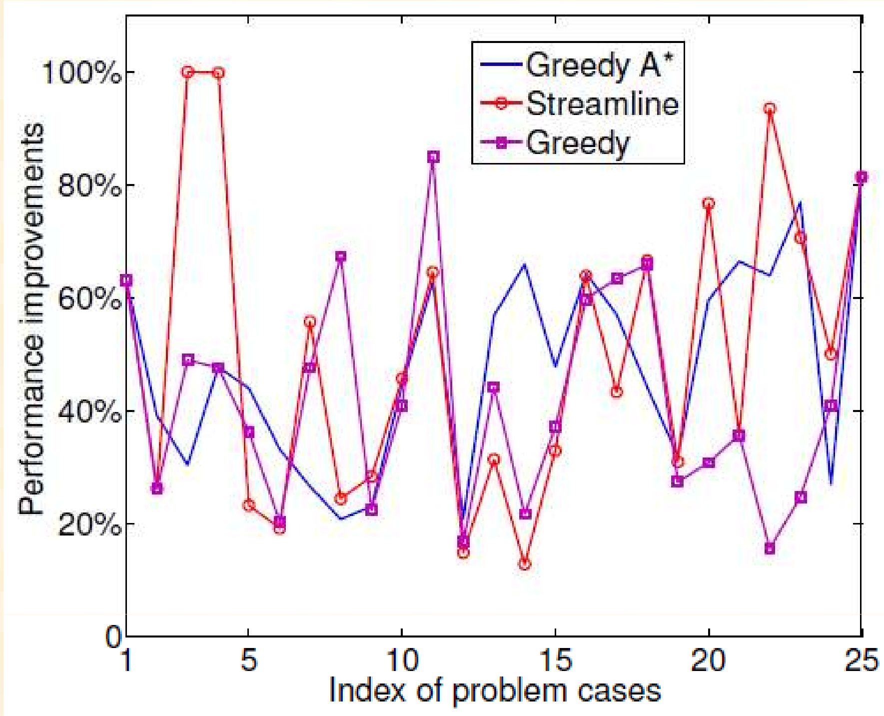
- **Simulation results**

MFR measurement among Greedy LDP, Greedy A*, Streamline and Greedy.

Prb Idx	Prb Size $m, \|E_w\|, n, \|E_c\|$	MFR (frames/sec) Greedy LDP	Greedy A^*	Streamline	Greedy
1	4,6,6,29	1.6925	1.0382	1.0382	1.0382
2	6,10,10,86	1.7403	1.2506	1.3789	1.3789
3	10,18,15,207	2.0003	1.5341	1.0001	1.3426
4	13,24,20,376	2.0252	1.3707	1.0133	1.3716
5	15,30,25,597	2.1408	1.4870	1.7367	1.5713
6	19,36,28,753	1.3697	1.0280	1.1490	1.1374
7	22,44,31,927	1.4748	1.1652	0.9467	0.9988
8	26,50,35,1180	1.2914	1.0690	1.0377	0.7718
9	30,62,40,1558	0.8936	0.7262	0.6962	0.7296
10	35,70,45,1963	0.9047	0.6261	0.6210	0.6418
11	38,73,47,2153	1.1613	0.7134	0.7057	0.6275
12	40,78,50,2428	0.9284	0.7677	0.8082	0.7952
13	45,96,60,3520	0.7790	0.4967	0.5930	0.5403
14	50,102,65,4155	0.6432	0.3876	0.5698	0.5284
15	55,124,70,4820	0.5953	0.4028	0.4478	0.4342
16	60,240,75,5540	0.5273	0.3208	0.3217	0.3300
17	75,369,90,7990	0.2954	0.1881	0.2061	0.1808
18	80,420,100,9896	0.5616	0.3895	0.3371	0.3385
19	90,500,150,22346	0.2132	0.1616	0.1628	0.1673
20	100,660,200,39790	0.3285	0.2058	0.1859	0.2512
21	120,680,220,48168	0.3647	0.2191	0.2687	0.2687
22	150,720,250,66235	0.3756	0.2291	0.1941	0.3248
23	165,850,280,78112	0.3241	0.1832	0.1899	0.2598
24	180,1000,300,89695	0.2566	0.2021	0.1710	0.1820
25	200,1200,500,249487	0.2859	0.1576	0.1576	0.1576

THE U ME





MFR comparison between Greedy LDP and other three mapping algorithms.

$$Speedup = \left| \frac{MFR_{LDP} - MFR_{other}}{MFR_{other}} \right|$$

MFR Performance speedups of Greedy LDP over other three algorithms.

THE UNIVERSITY OF MEMPHIS

Southern Illinois University Carbondale

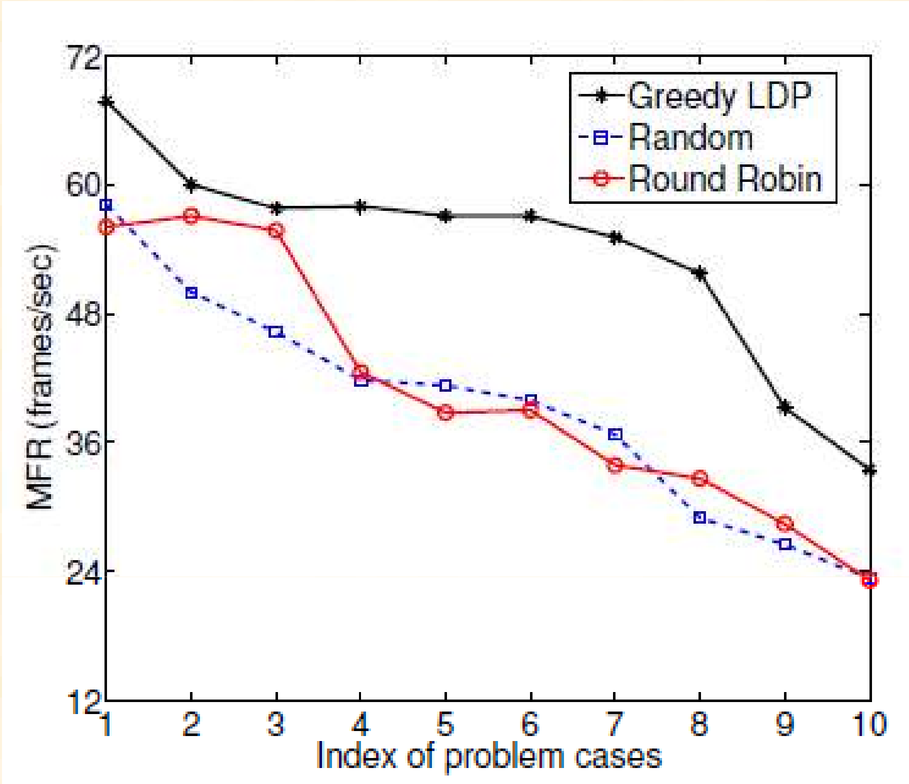
BROOKHAVEN NATIONAL LABORATORY

- **Experimental results using climate modeling workflow**
  - **A distributed heterogeneous network testbed**
    - **9 PC workstations at University of Memphis**
    - **3 PC workstations at Southern Illinois University at Carbondale**
    - **CPU frequency varies from 1.2 GHz to 3.4 GHz**
    - **Create an arbitrary network topology by configuring different firewall settings**
    - **Use “tc” command to allocate different bandwidth**

Input data size for 10 SCAM experiments.

Idx. of prb. cases	1	2	3	4	5	6	7	8	9	10
No. of days	2	4	6	8	10	12	14	16	18	20

THE UNIVERSITY OF MEMPHIS ® Southern Illinois University Carbondale BROOKHAVEN NATIONAL LABORATORY



MFR comparison between Greedy LDP, Random, and Round Robin for climate modeling workflow in SWAMP

THE UNIVERSITY OF MEMPHIS

Southern Illinois University Carbondale

BROOKHAVEN NATIONAL LABORATORY

Conclusion and Future Work

- Constructed cost models for workflows and networks
- Proposed a layer-oriented LDP algorithm for MFR
- Developed a workflow management system – SWAMP
  - A generic framework to support different applications in various network environments
  - A special network-aware workflow mapper
- Implemented a real-life workflow in SWAMP system
- Future work
  - Investigate more sophisticated and accurate cost models
  - Conduct more real-life workflow experiments in wide-area networks

THE UNIVERSITY OF MEMPHIS

Southern Illinois University Carbondale

BROOKHAVEN NATIONAL LABORATORY