BReW: Blackbox Resource Selection for e-Science Workflows

Yogesh Simmhan | USC

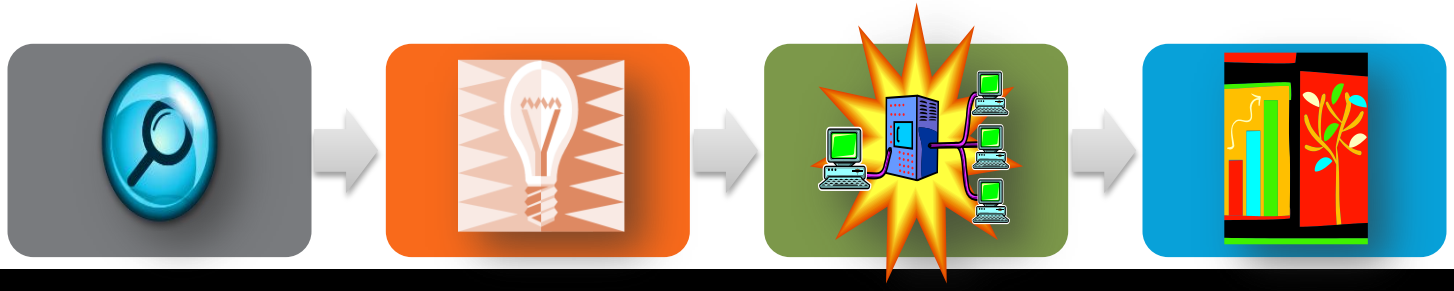
Emad Soroush | UW

Lavanya Ramakrishnan , Deb Agarwal | LBL

Catharine van Ingen | MSR

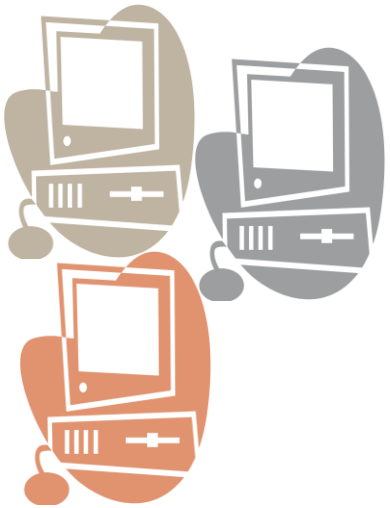
Growth of Workflows for eScience

- Complex workflows to model science experiments
  - Evolving over time
  - Different task characteristics
- Compute & Data intensive
  - 4th Paradigm of Science: Increasing size of science problems
  - Resource needs often exceed available ones



Diversity of Resource Platforms

- Workstation, Local Cluster, HPC Center, Cloud
- Different resource features
  - #/speed of cores, bandwidth
  - Resource acquisition, batch queues, policies
- Different programming characteristics
  - Requires effort to port



Platform Selection for Workflows

Precise workflow scheduling vs. *Approximate* platform selection

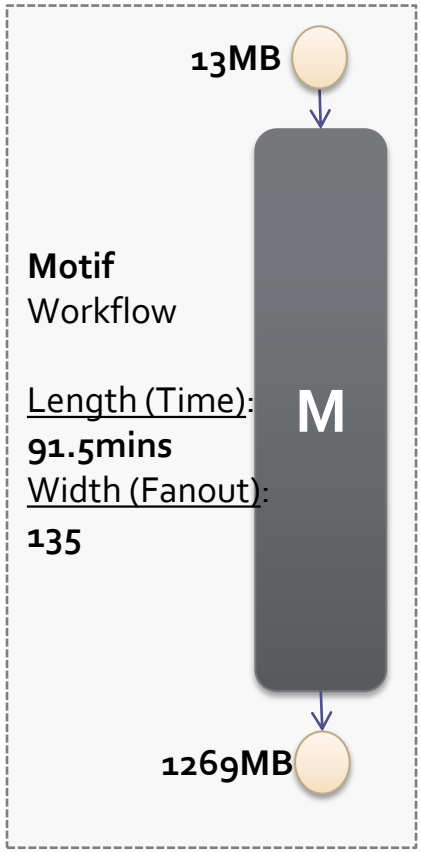
- DAG Scheduling
  - Map tasks or workflows to resources
  - Optimize to minimize makespan, maximize resource usage, …
  - Keep track of available resources, clusters
  - Built into workflow engines (Pegasus, Swift, Trident, DAGMan)
  - Requires complete description of workflow
- Approximate platform selection
  - Developers select platforms to develop, migrate workflows
  - Scientists choose resources to sign up for, feasibility study
  - Policy makers to choose for requesting grants

Can we make *a priori* decisions about suitability of a platform for a workflow with limited workflow knowledge

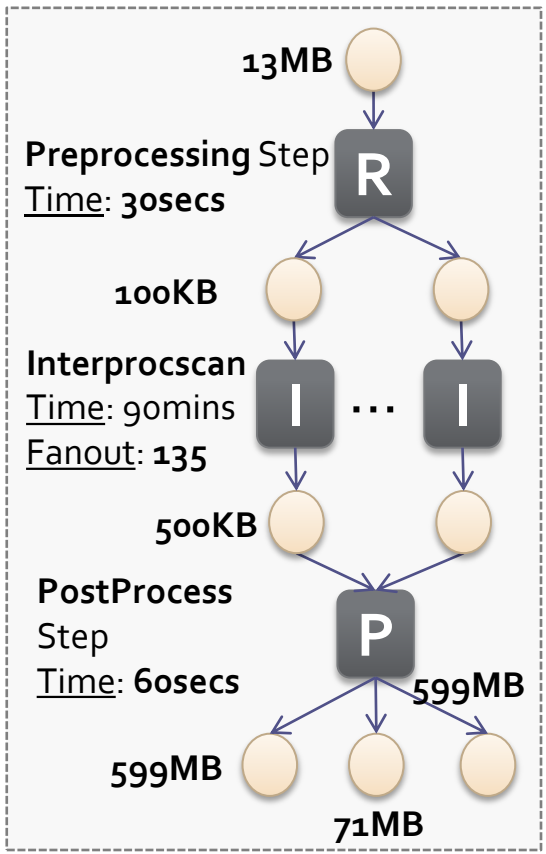
Workflow Attributes

- **Blackbox**
  - **Width**: Max fanout across workflow
  - **Length**: Time to run full workflow at scale
  - **Data I/O**: Total data input & output
  - ***MinCore***: Min # of concurrent cores reqd
- Whitebox
  - Tasks, dataflow, stages, …

BReW Blackbox



Whitebox



BReW Blackbox Model

- Determine workflow *makespan* for a platform using blackbox information

$$M_{Workflow} = T_{LatencyMax} + T_{DataSum} + \frac{\left(T_{WorkflowLength} \times N_{WorkflowWidth}\right)}{N_{Cores}}$$

- Time to acquire required number of cores (≤ width)
- Time to transfer data
- Time to perform computation on acquired cores

Comparative Whitebox Approach

- Makespan assuming full workflow knowledge

$$M_{Workflow} = \sum_{i=0}^{\#\ stages} M^i_{Stage}$$

$$M^i_{Stage} = T^i_{LatencyOne} \times Ceil\left(\frac{N^i_{TaskWidth}}{N^i_{Cores}}\right) + T^i_{Data} + \frac{\left(T^i_{TaskLength} \times N^i_{TaskWidth}\right)}{N^i_{Cores}}$$

- Sum of time taken in each workflow stage
- At each stage: Time to Acquire cores, Transfer data, Perform compute

BReW Framework

- Tool to make blackbox platform selection given coarse grained workflow details
- Initial support for 3 platforms
  - HPC (SDSC TeraGrid, IU BigRed)
  - Clouds (Azure)
  - Generic local Cluster
- Platform knowledge available
  - TeraGrid QBETS batch queue prediction system
  - µBenchmarks for Azure latency, bandwidth
- Also does whitebox selection if given DAG

- Parameter sweep across workflow attributes
- Analysis for eScience workflows

EVALUATION OF BREW

How closely can BReW make the same platform selection decisions as the whitebox model for different workflows ?

Assumptions

- Platforms
  - SDSC TeraGrid, IU BigRed: 2048 cores, 2.5GHz, 100Mbps, QBETS@50%
  - Azure: 2048 cores, 1.6GHz, 10Mbps, (200+20x*n*)s
  - Cluster: 512 cores, 2.5GHz, 1Gbps, instantly avail.
- Policies
  - Min of available and required cores for latency
  - BReW acquires once, retains cores
  - Whitebox acquires cores at each stage
  - Complete workflow runs on single platform
  - Data transfer from desktop to platform
    - BReW: At boundary; Whitebox: Includes intermediate

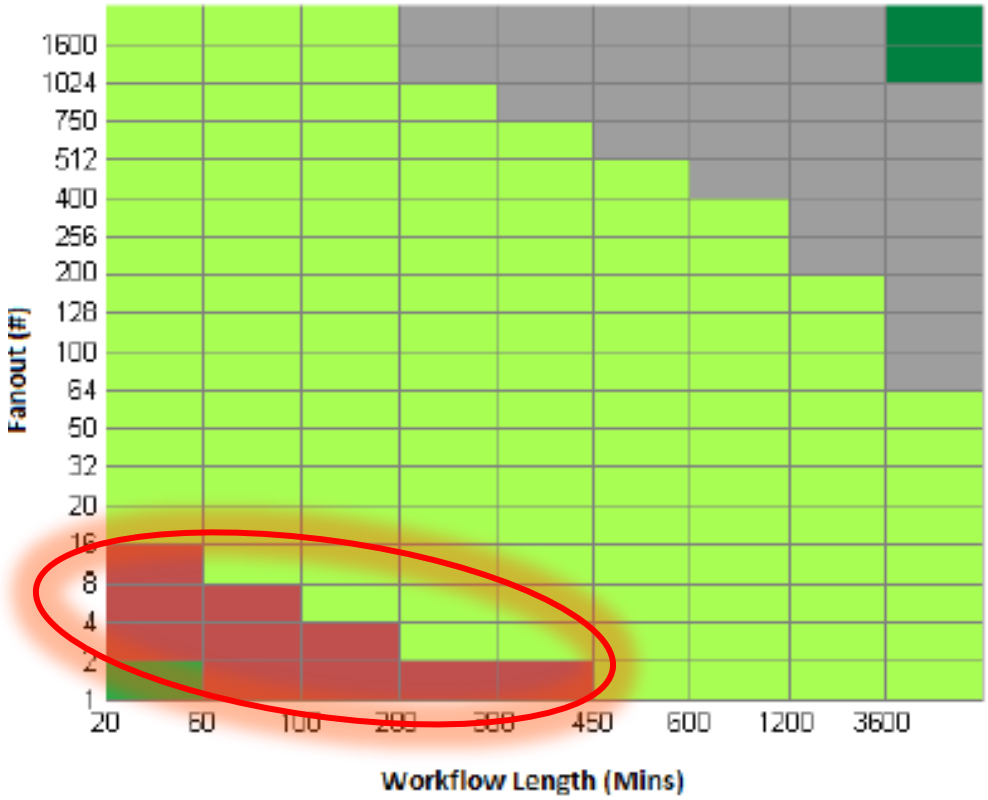
Workflow Parameter Space

- Synthetic workflows created for:

WF Length (L)	20mins – 60hrs
Stage Length	30secs – 6hrs
# of Stages	4, 10, 50, 100
WF Width (W)	1 – 1600
Mincores per Stage	1, 0.25×W, 0.50×W, 0.75×W, W

Effect of Total Workflow Length

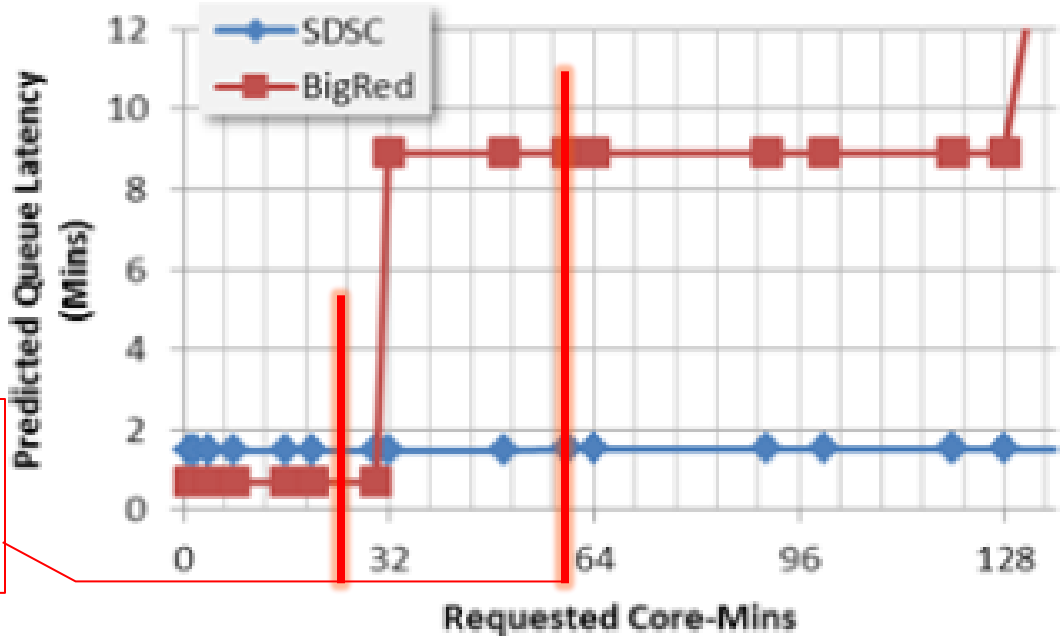
- 10 Stage Workflows
  - Length in mins ↑ on X axis, Fanout width ↑ on Y axis
- Green shades indicate same platform selected by BReW & Whitebox



- Consistent selection for middle region (Lime)
- Poor for large width/length (Gray)
  - Lack of QBETS information
- Mixed for small width, length (Red)
  - HPC behavior

QBETS times for SDSC, BigRed

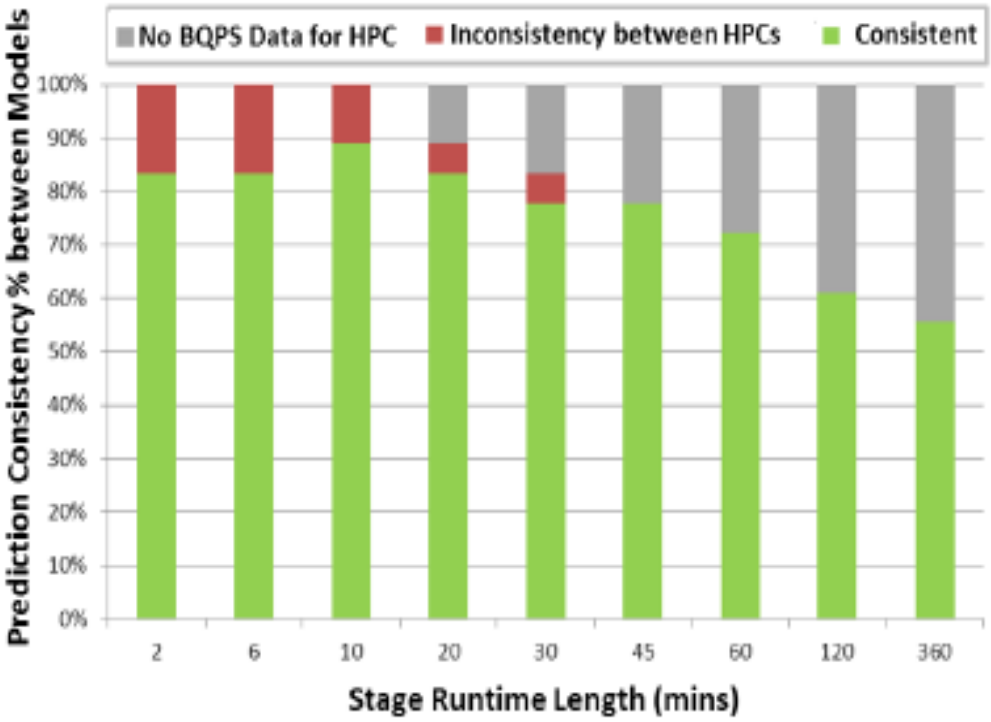
- HPC Queue times cross-over past 32 core-mins
- Coarse grained BReW information causes over estimation of core-mins



2 stage WF,
30min/stage,
1 core/stage

Effect of Length per Workflow Stage

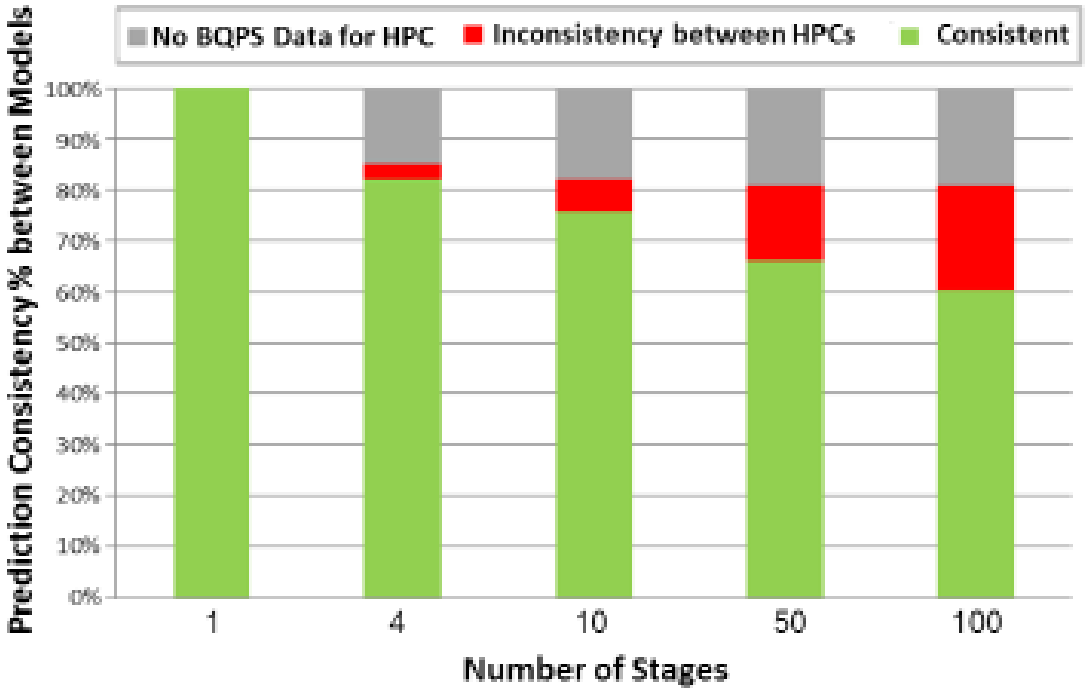
- % of prediction consistency for 162 workflows
- Length per stage from 2min-6hrs for constant WF length



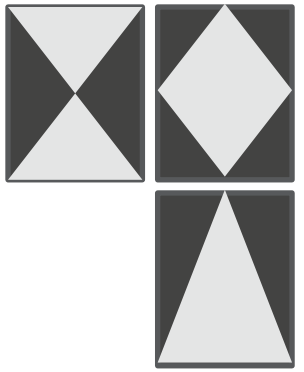
- Small stage lengths ≤ 20min are good @ 80%+
  - Peak @ 95% for 100mins: 4 stage, 25mins/stage
  - HPC errors for very small stages
- No QBETS data for large stage lengths (Gray)

Effect of # of Stages in Workflow

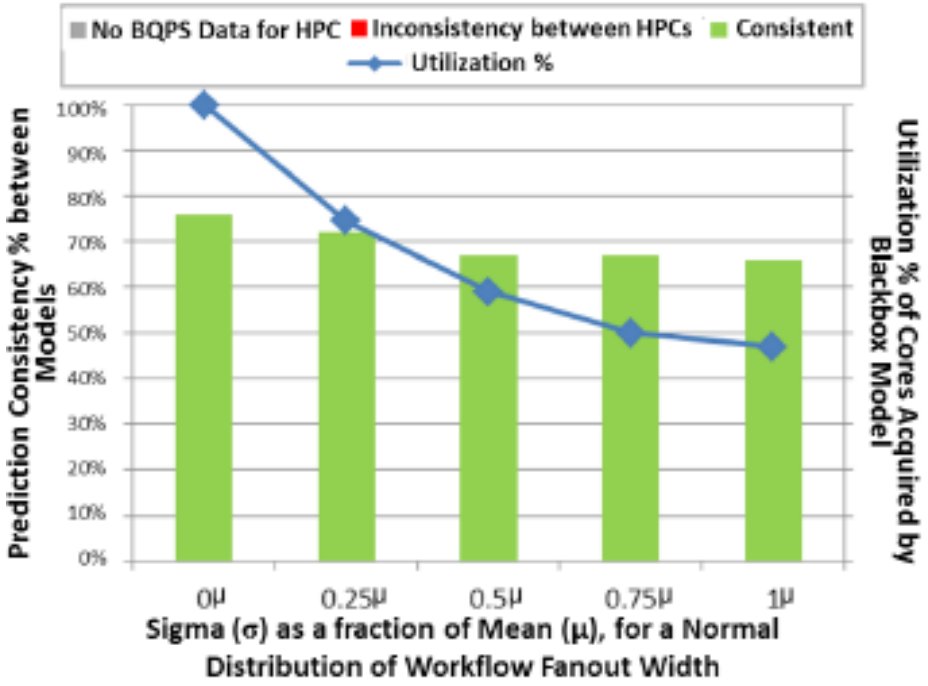
- 810 workflows run with 1-100 stages
  - 1 Stage whitebox ~= Blackbox model
- Both HPC and QBETS errors are seen to increase
- Latency errors accumulate at each stage



Effect of Workflow Width Variability



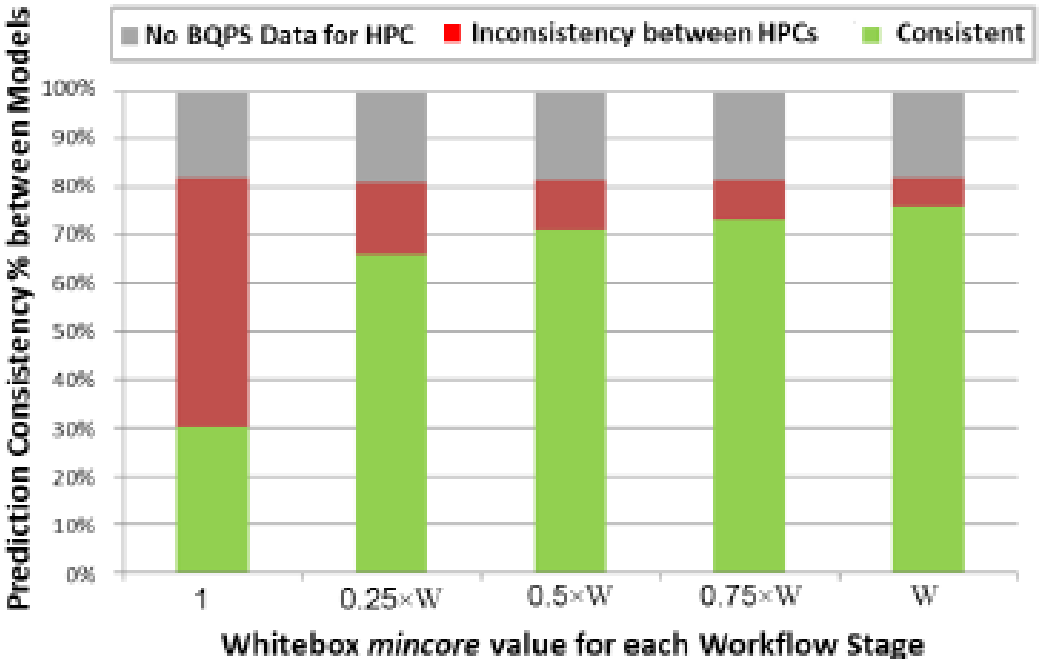
- Non-uniform fanout across stages
- Use normal distr for varying stage width:
  - μ±3σ for different σ; Use max for BReW
- Plot consistency and resource utilization



- Increasing variability has small impact on consistent platform prediction
- But, prediction accuracy of resources required reduces

Effect of MinCore per Workflow

- MinCore decides if tightly or loosely coupled
  - 1 = loose, w = tight
- *BReW model always uses tightly coupled*



- For MinCore=1, whitebox get core at a time. So drastic variation in latency.
- For BReW using static MinCore=1, difference is lesser

eScience Workflows

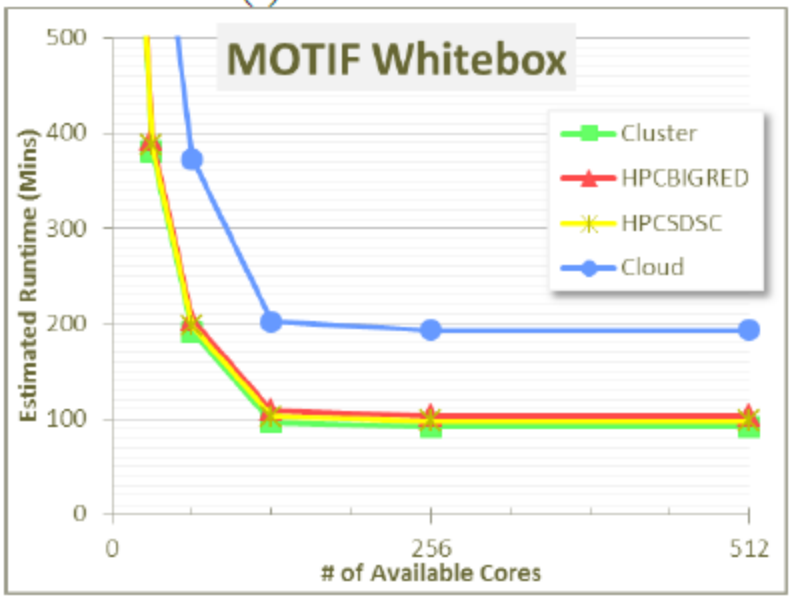
- Space of eScience workflows evaluated

	Domain	Length (h:m:s)	Width	Stages	Data In/Out
Motif	Genomics	1:31:30	135	3	13M/1.2G
Montage	Astronomy	0:06:10	662	9	700M/1.5M
MODIS	Environ Sci.	0:29:40	60,000	4	400G/1G
GWAS	Comp. Bio.	0:19:00	1100	7	150M/10M

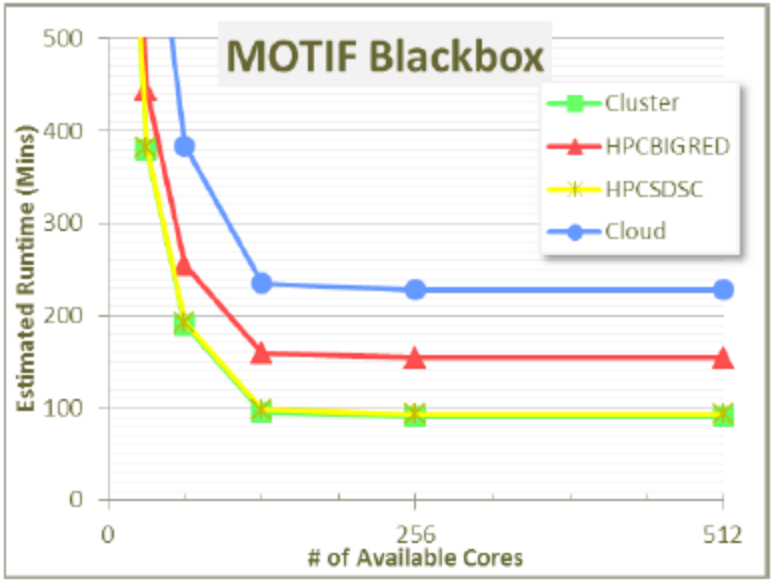
eScience Workflows: Motif N/W

- Goal is to perform same platform ordering for different resource availability
- Motif: 3 stage, 135 tasks, long running >1hr
- Variations between platforms large enough to order
  - Even absolute values are similar

(a) MOTIF Whitebox



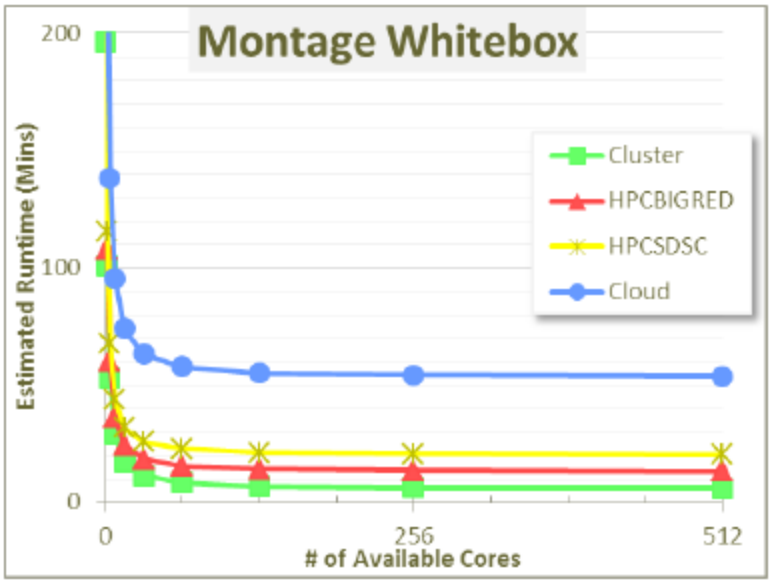
(b) MOTIF BReW Blackbox



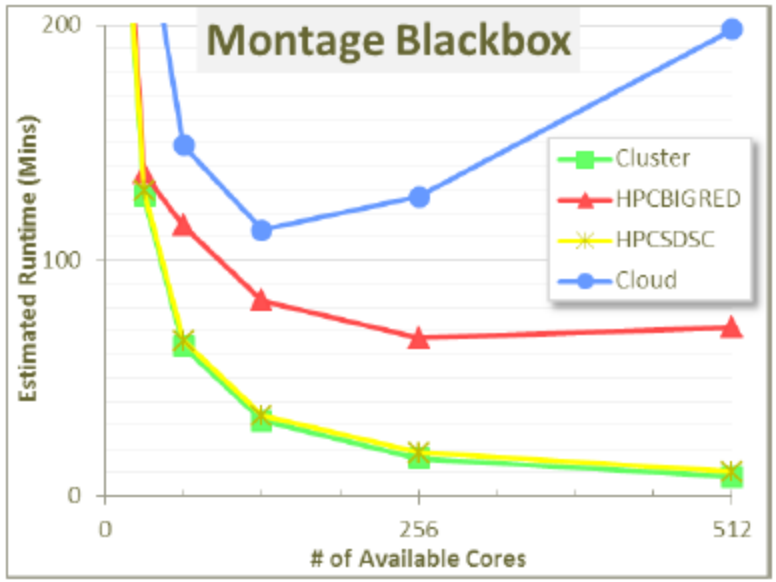
eScience Workflow: Montage

- 7 short stages <100s, up to 662 fanout
- Consistent ordering except for HPCs
- Azure startup time outstrips performance gains from cores

(a) Montage Whitebox



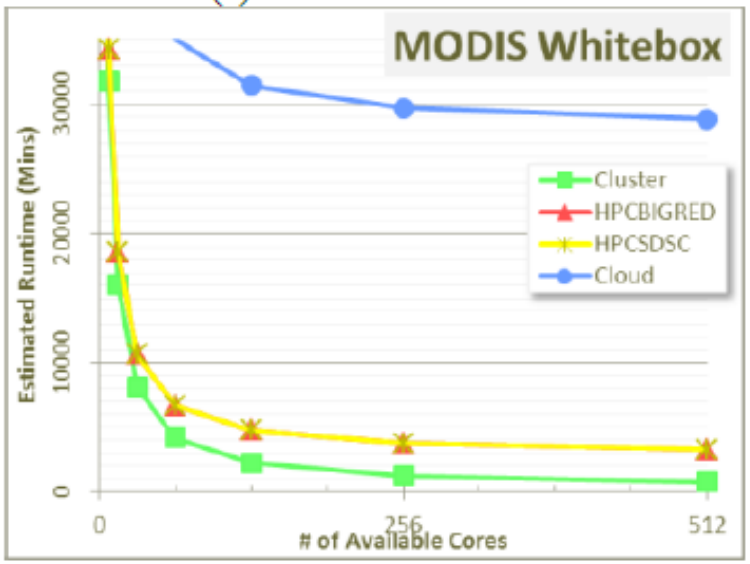
(b) Montage BReW Blackbox



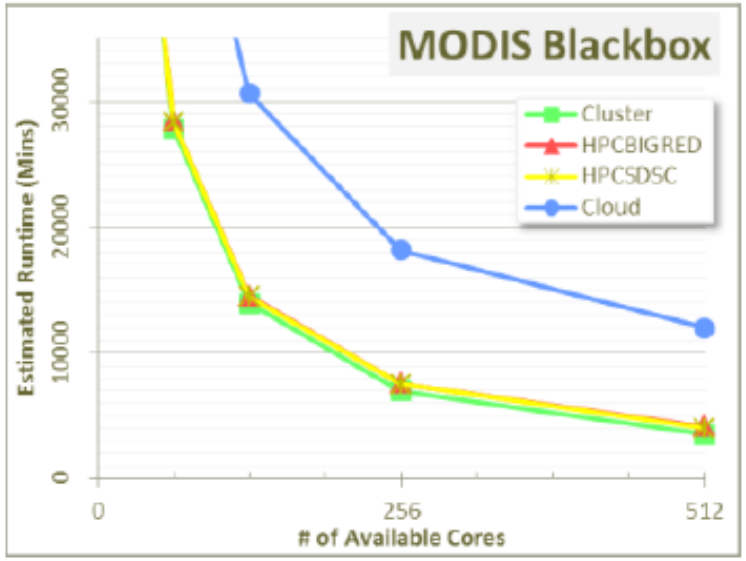
eScience Workflow: MODIS

- 4 stage, data parallel, 400GB, 60K tasks
- Consistent predictions
- Data transfer dominates
  - Whitebox intermediate transfers costs are high
- BReW compute prediction high

(a) MODIS Whitebox



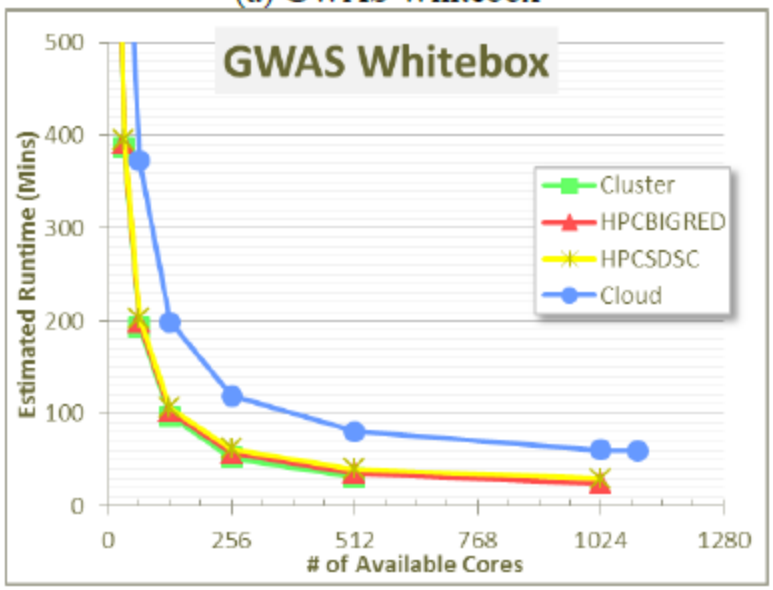
(b) MODIS Blackbox



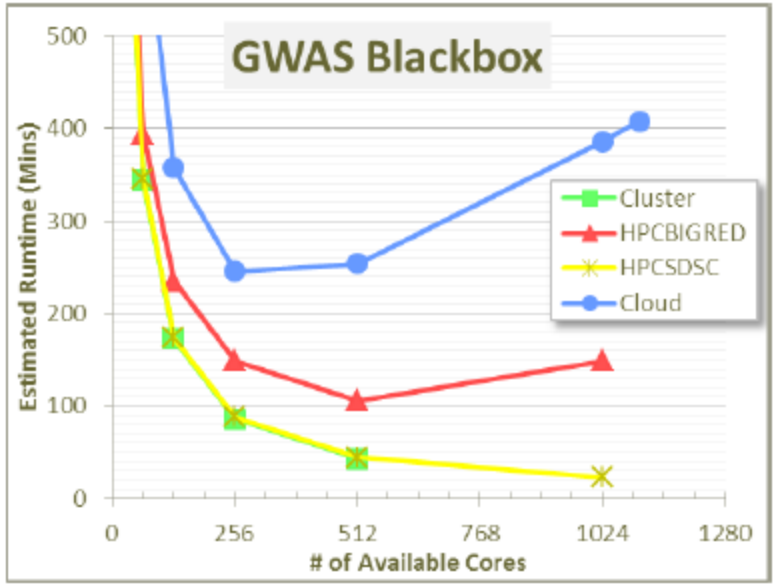
eScience Workflow: GWAS

- 2 MapReduce stages of 1100, 150 tasks
  - compute intensive, small data, <10min tasks
- Short runtimes cause HPC errors
- BigRed cross at 1024 cores; Azure at 256 cores

(a) GWAS Whitebox



(b) GWAS Blackbox



Conclusions & Future work

- Runtime estimated from Blackbox suitable for several classes of workflows for relative comparison
  - Absolute values vary
- Queue latencies have major impact on selections
  - Azure linear, HPC step times

- Detect suitability of workflows for selection
- More complex workflows, whitebox scheduling
- Other WF features that have impact
  - Should min required cores be part of BReW?
- Impact of HPC policies

CFP: IJCA Special Issue on
Scientific Workflows, Provenance and Their Applications
Guest Eds: Artem Chebotko, Yogesh Simmhan, and Paolo Missier
Submission deadline: **December 15, 2010**
Notification of acceptance: April 1, 2011
www.cs.panam.edu/~artem/ijca

CFP: ScienceCloud2011: 2nd Workshop on Scientific Cloud Computing co-located with ACM HPDC 2011, San Jose CA – June 8th, 2011
Chairs: Ioan Raicu, Pete Beckman, Ian Foster, Yogesh Simmhan
Abstract Due: **January 25th, 2011**
Papers Due: February 1st, 2011
www.cs.iit.edu/~iraicu/ScienceCloud2011

Job Opening: Post Doctoral Research Associate
Center for Energy Informatics, Viterbi School of Engineering, University of Southern California
Information integration, data management, Semantic web, Software systems integration, Social and/or Scalable computing in Smart Oil Fields
cei.usc.edu/jobs

Thank you!

Questions

Yogesh Simmhan

Center for Energy Informatics, Viterbi School of Engg.

University of Southern California

simmhan@usc.edu

ceng.usc.edu/~simmhan