Comparison of Resource Platform Selection Approaches for Scientific Workflows

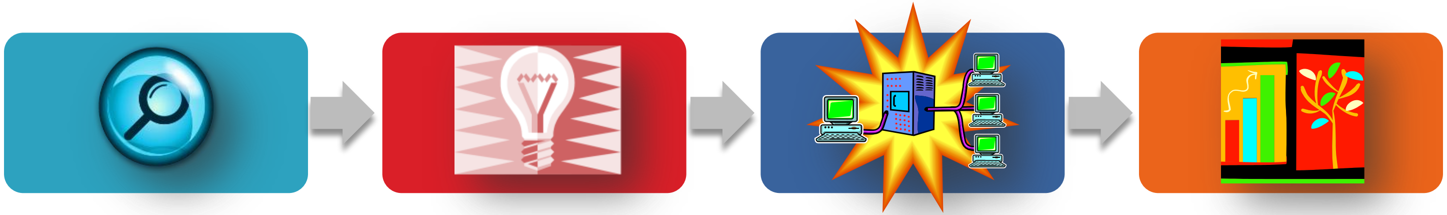
Science Cloud 2010

Yogesh Simmhan, MSR & **Lavanya Ramakrishnan**, LBL

Workflows for Modeling eScience

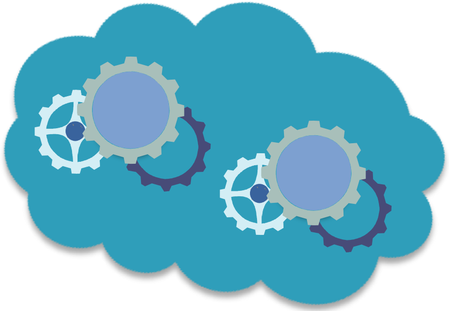
- Workflows common for *in silico* experiments
  - DAGs & dataflows
- Allow easy composition
  - Change often
- Loosely-coupled tasks, Tighly-coupled MPI
  - Different task characteristics

- Compute & Data intensive
  - Challenge of eScience problem sizes
- Resource needs often exceed available ones
  - Scale-out beyond current resources



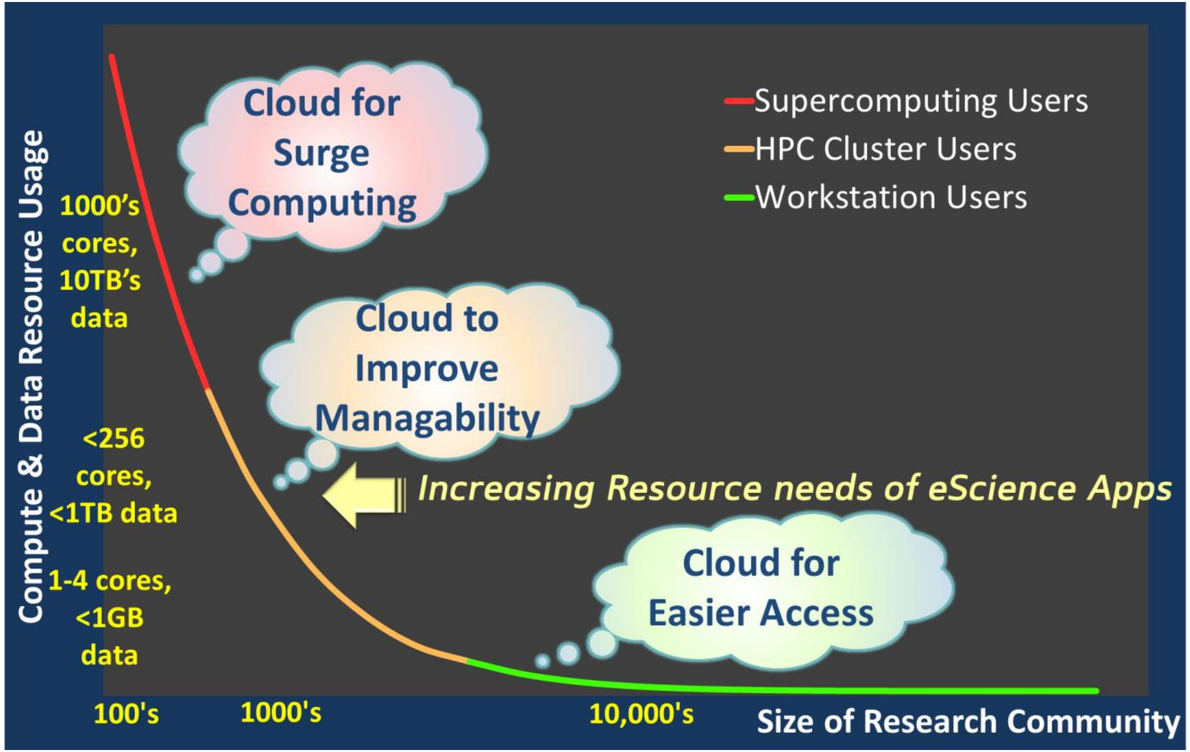
Flourishing Space of Resource Platforms

- Cluster, Cloud, HPC, Desktop
- Local, captive resources
- Batch systems
- On-demand platforms
- Different characteristics of resource platforms



Cloud Platform for eScience

- ❑ On-demand
- ❑ Scale out
- ❑ Available
- ❑ Management Ease
- ❑ Economical (TCO)
- ❑ Simple APIs



Resource Selection for Workflows

- Scientists need to select from existing & emerging resources
- *Ad hoc*, Rule of thumb, based on familiarity
  - Can be sub-optimal, punitive
- Different characteristics of resource platforms
  - Dynamic over short, long term
- Different goals
  - Makespan, usage, co$t

- DAG Scheduling Algorithms
  - Tasks/WF scheduled to a platform
- Automatic WF Scheduling
  - Pegasus, Swift, Trident, etc. schedule WFs to remote resources
  - Support various platforms: Cluster, HPC, EC2.
  - Mandal, et al. Perf-based advance reservation for GrADS
  - Batch Queue Prediction Service
  - Blythe, et al. Task level greedy algorithm v. WF level optimization

Resource Selection for Workflows

- These need information about WFs
  - Structural, task level details, data flow
- Fine grained details hard to determine & specify
  - Provenance mining, perf models
- Different granularity of WF details
  - Blackbox, Graybox, Whitebox

- *Fine grained workflow specs, evolving resource platforms pose overhead for users*

Hypothesis —

- □ *Can we make intelligent resource platform selection with limited workflow information?*
  - Length
  - Width
  - Data In/Out

- □ *What are trade-offs of running applications on different platforms?*

Overview

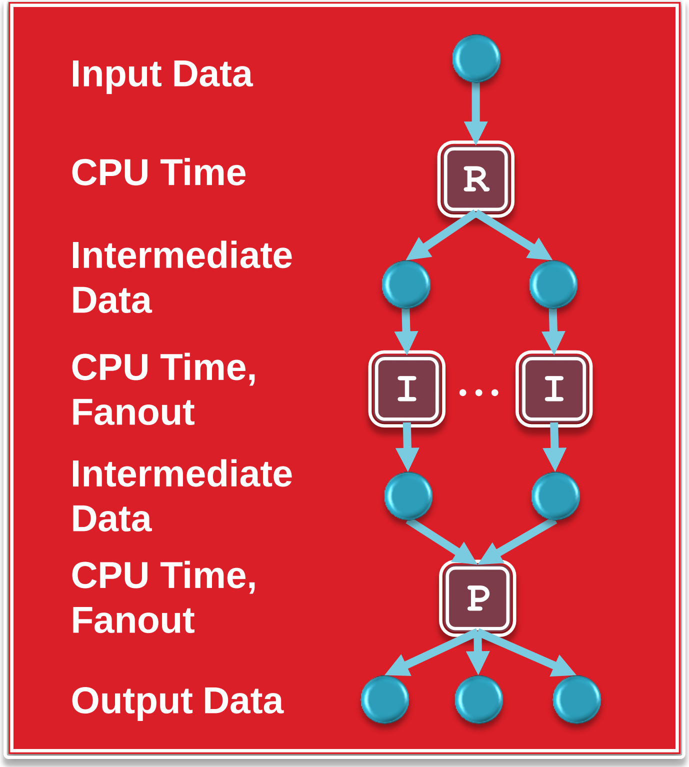
Workflow Characteristics

Structural Information

- **Pattern**: Sequential, Fork-Join, Control flow
- **Length**: # of stages, length per stage, total length
- **Width**: Fanout

Resource Usage

- **Data**: In/Out
- **Compute**: Cores required



Input Data

CPU Time

Intermediate Data

CPU Time, Fanout

Intermediate Data

CPU Time, Fanout

Output Data

Resource Platforms & Characteristics

Desktop

- Full application control
- Growth of multi-core
- eScience beats Moores Law

Cluster

- Small-Mid Clusters (~256 core)
- Under-subscribed, instant use
- Large science apps don't fit

HPC

- Shared, national centers
- Large # of cores (>1000)
- Over-subscribed queues, policies

Cloud

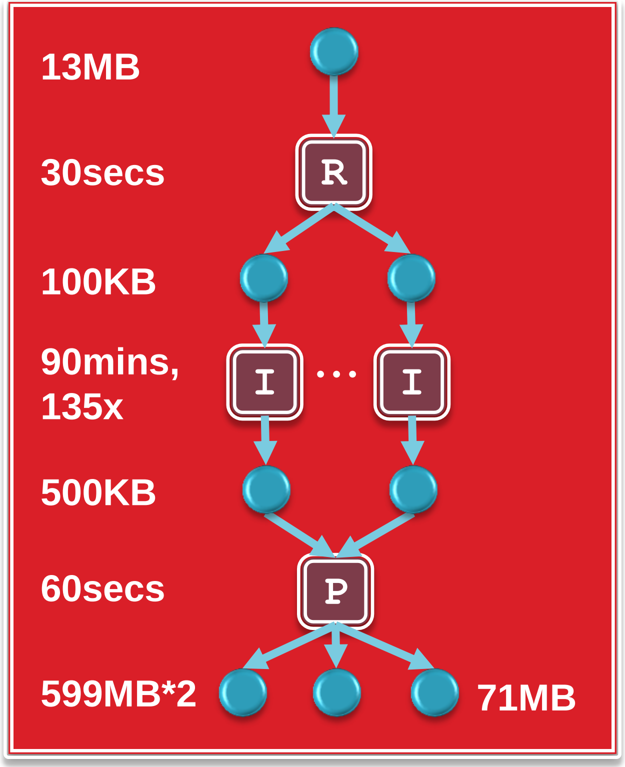
- Infra. & Platform as Service
- On-demand, customizable
- Virtualization impact, Bandwidth

● Available cores ● Queue/VM Latency ● N/W Bandwidth ● Core Speed

Platform Selection for Workflows

Whitebox Selection (*Fine Grained*)

MOTIF WORKFLOW



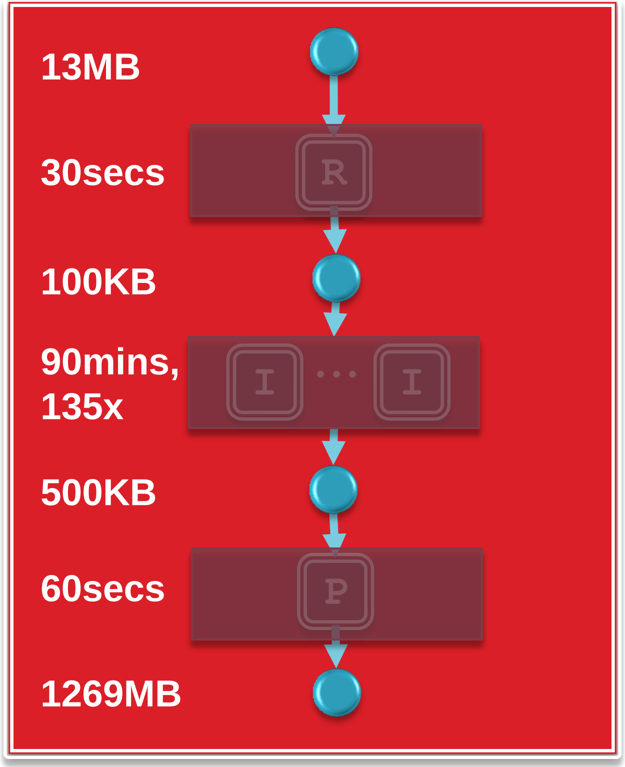
- Full workflow & *Task* details available
- Total runtime due to ■ CPU Time ■ I/O data transfer ■ Queue/VM Overhead
- Time is from by each independent task

$$F_{WorkflowTime} = \sum_{i=0}^{\#\,stages} F^{i}_{StageTime}$$

$$F^{i}_{StageTime} = T^{i}_{OverheadOne} + T^{i}_{Data} + \frac{\left(T^{i}_{TaskLength} \times N^{i}_{TaskWidth}\right)}{N^{i}_{Cores}}$$

Graybox Selection (*Hybrid*)

MOTIF WORKFLOW



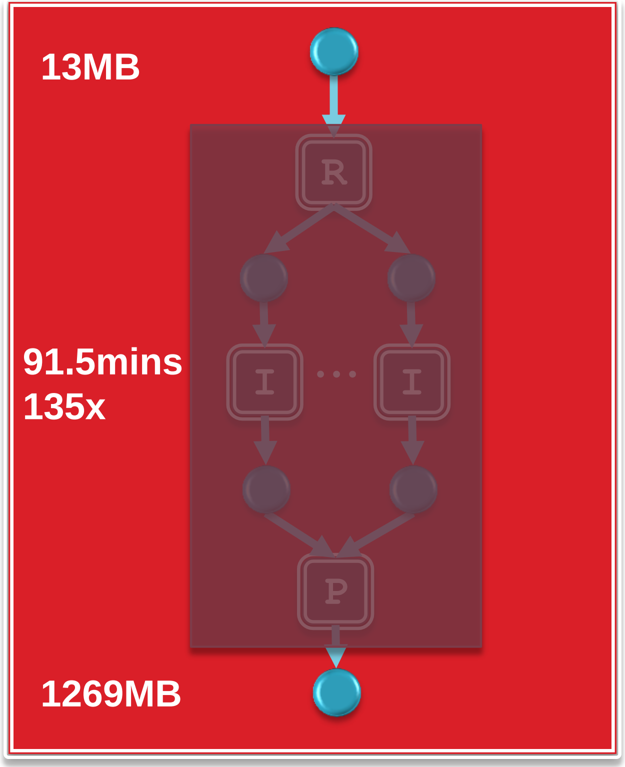
- Workflow *Stage* details available
  - Stages opaque. No task details available.
- Time is from by each independent stage
- Overhead time only for longest stage
  - Queue/VM times pipelined

$$F_{WorkflowTime} = T_{OverheadMax} + \sum_{i=0}^{\#\,stages} F^{i}_{StageTime}$$

$$F^{i}_{StageTime} = T^{i}_{Data} + \frac{\left(T^{i}_{StageLength} \times N^{i}_{StageWidth}\right)}{N^{i}_{Cores}}$$

Blackbox Selection (Coarse Grained)

MOTIF WORKFLOW



- Only *Workflow outline* details available
  - Workflow internals opaque
  - Total CPU Time
  - Data xfer at boundary
- Overhead time for entire workflow
  - All required cores for workflow acquired

$$F_{WorkflowTime} = T_{OverheadMax} + T_{Data} + \frac{\left(T_{WorkflowLength} \times N_{WorkflowWidth}\right)}{N_{Cores}}$$

Early Evaluation

eScience Workloads for Evaluation

MOTIF Network Workflow

- Gene regulation dependency networks
- Compute & data intensive
  - 13MB input, 1300MB output
  - 90mins long, 135 task wide
  - 3 Stages: Fork, Compute fanout, Join

GWAS Workflow

- Genome wide association study
- Compute intensive & wide
  - 150MB input, 160MB output
  - 19mins long, 1100 task wide
  - 6 stages: Two compute fanouts 1100 and 150 tasks wide

Resource Platforms for Evaluation

Local Workstation

- 1 Core, 2.5GHz
- All data local

Local Cluster

- Up to 256 cores of 2.5GHz
- Data remote on client
- 1Gbps LAN bandwidth

Teragrid HPC Clusters

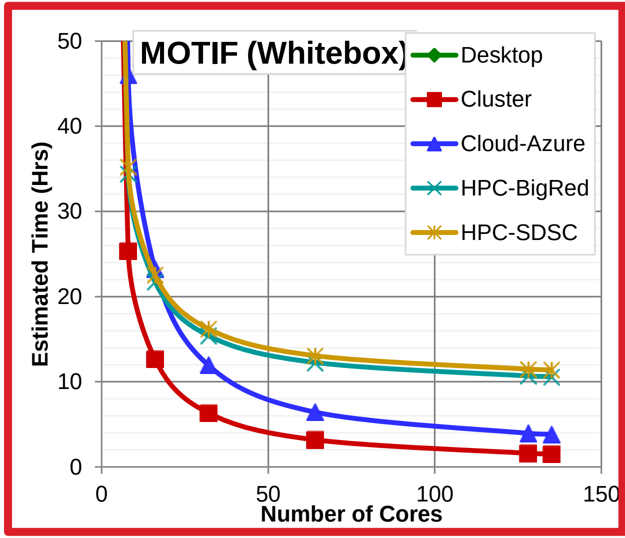
- SDSC & BigRed clusters
- 1 – 2048 cores of 2.5GHz
- NWS Batch Queue Prediction Service (95% Quantile)
- Data remote. 10Mbps WAN.

Azure Cloud

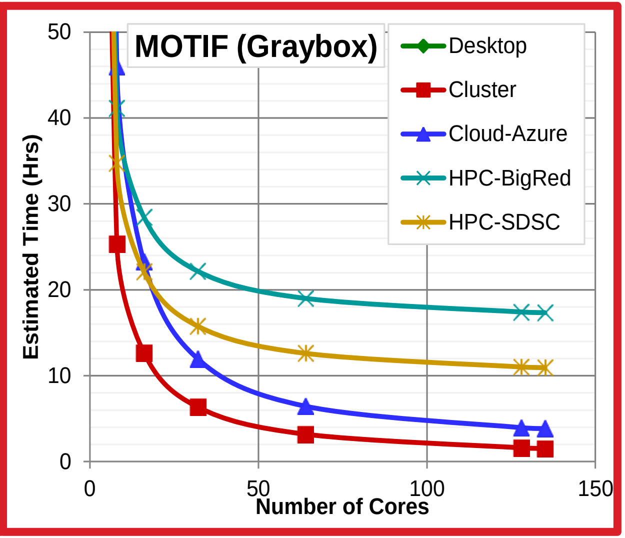
- Small VM, 1 core, 1.6GHz
- VM start time ~$200 + 20c$ secs
- Data remote. 10Mbps WAN.

Results: *Motif Workflow*

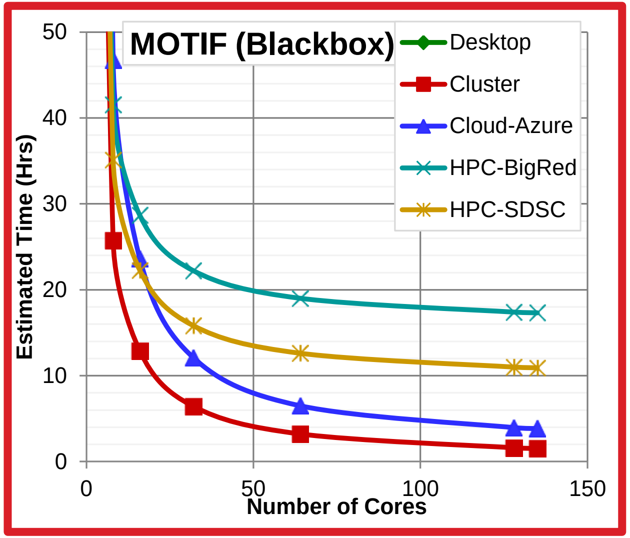
- Black & Graybox **ordering** of platforms **same** for different # cores
- Black & Whitebox **ordering *similar*** … except for the two HPC’s



Whitebox Estimate



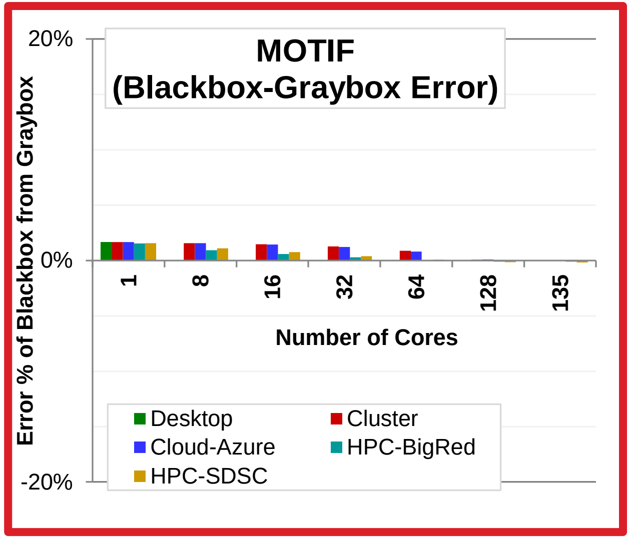
Graybox Estimate



Blackbox Estimate

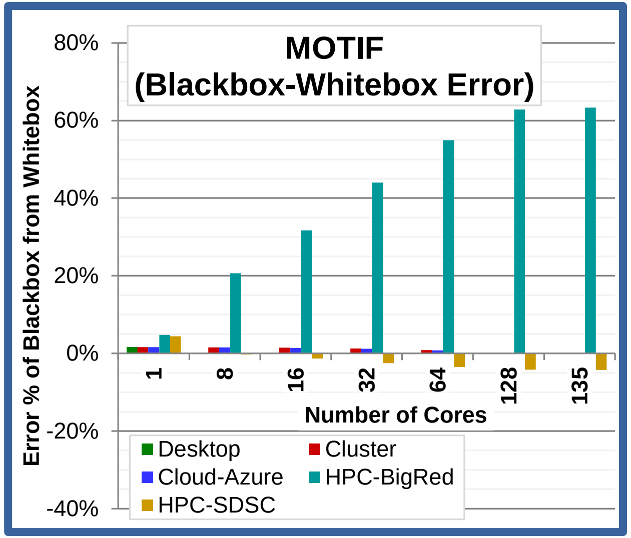
Results: *Motif Workflow*

Black & Graybox absolute difference is small across platforms



Blackbox – Graybox Absolute Difference

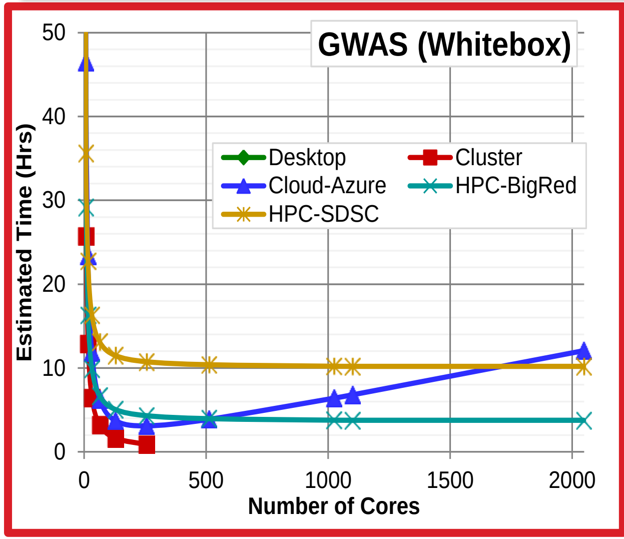
Black & Whitebox absolute difference for BigRed is large



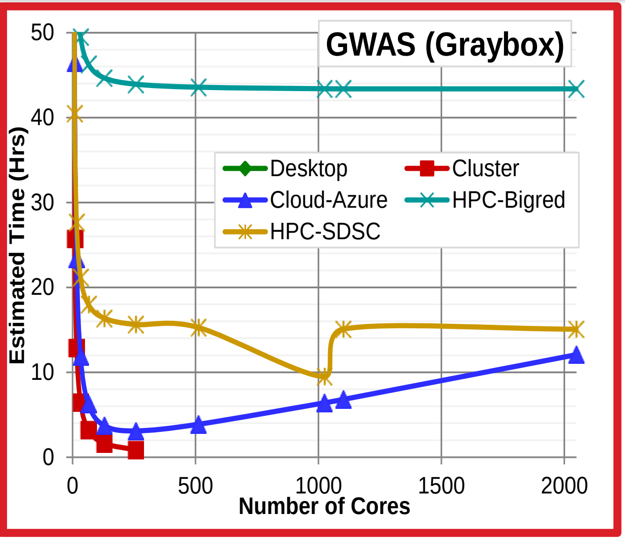
Blackbox – Whitebox Absolute Difference

Results: GWAS Workflow

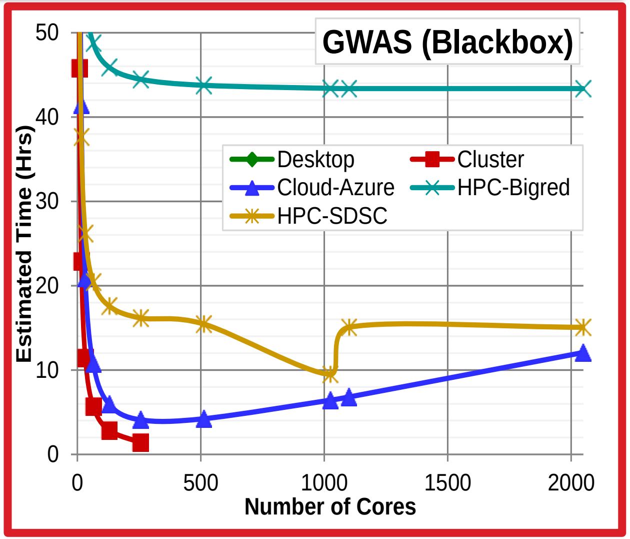
- Blackbox & Whitebox ordering *similar*
  - BigRed: *1 core job queue time* in WBox faster than *width-core job* for BBox
  - SDSC/Azure: Azure linear time; SDSC has step at 1024 cores for *BBox*



Whitebox Estimate



Graybox Estimate



Blackbox Estimate

Conclusion & Future Work

Conclusions & Future work

- Runtime estimated from Blackbox good enough for relative comparison
  - Absolute values vary
- Queue overhead for task v. WF as a job has impact
  - Azure linear, HPC step times
- Graybox ~= Blackbox

- More complex workflows
  - Simulation v. Calculation
- Synthetic workflow runs
  - Effect of each workflow attribute on estimate
- Other WF features that have impact
  - E.g. Min required cores per stage

Thank you!

Questions

Comparison of Resource Platform Selection Approaches for Scientific Workflows

Yogesh Simmhan, Microsoft Research

Lavanya Ramakrishnan, Lawrence Berkeley Lab

Science Cloud Workshop, 2010

Acknowledgements

Dept of Energy | C. van Ingen, R. Barga, J. Listgarten **MSR** | K. Jackson, D. Agarwal **LBL** | E. Soroush **UW**