Presentation about: Rodrigo Álvarez, G.P, Elmroth, E., Östberg, P.O., Ramakrishnan, L. Enabling workflow aware scheduling on HPC systems. 26th International Symposium on High-Performance Parallel and Distributed Computing (HPDC 2017)

Enabling Workflow-Aware Scheduling on HPC Systems

HPDC'17

Gonzalo P. Rodrigo Álvarez
gprodrigoalvarez@lbl.gov

Open Source patch for slurm **available at:**
http://frieda.lbl.gov/download

U.S. DEPARTMENT OF ENERGY | Office of Science

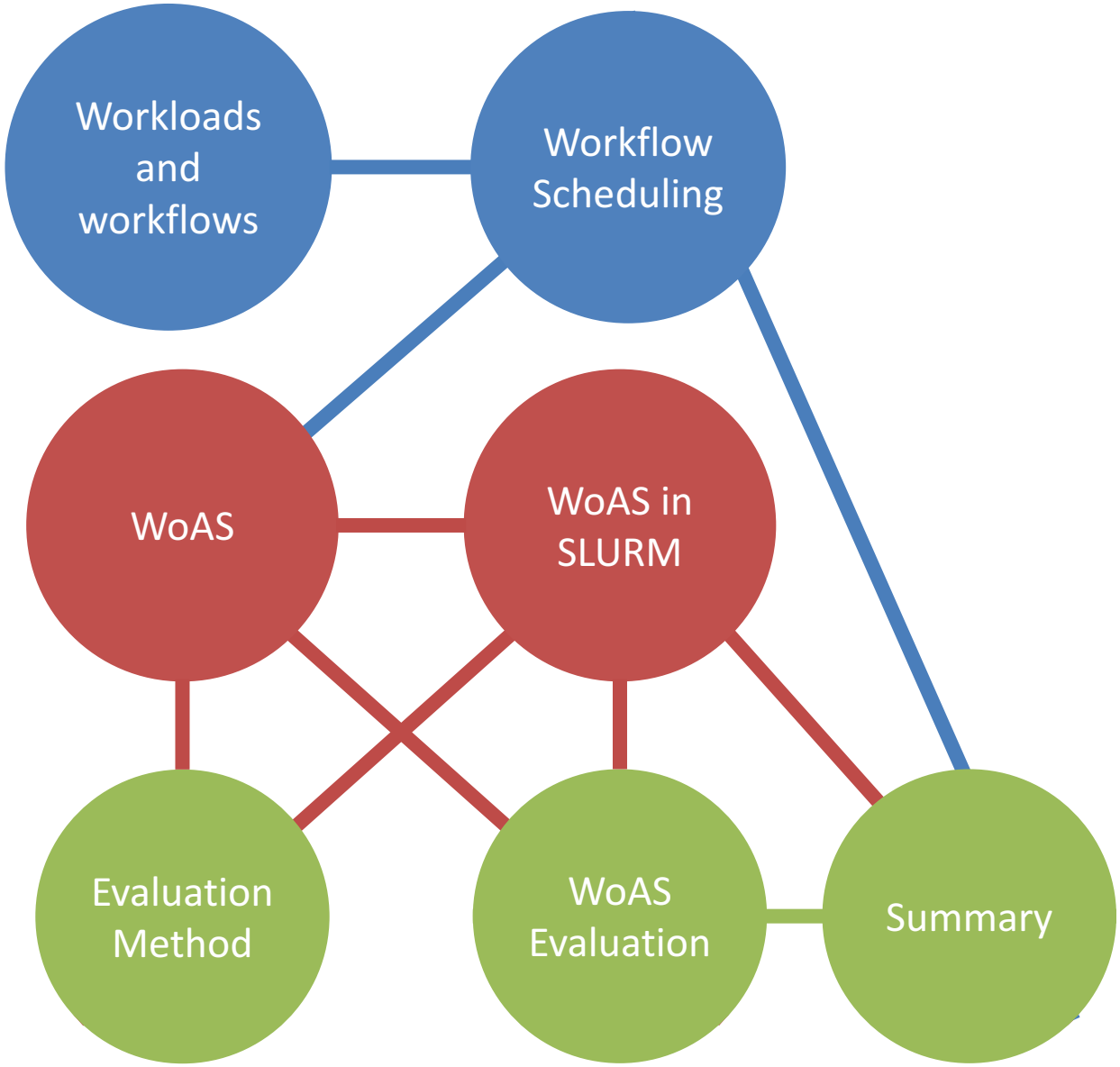
Outline

Introduction

Contributions

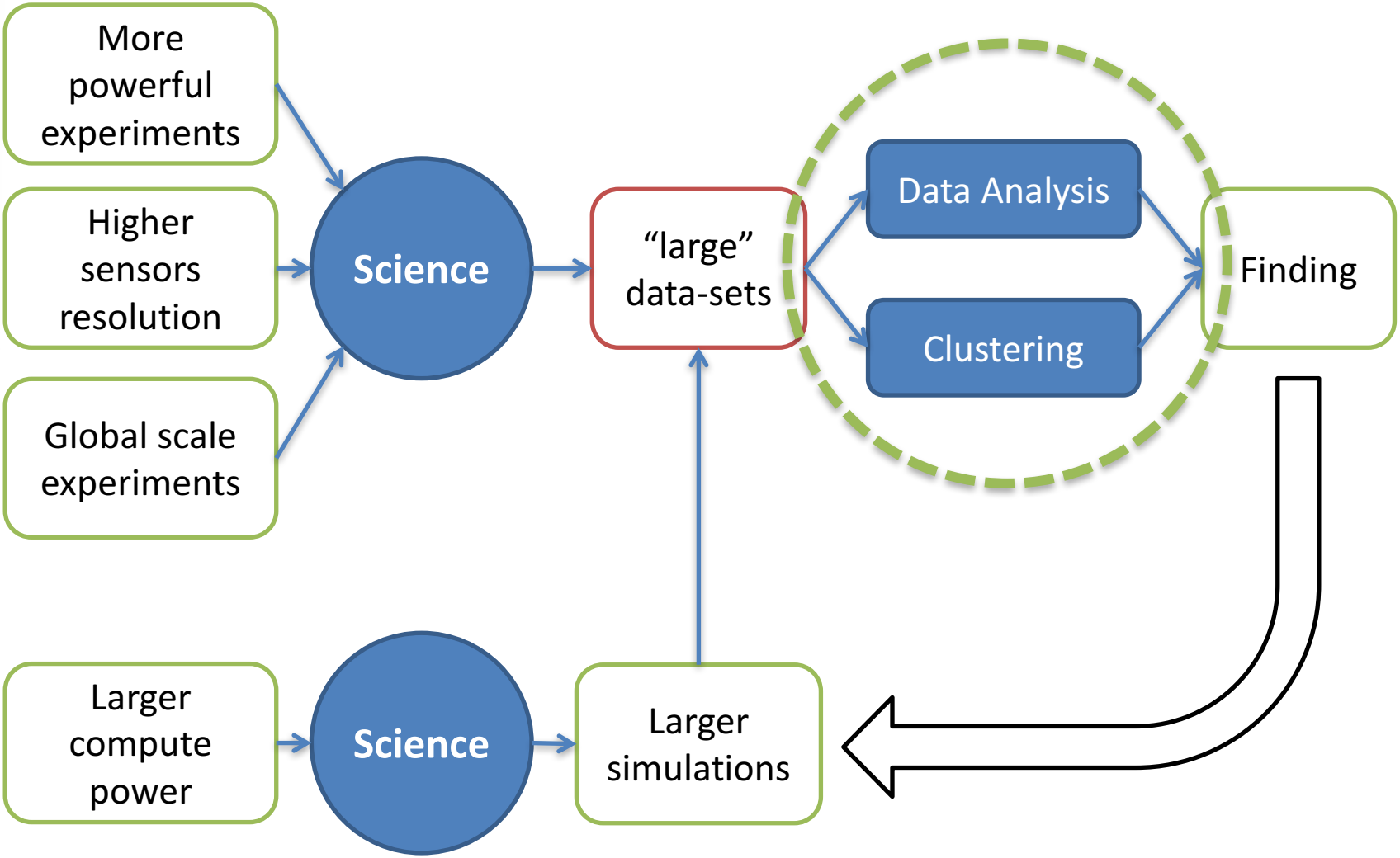
Workflow Aware Scheduling

Evaluation

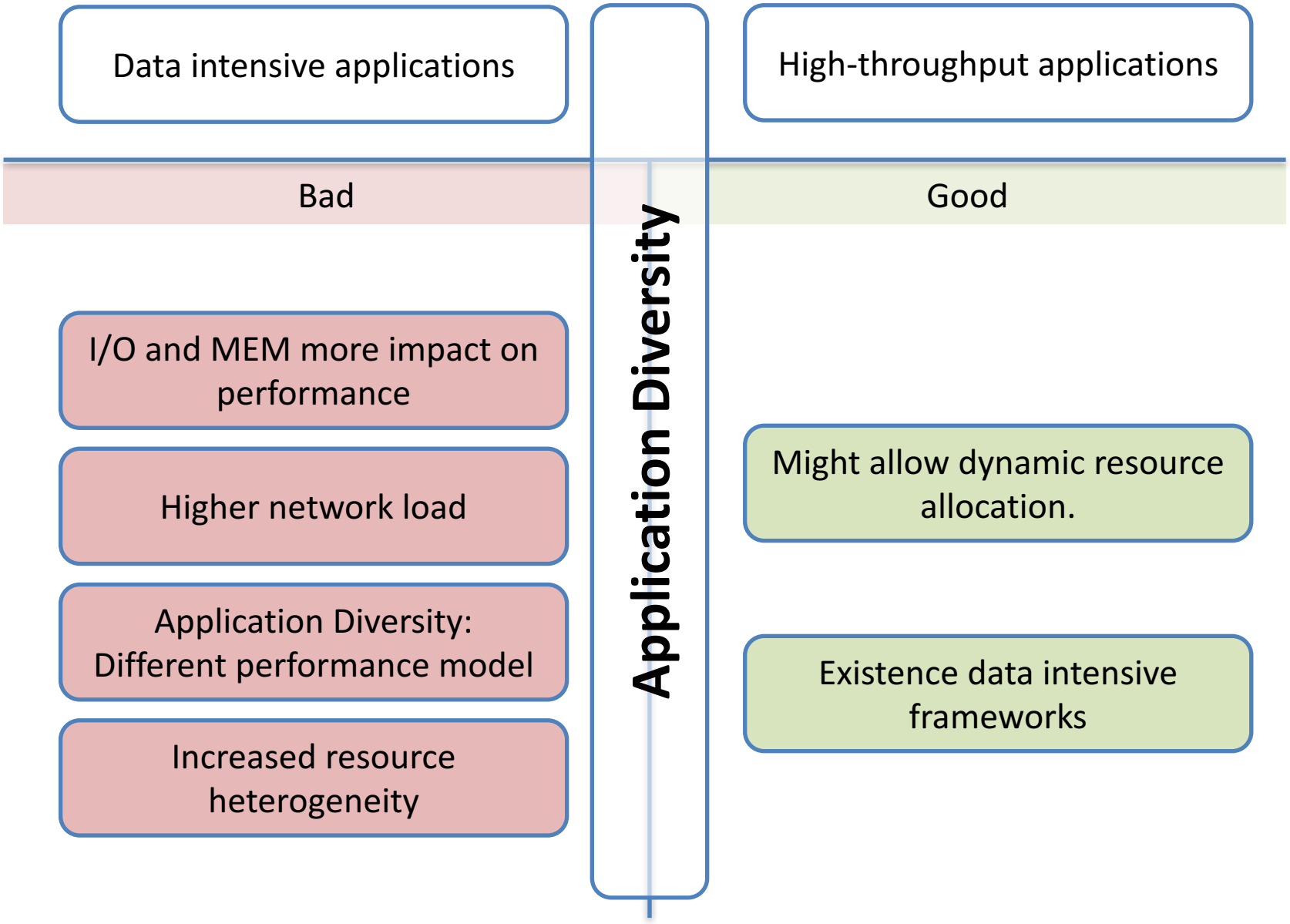


 gprodrigoalvarez@lbl.gov

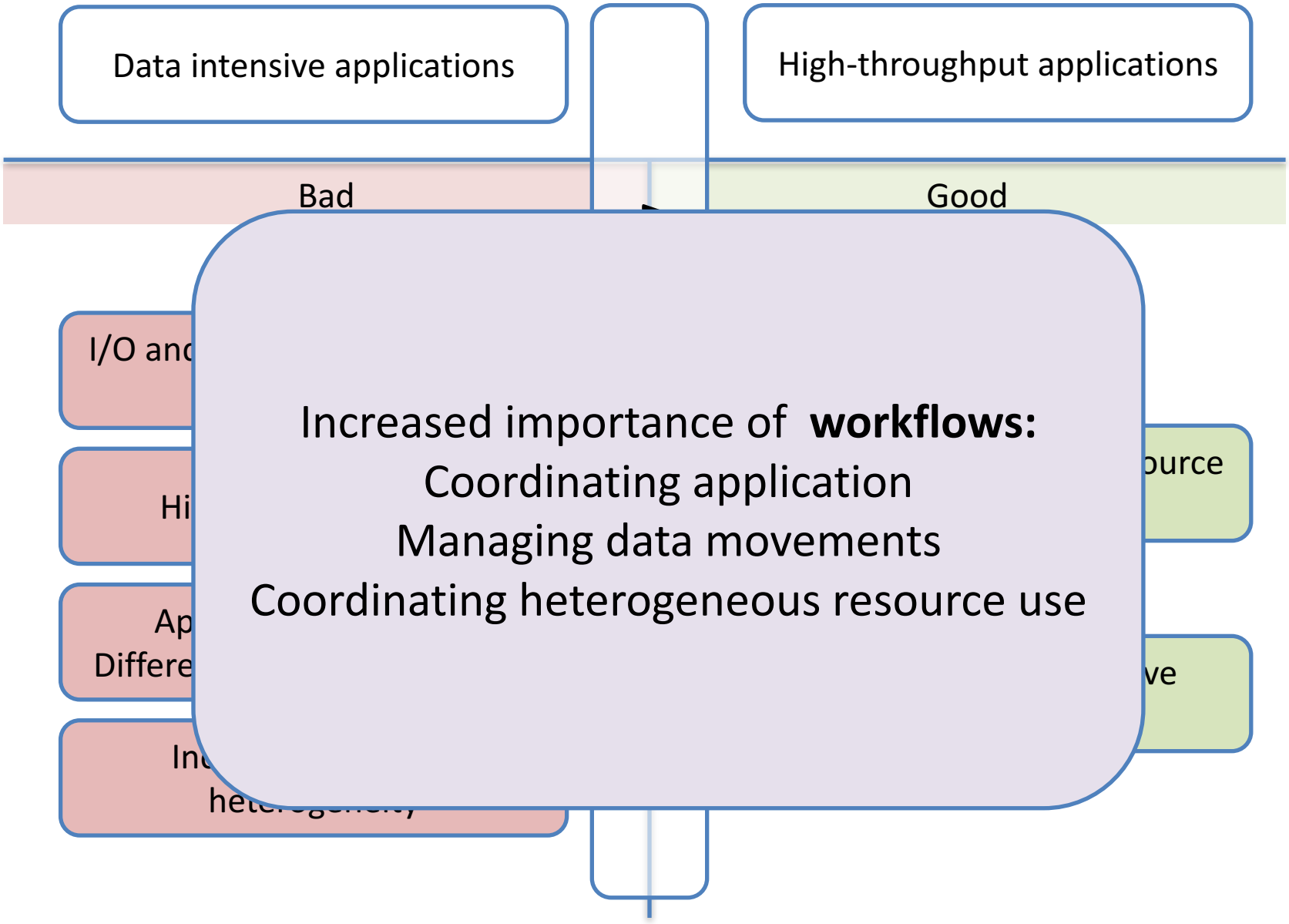
Welcome to the 4th Paradigm of Science: Big Data



Data more important in HPC workloads

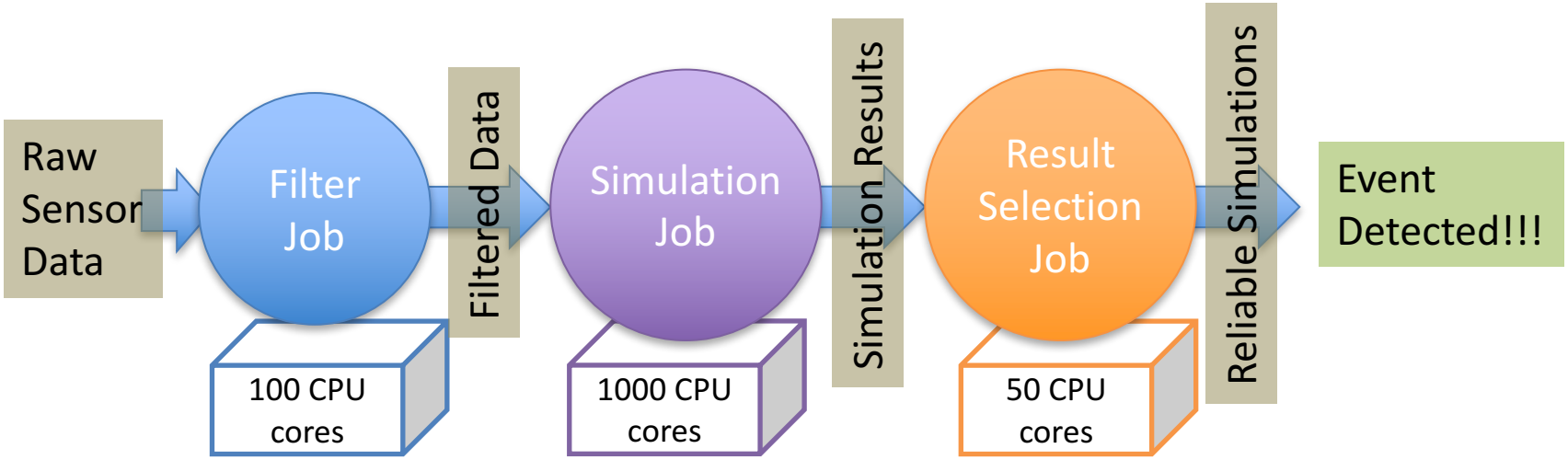


Data more important in HPC workloads



What is a workflow?

"… a composition of jobs with data or control dependencies…"

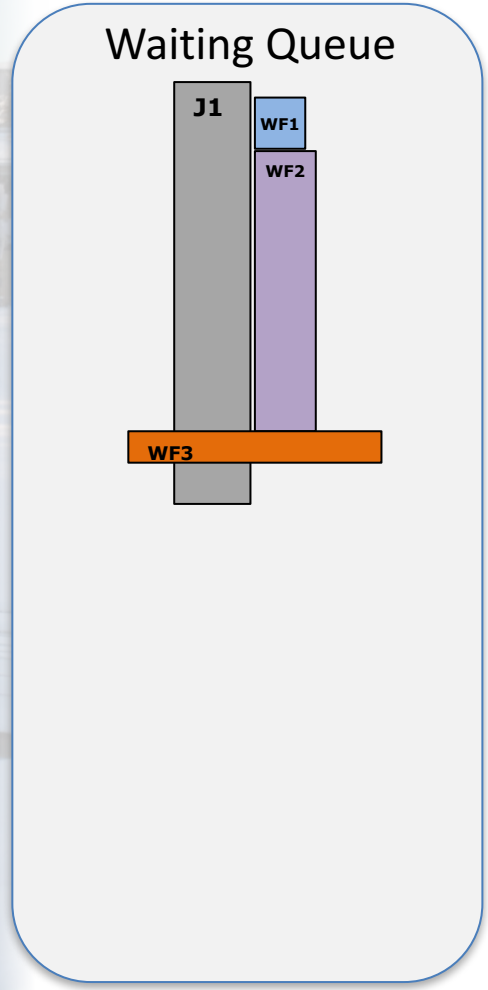


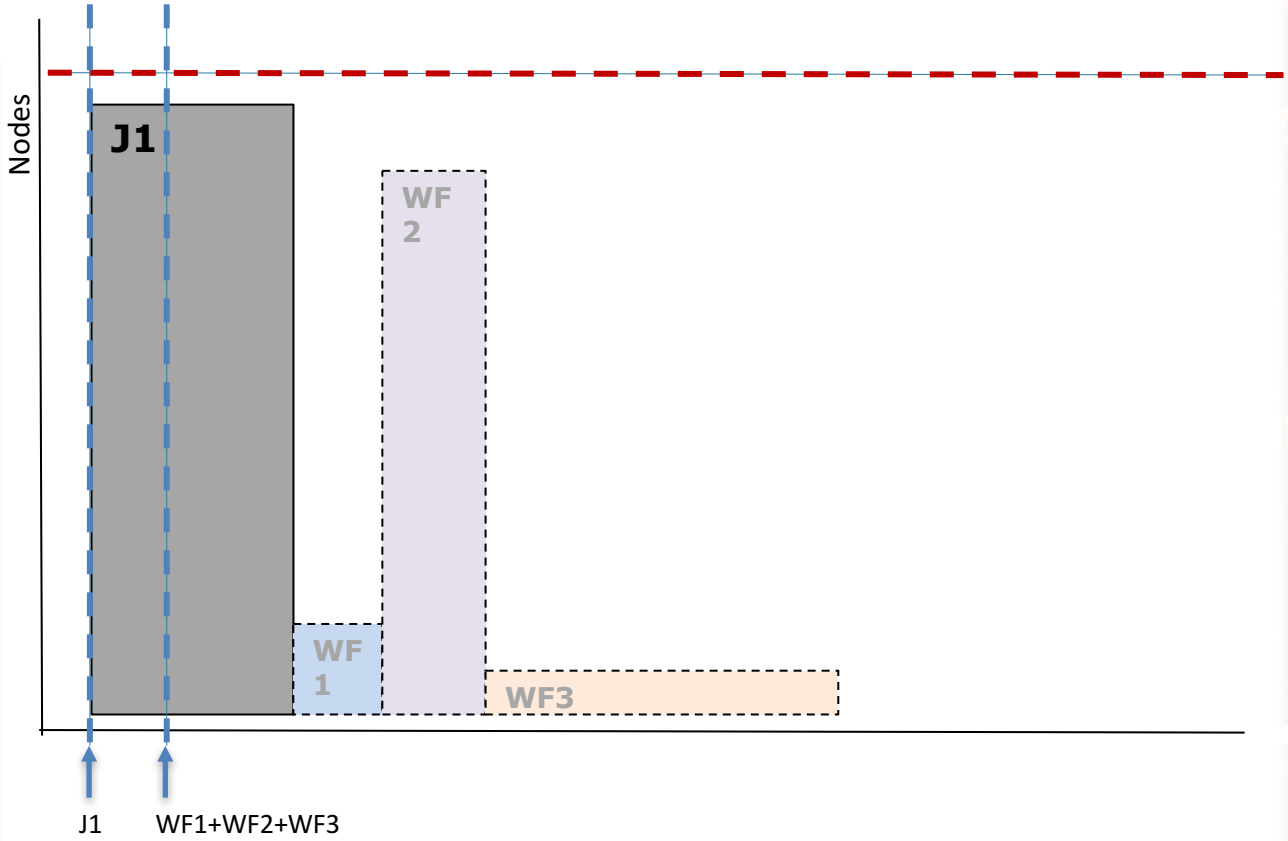
Generic workflow for a sensor in a live experiment

Workflows and HPC Schedulers

Schedulers are not aware of workflows

"Chained Jobs"/Wait Approach

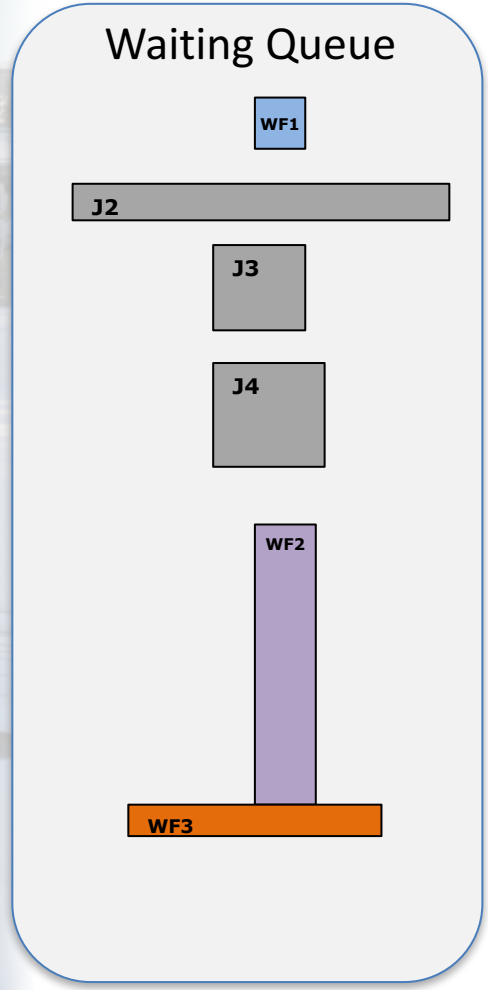


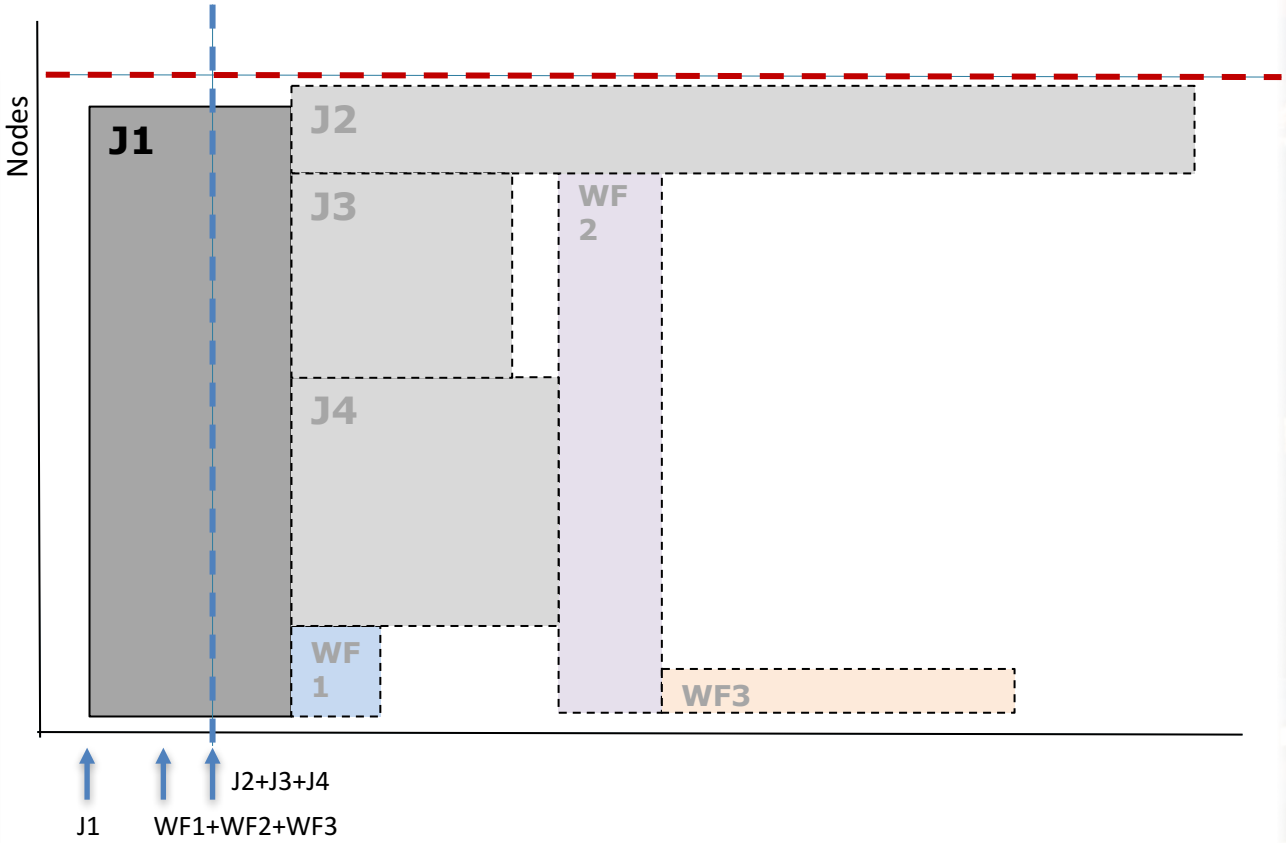


Workflows and HPC Schedulers

Schedulers are not aware of workflows

"Chained Jobs"/Wait Approach

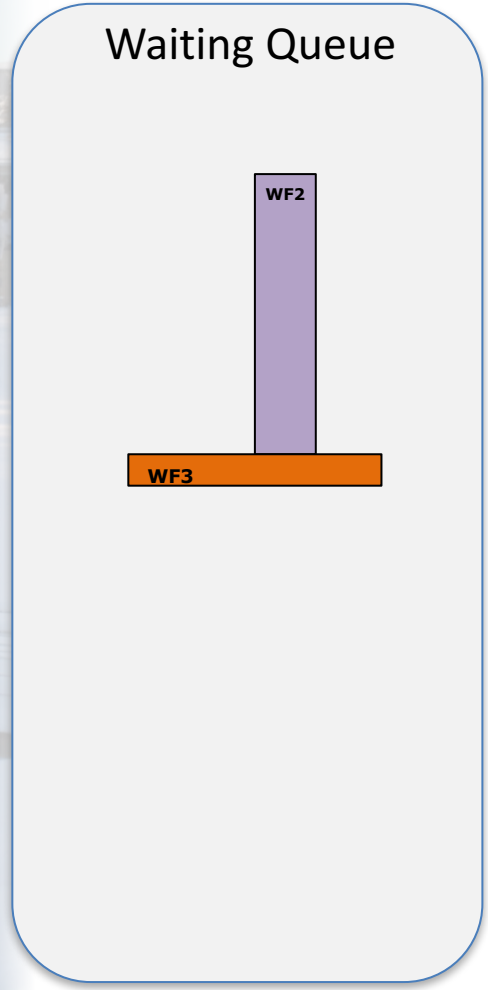


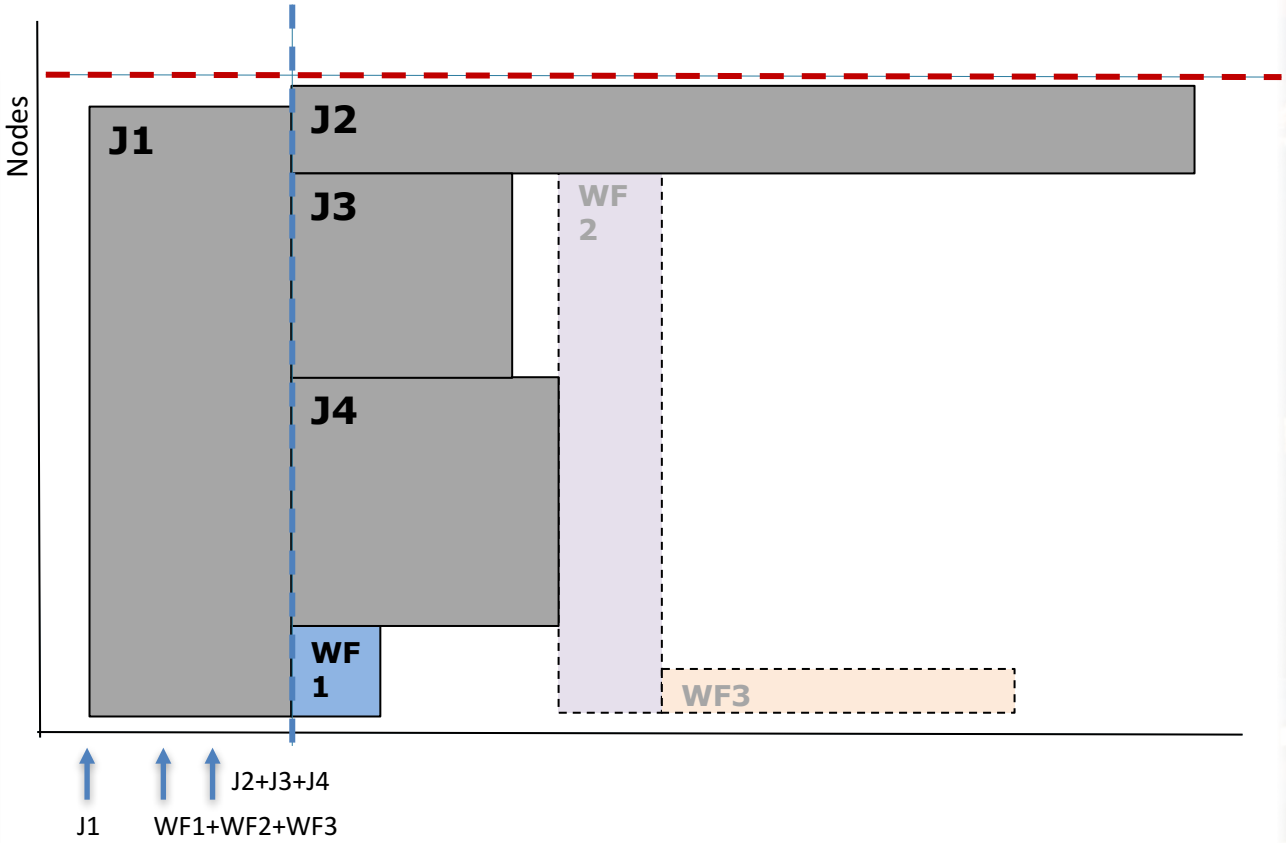


Workflows and HPC Schedulers

Schedulers are not aware of workflows

"Chained Jobs"/Wait Approach

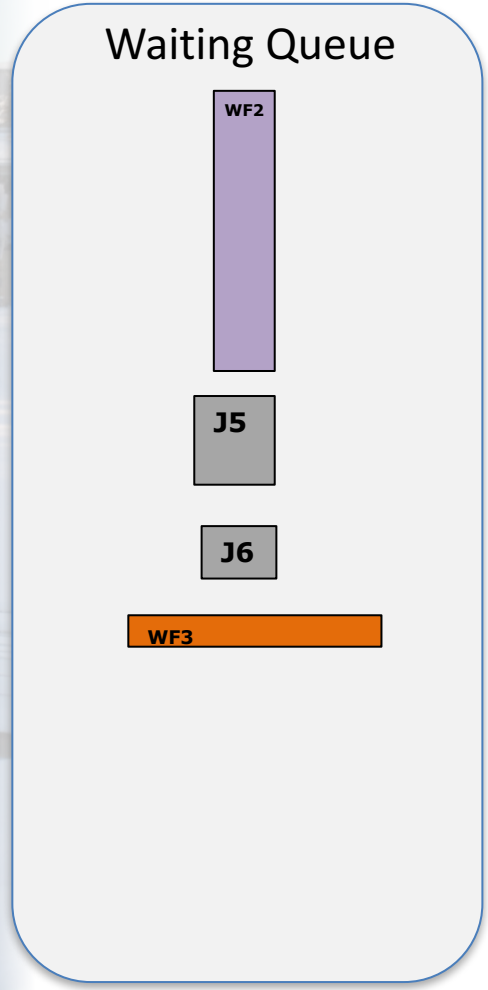


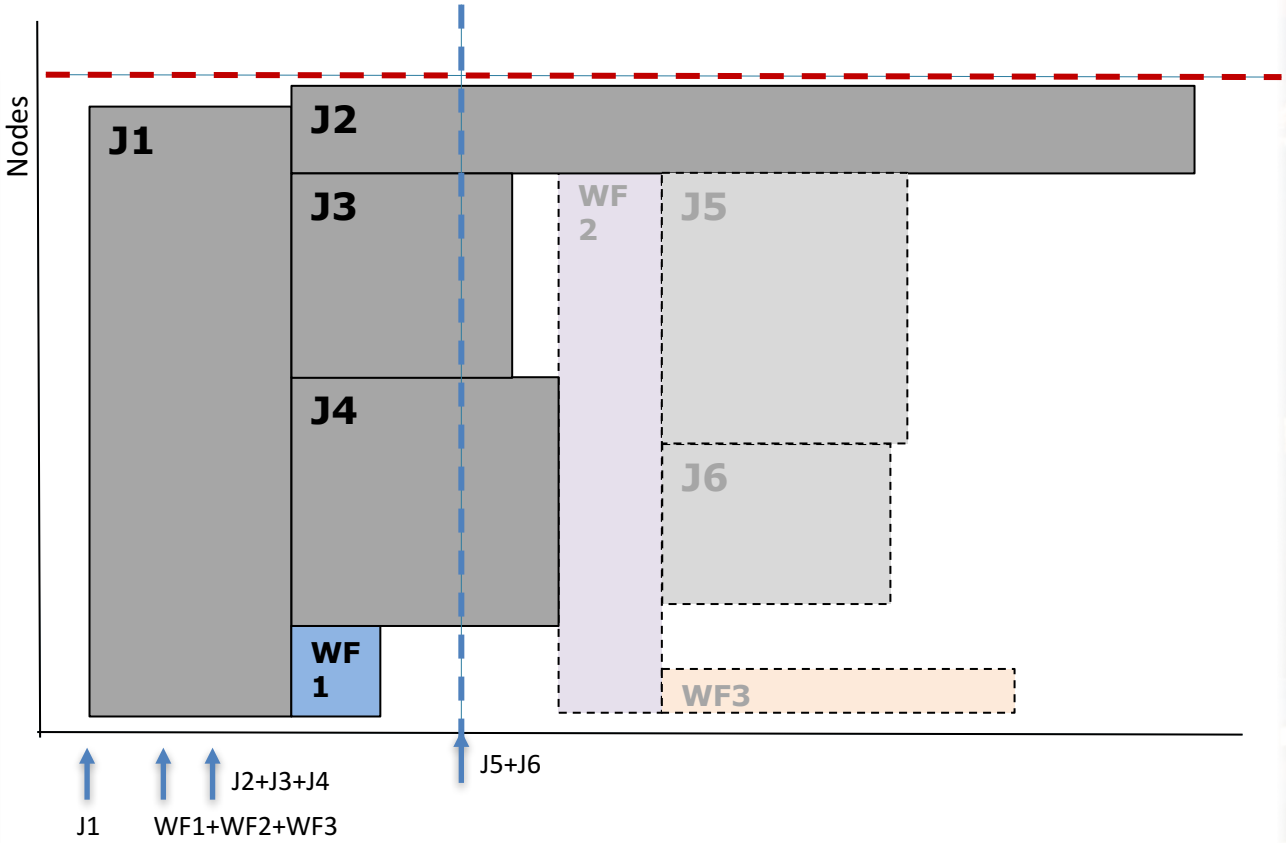


Workflows and HPC Schedulers

Schedulers are not aware of workflows

"Chained Jobs"/Wait Approach

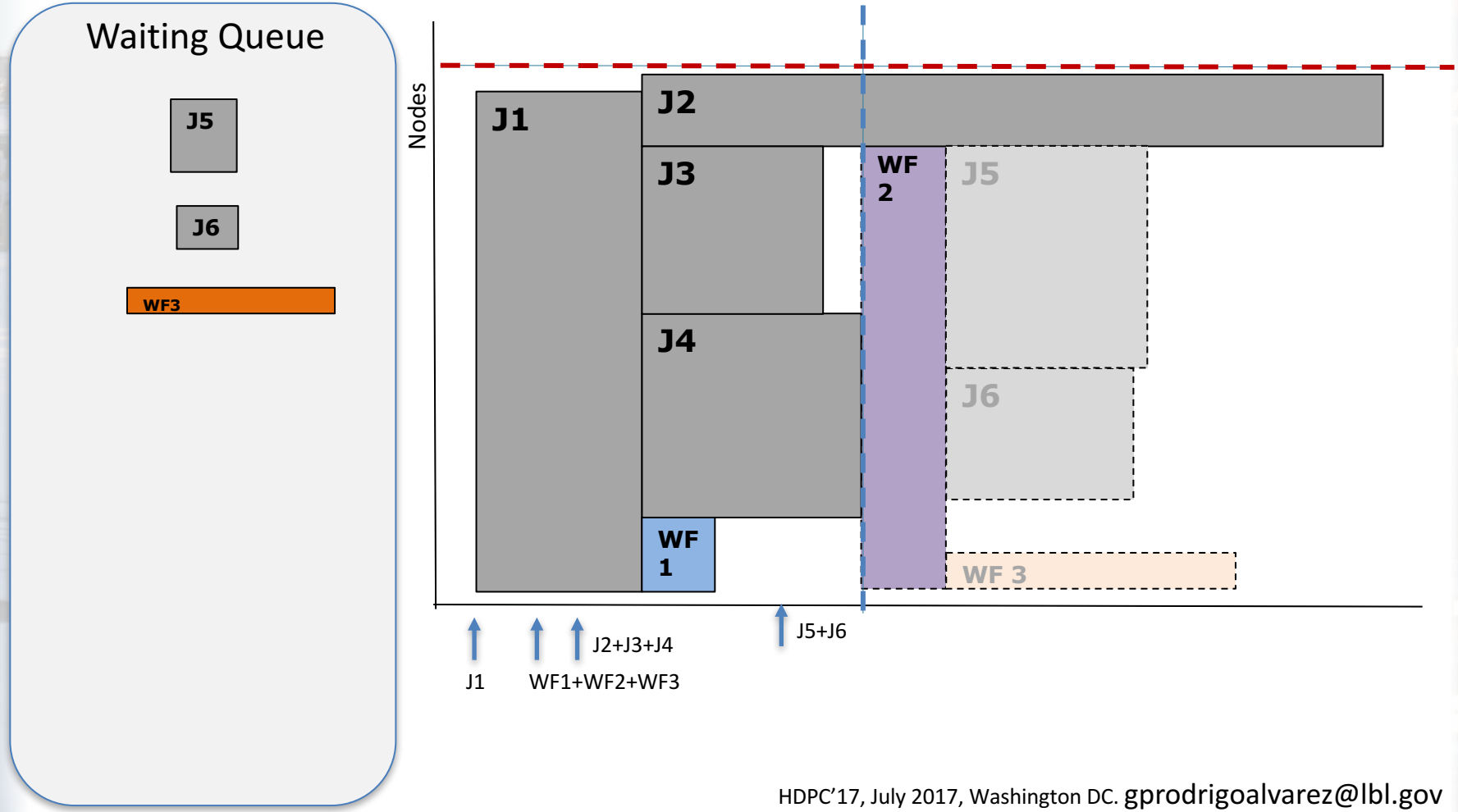




Workflows and HPC Schedulers

Schedulers are not aware of workflows

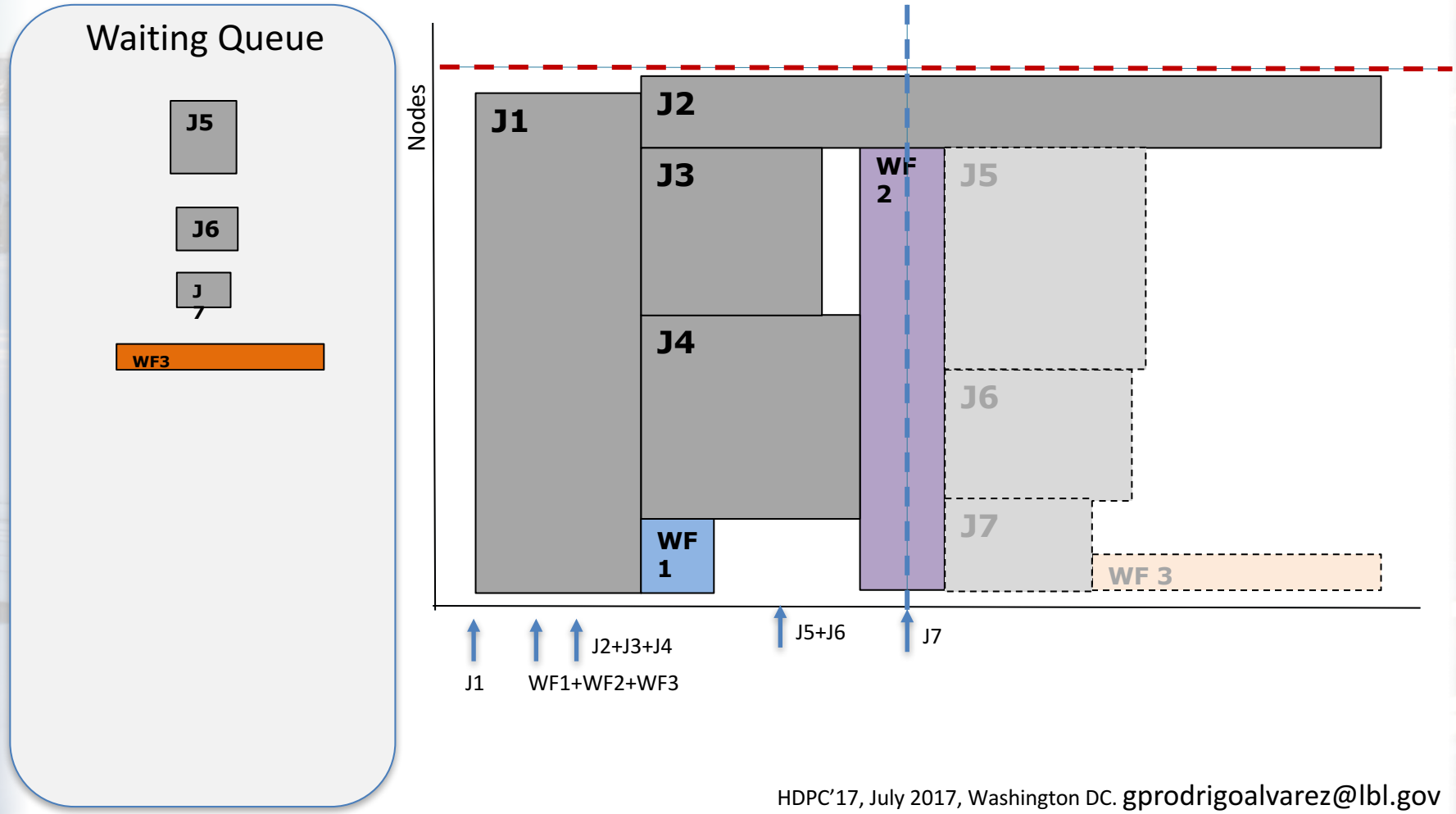
"Chained Jobs"/Wait Approach



Workflows and HPC Schedulers

Schedulers are not aware of workflows

"Chained Jobs"/Wait Approach

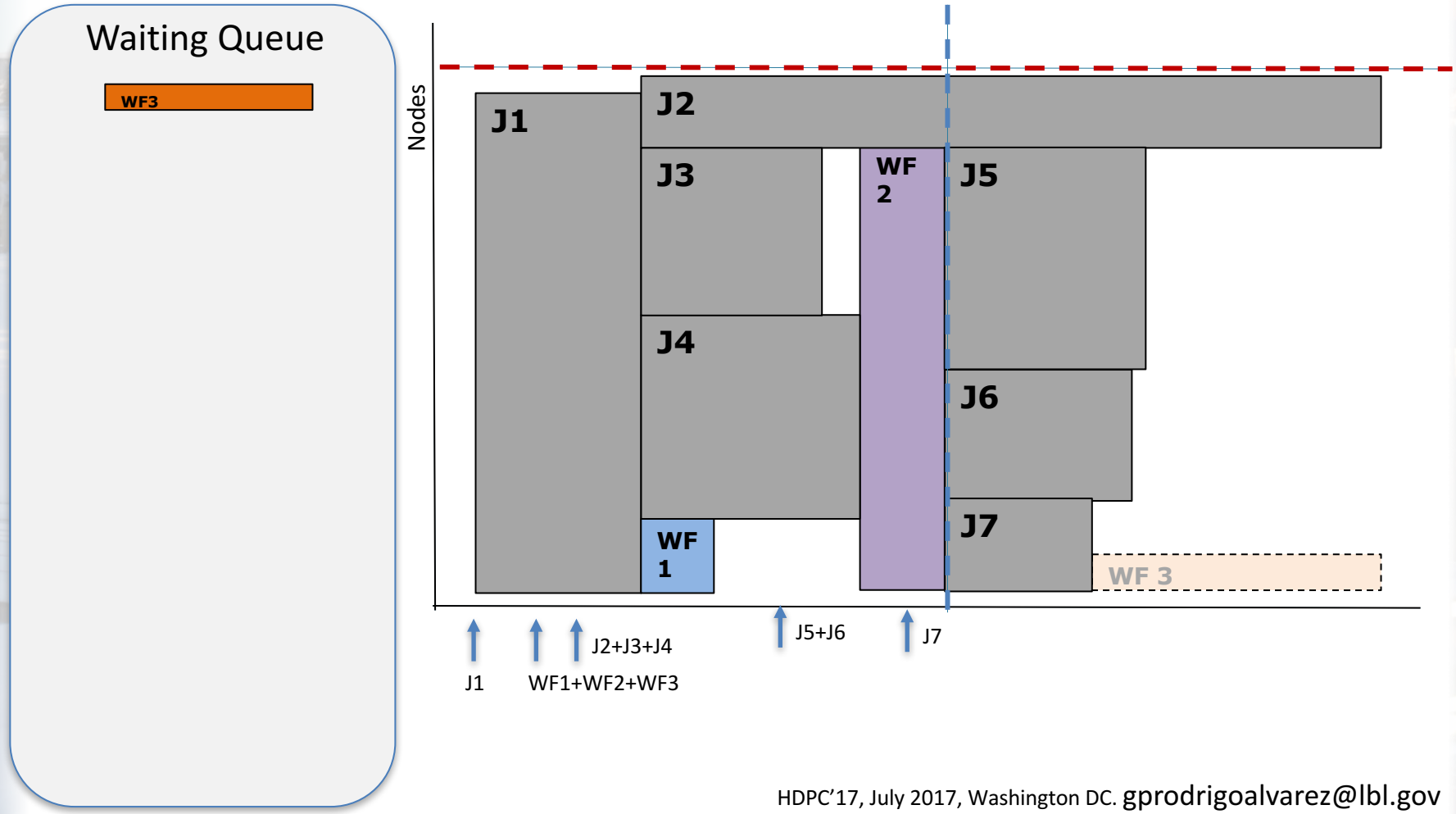


 gprodrigoalvarez@lbl.gov

Workflows and HPC Schedulers

Schedulers are not aware of workflows

"Chained Jobs"/Wait Approach

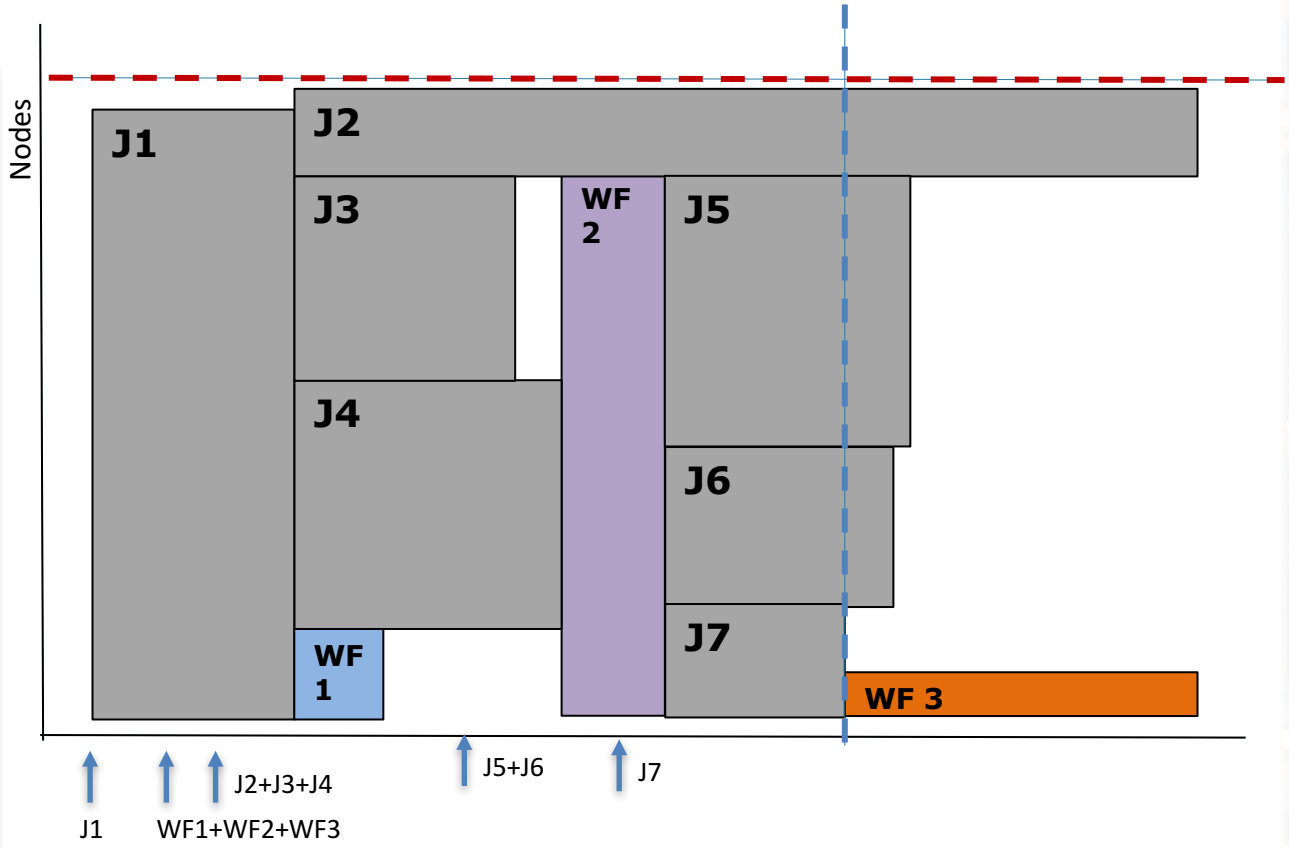


Workflows and HPC Schedulers

Schedulers are not aware of workflows

"Chained Jobs"/Wait Approach

Waiting Queue



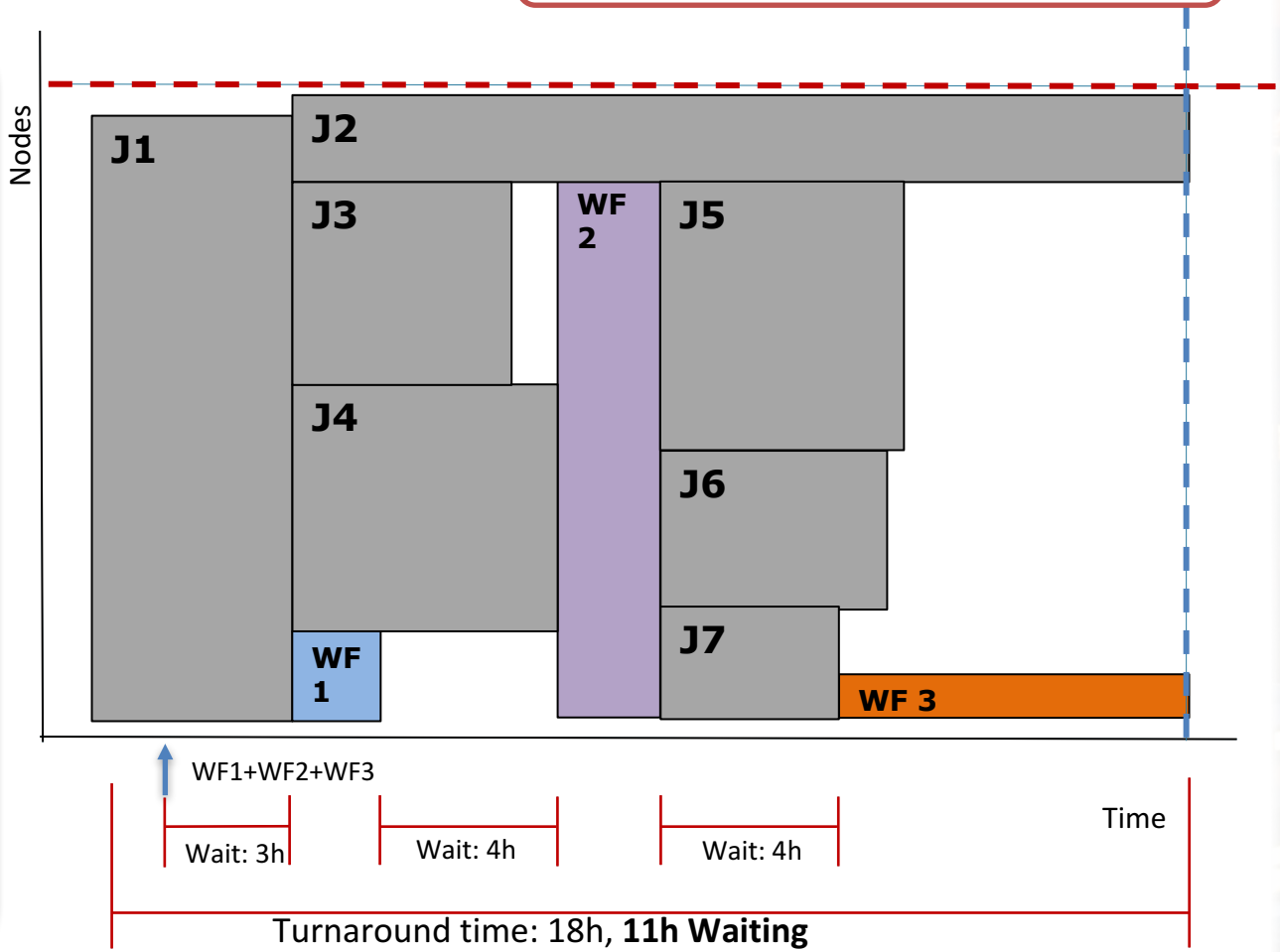
Workflows and HPC Schedulers

Schedulers are not aware of workflows

"Chained Jobs"/Wait Approach

Long turnaround times

Waiting Queue

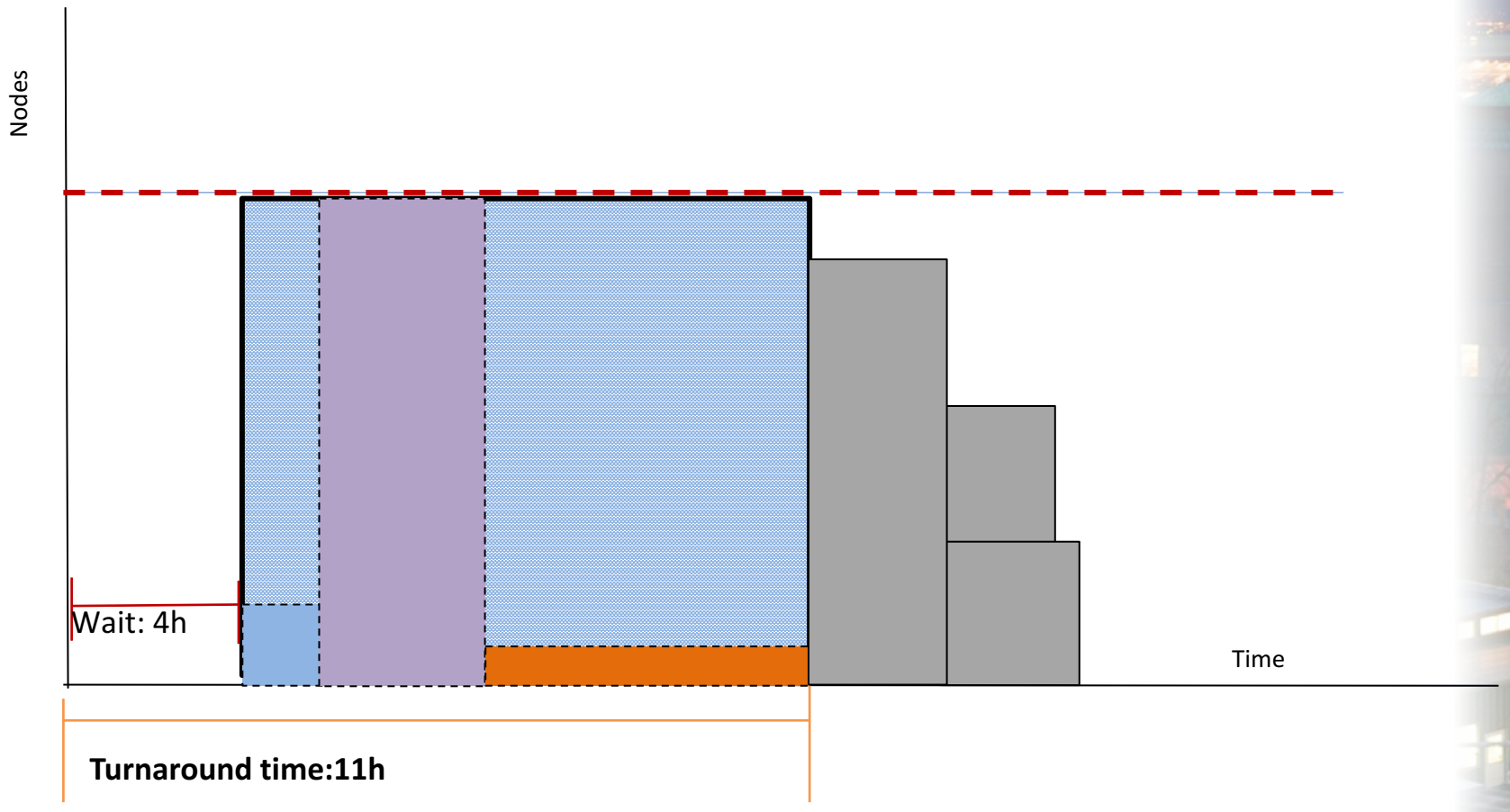


 gprodrigoalvarez@lbl.gov

Workflows and HPC Schedulers

Schedulers are not aware of workflows

"Pilot job"/Waste Approach

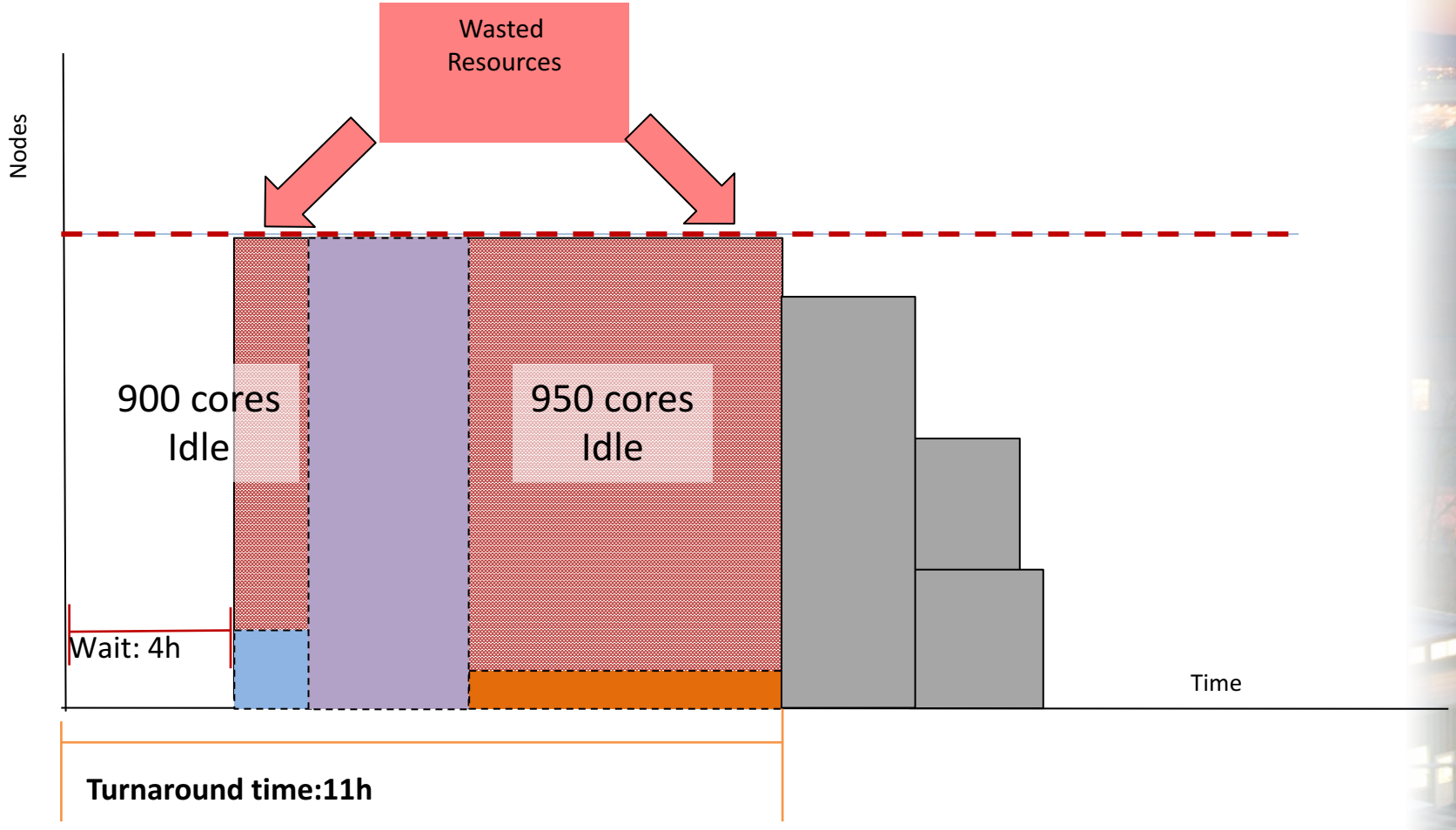


Workflows and HPC Schedulers

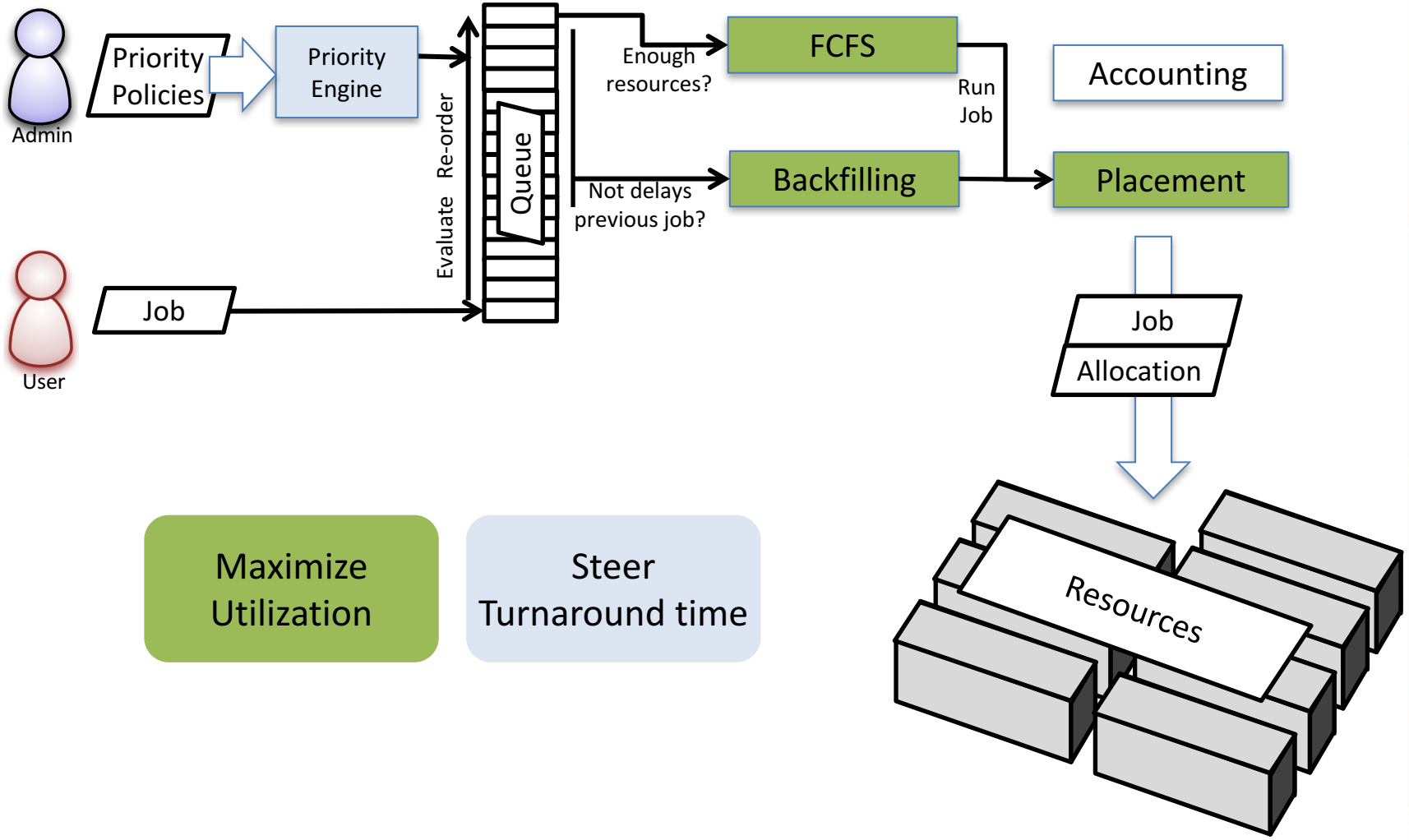
Schedulers are not aware of workflows

"Pilot job"/Waste Approach

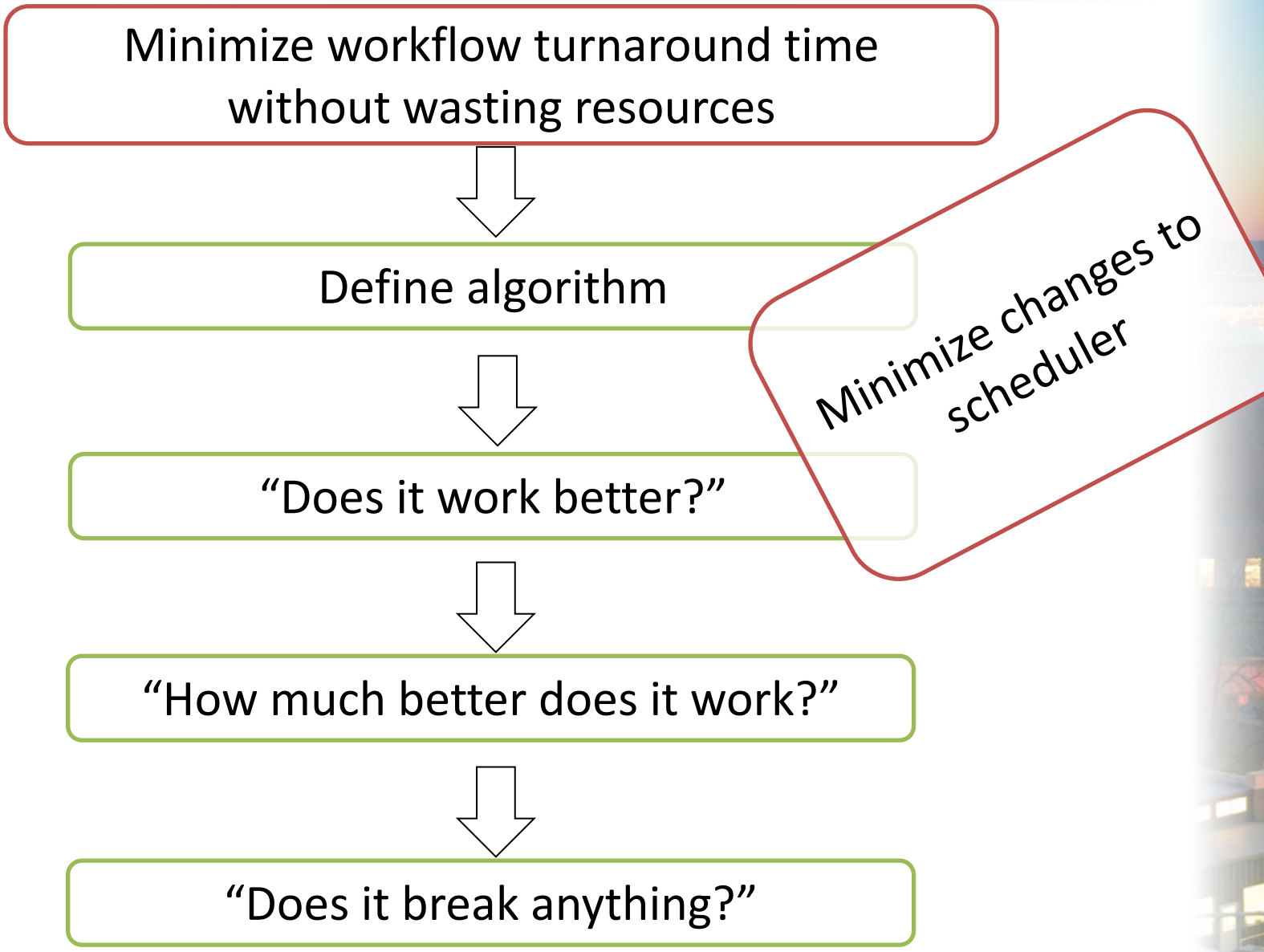
Idles resources wasted



Generic HPC Scheduler



Improving Workflow Scheduling



WoAS: Workflow Aware Scheduling

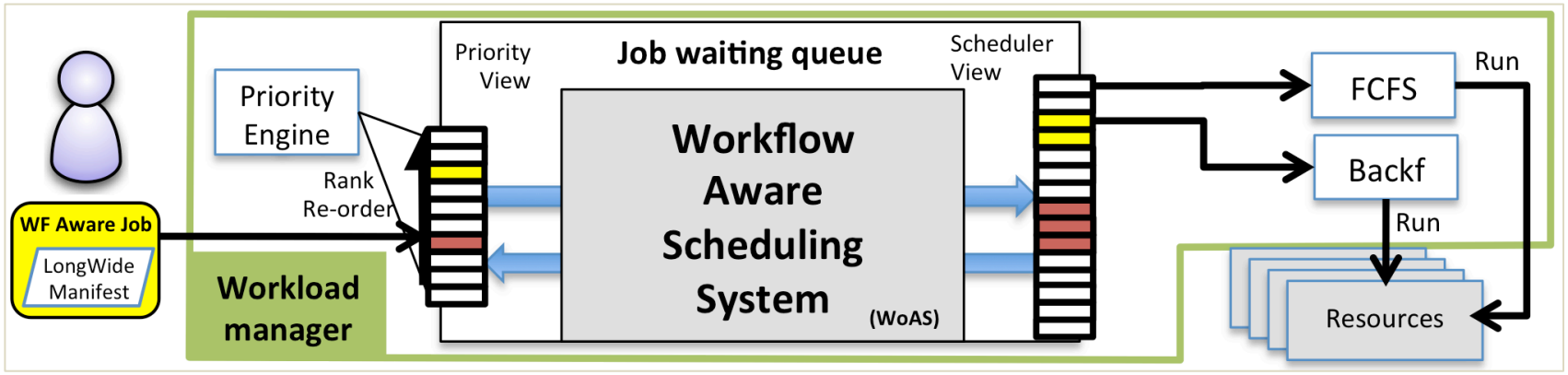
Best of both worlds

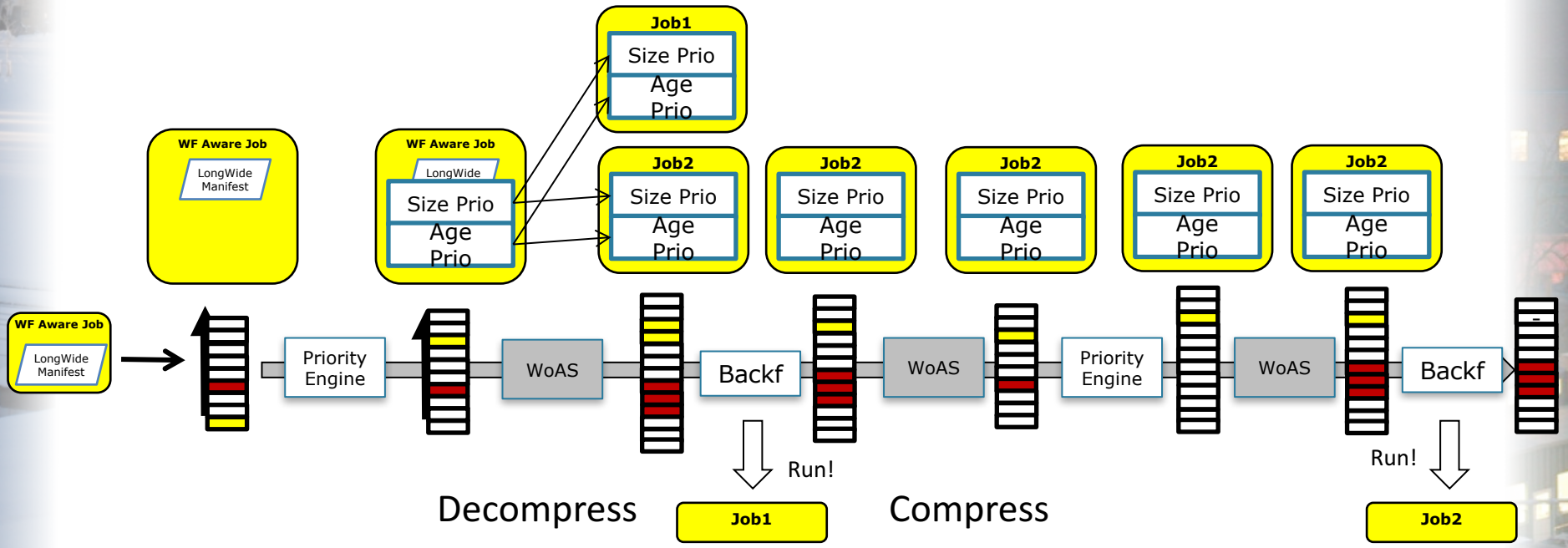
- Pilot Job
- Scheduler aware of “idle resources”

 gprodrigoalvarez@lbl.gov

WoAS: In a real Scheduler

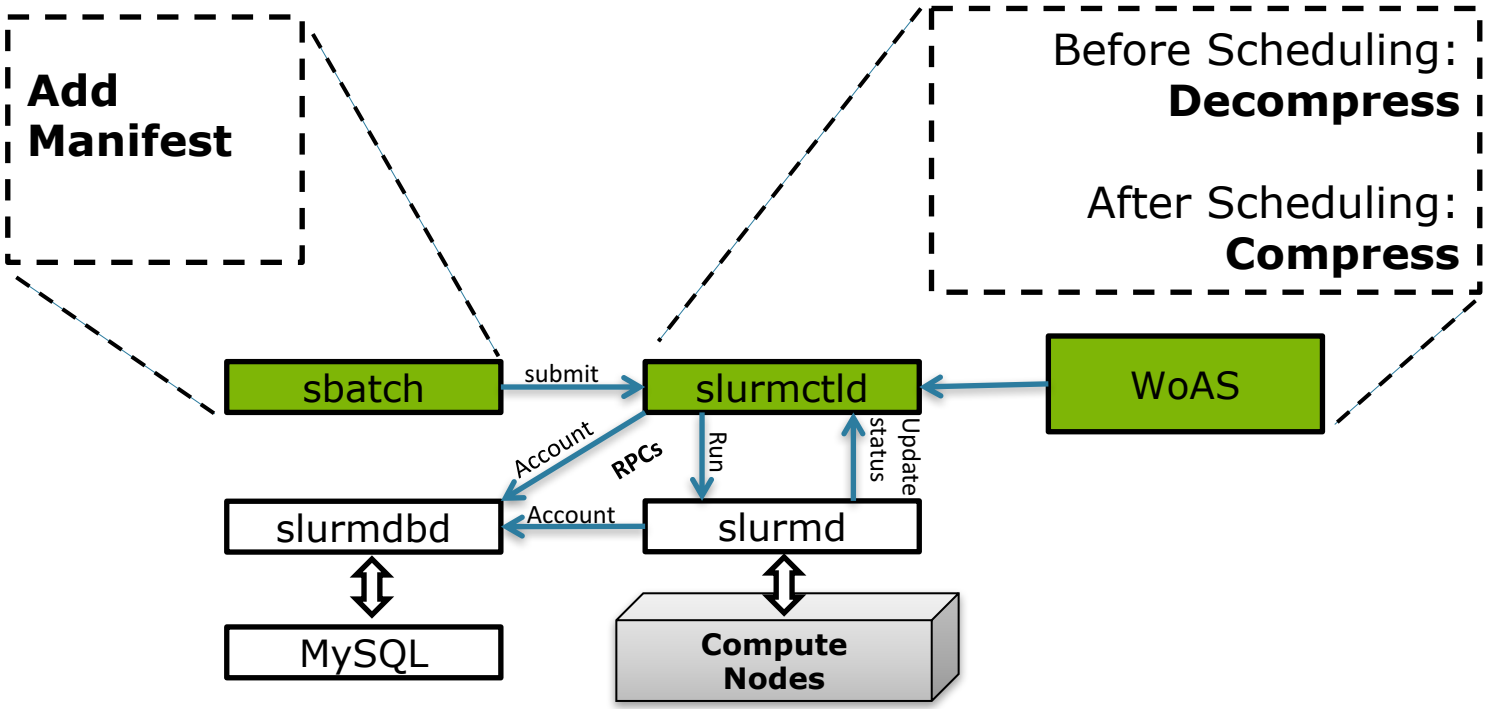
The "views" system





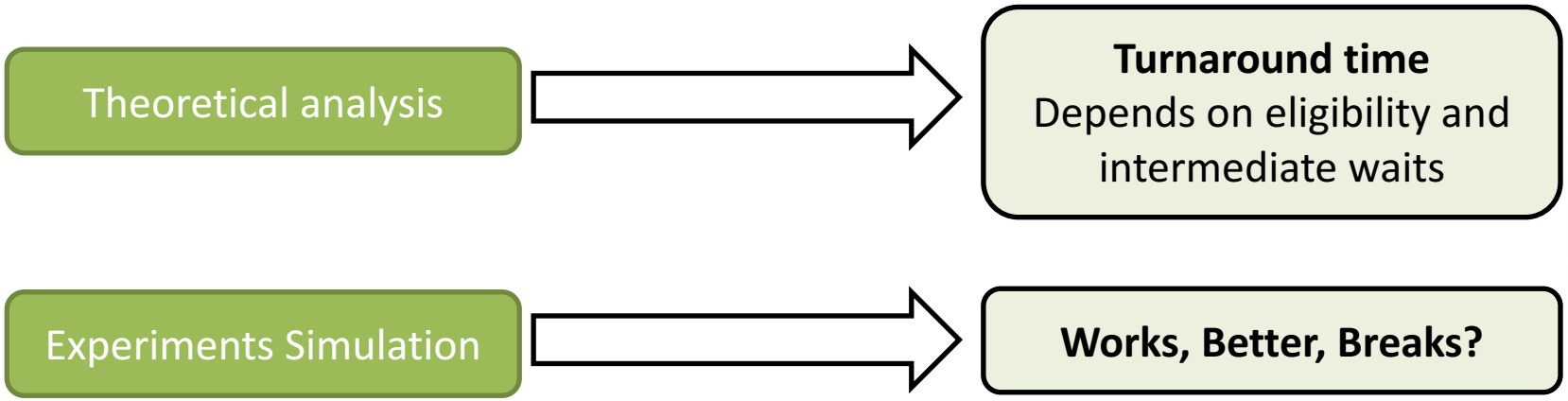
WoAS for Slurm

Main HPC scheduler and Open Source



Open Source Patch for Slurm 14.8.3

WoAS Evaluation

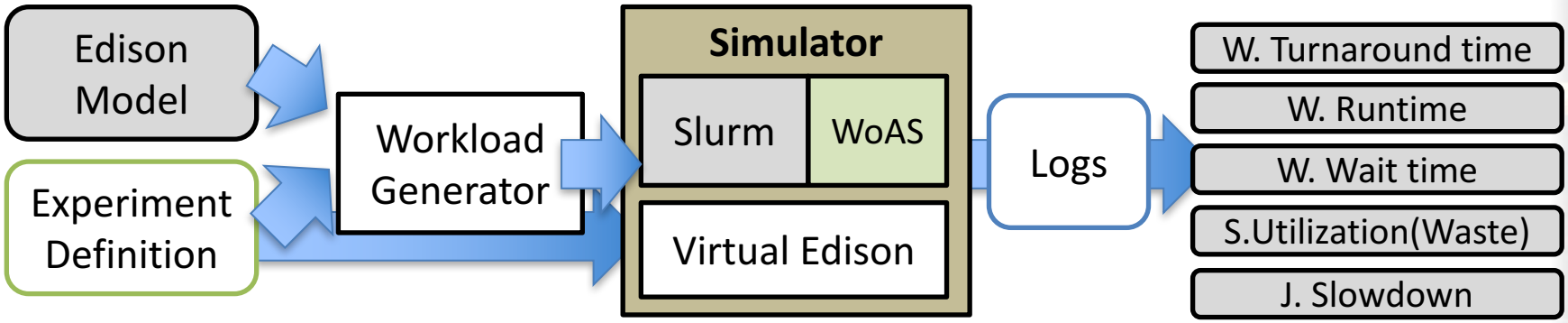


WoAS Evaluation: Simulations

Workload and system modeled after Edison

Workload

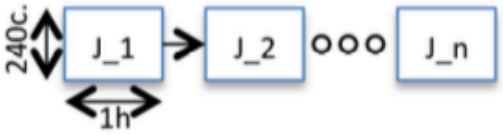
- Edison-like synthetic jobs
- Experiment specific workflows
- 5 +1 Days of simulated system



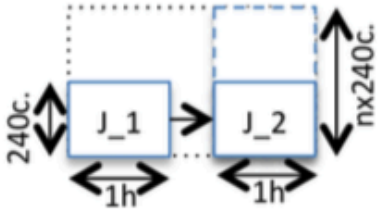
271 Scenarios, 1626 Experiments. 29 years of Edison: 3.8 Million Core-Years

Results: Does WoAS work?

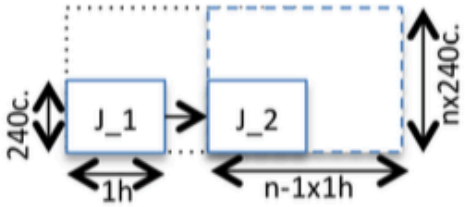
Type A

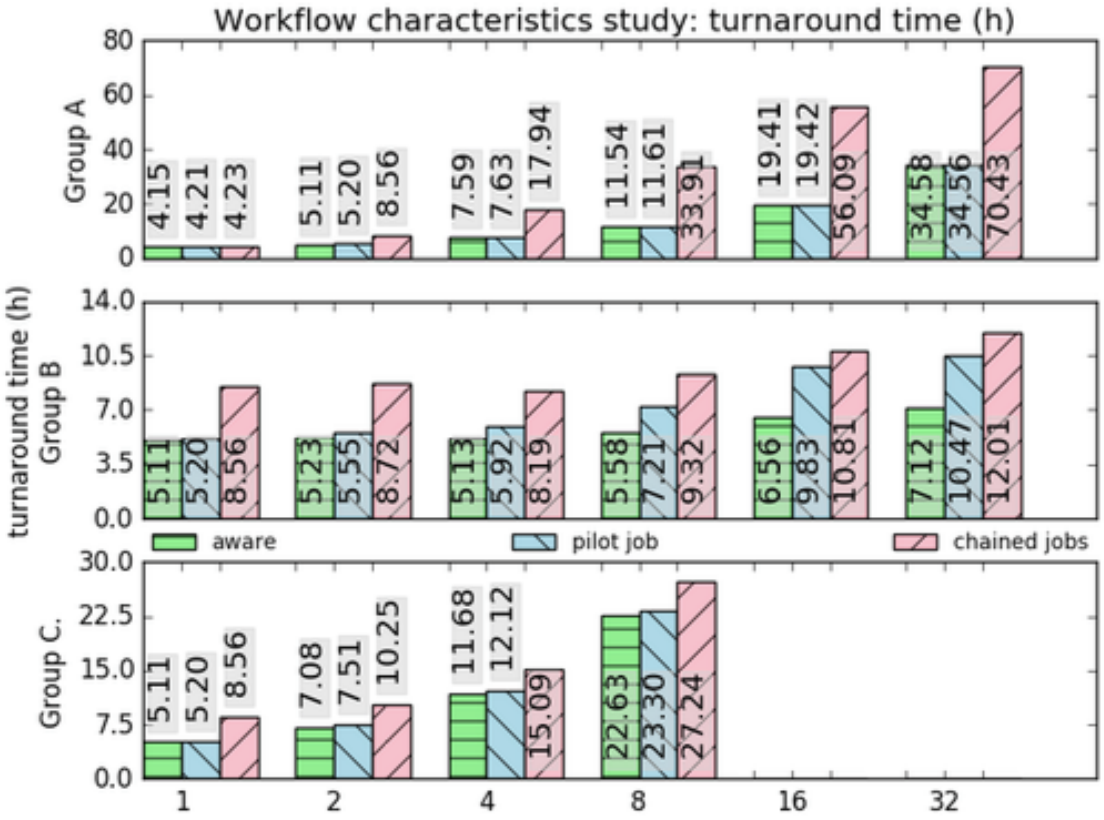


Type B



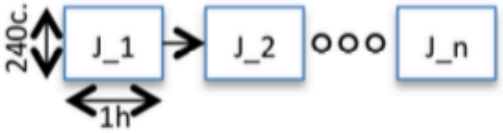
Type C



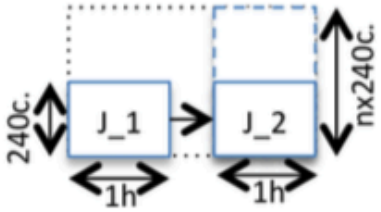


Results: Does WoAS work?

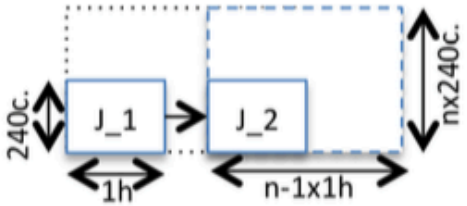
Type A

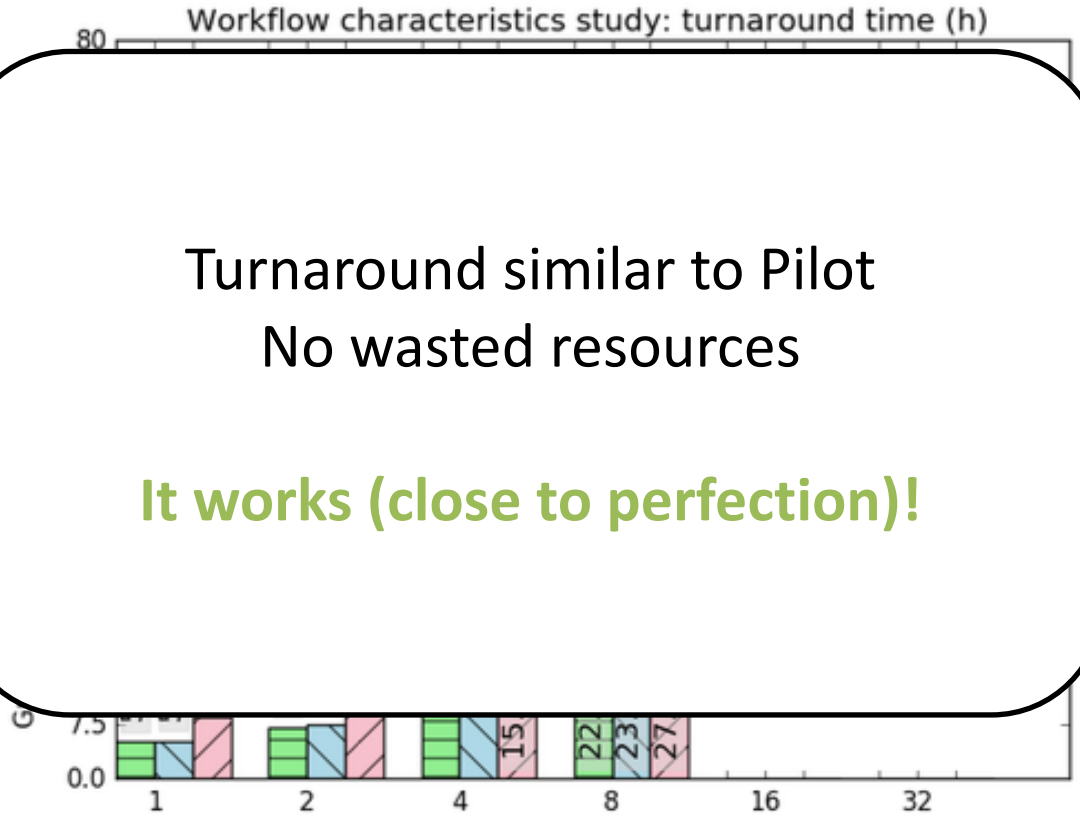


Type B



Type C





Turnaround similar to Pilot

No wasted resources

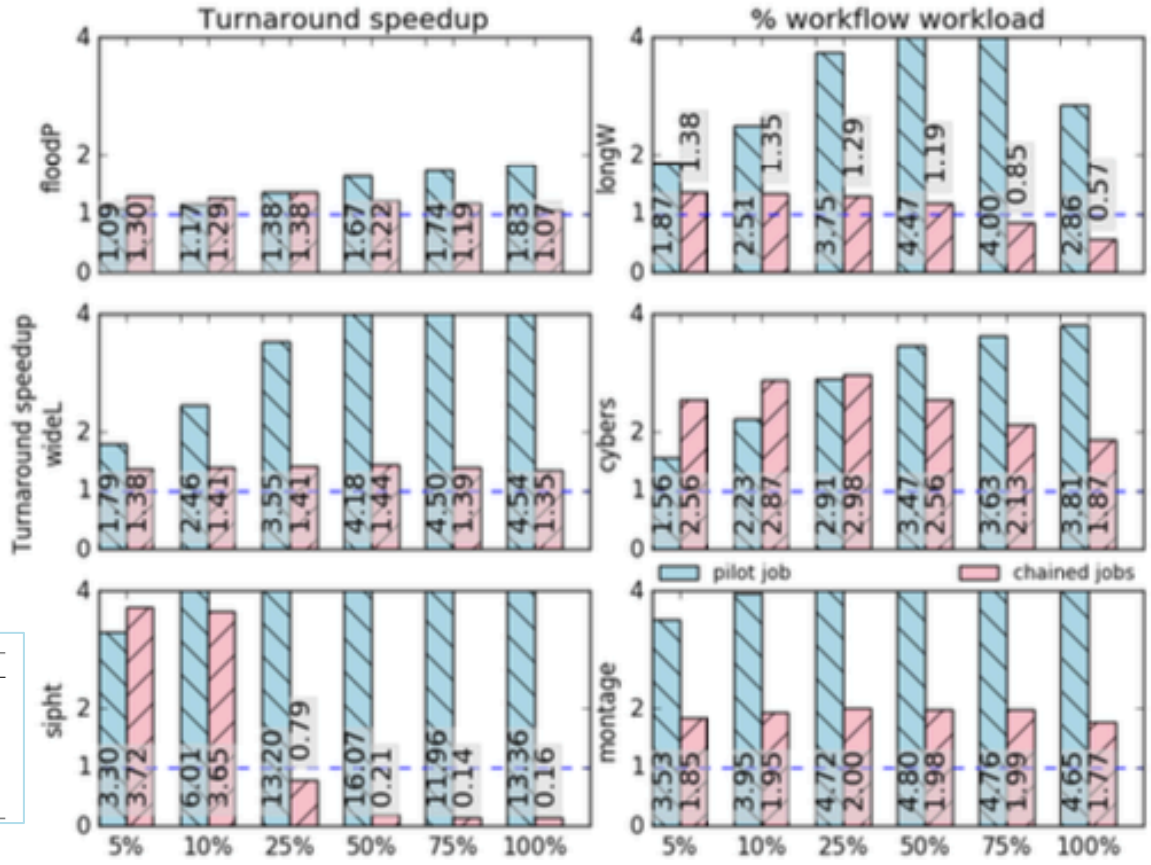
It works (close to perfection)!

Results: How much does WoAS work better?

Different core-hours % for workflows

Workflows

- LongWide
- WideLong
- FloodPlain
- Montage
- Cybershake
- Sipht



Gain(%)	1%	5%	10%	25%	50%	75%	100%
floodP	1.80	5.22	14.46	29.29	44.53	51.64	64.47
longW	2.30	8.33	18.93	30.84	40.25	31.99	27.18
wideL	0.33	10.64	19.74	32.35	48.22	57.19	66.16
cybers	1.66	7.72	13.92	25.58	36.72	44.45	52.83
sipht	2.55	11.41	18.16	34.85	42.77	37.27	35.83
montage	12.36	44.90	60.30	72.34	80.13	82.14	85.26

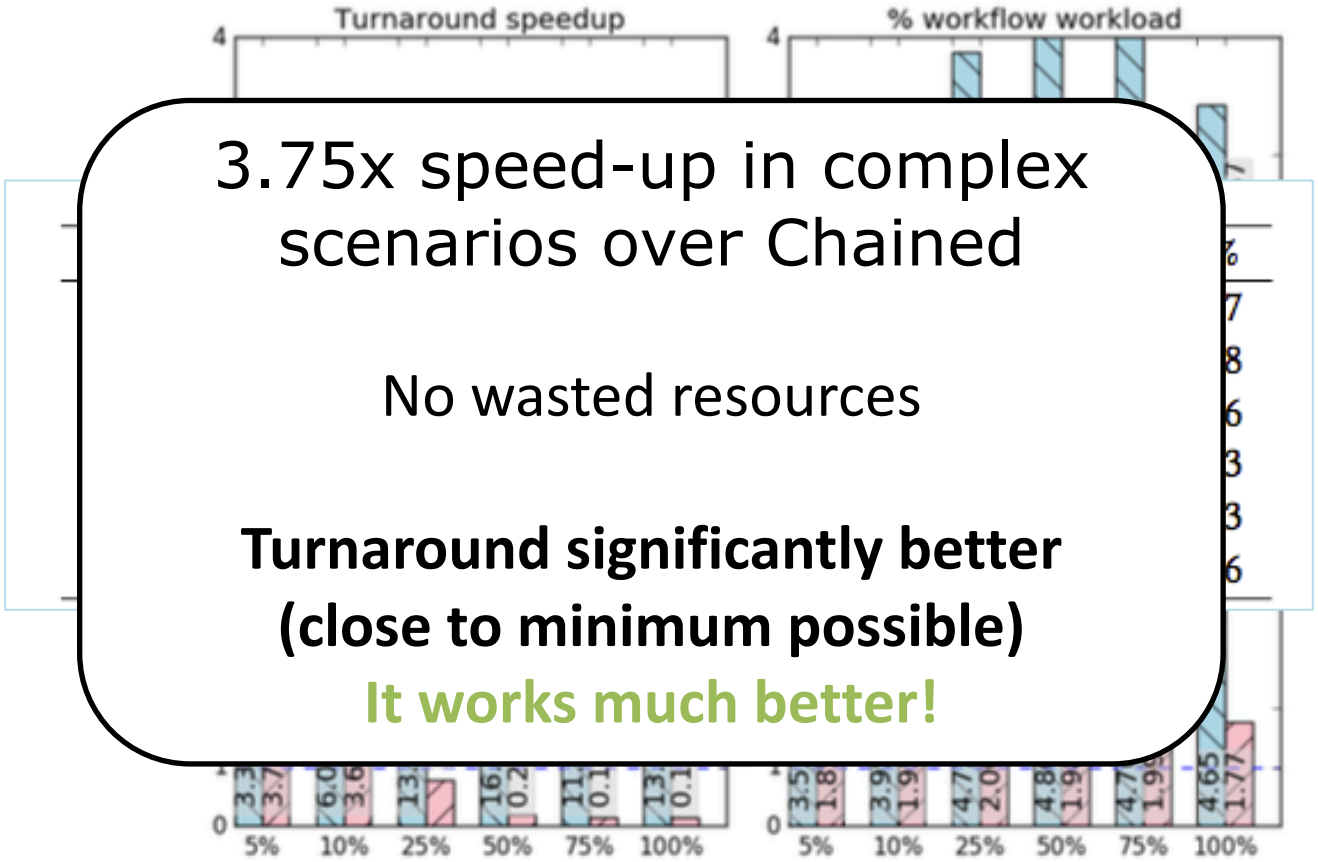
Results: How much does WoAS work better?

Different core-hours % for workflows

Workflows

LongWide | WideLong | FloodPlain

Montage | Cybershake | Sipht



3.75x speed-up in complex scenarios over Chained

No wasted resources

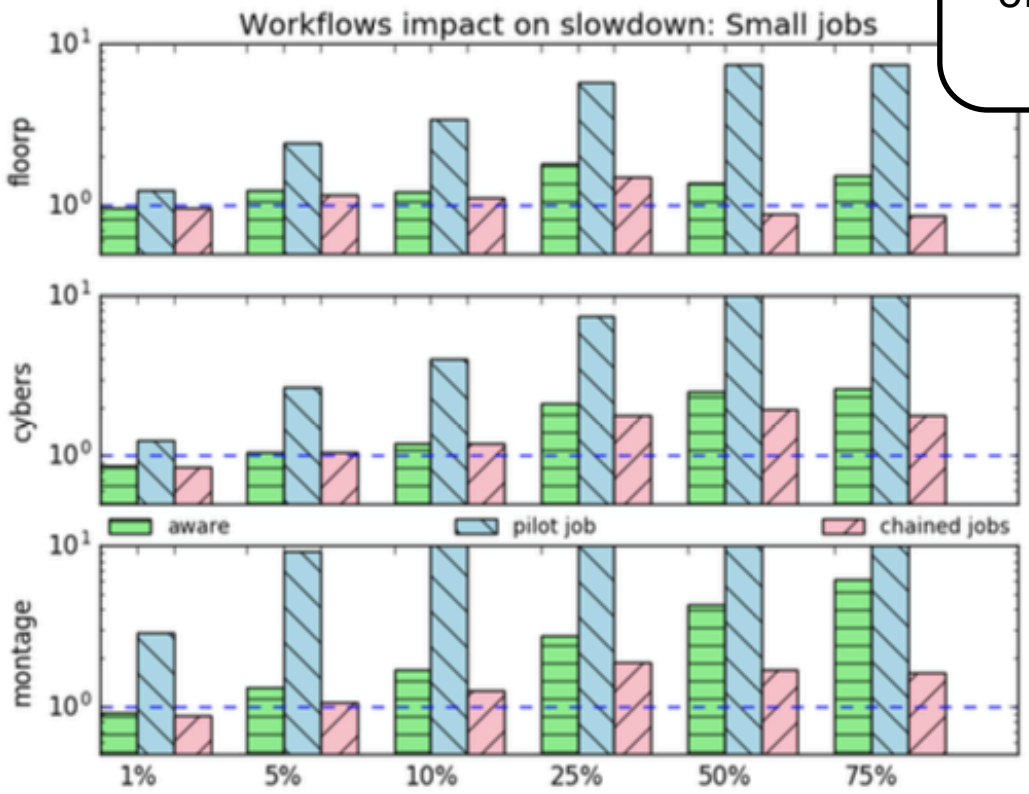
Turnaround significantly better (close to minimum possible)

It works much better!

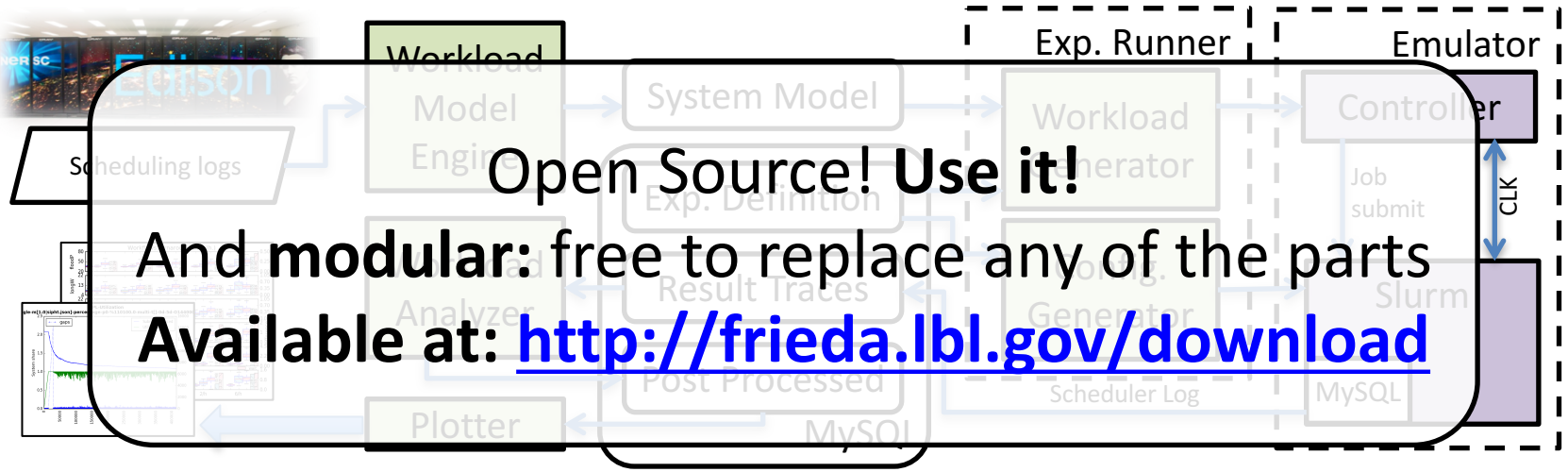
Evaluation: Does WoAS break the schedule?

Regular Jobs Slowdown Analysis

No significant effect on jobs' slowdown



ScSF: Scheduling Simulation Framework



Open Source! **Use it!**

And **modular:** free to replace any of the parts

Available at: http://frieda.lbl.gov/download

HPC Scheduling research cycle:
Model/generate workloads -> scheduling emulation -> analysis

Tools to run experiments in scale

Slurm simulator in its core: A production HPC simulator

WoAS: Take-Aways

In-site **scientific workflows are important** in HPC

Users forced to face **long turn around times**… or to **waste resources**

WoAS Minimizes turnaround time, without wasting resources

WoAs Requires minimum changes to the scheduler

Open Source patch for Slurm! **Use it!**
Download it at : http://frieda.lbl.gov/download

THANKS

For any questions, please contact:
gprodrigoalvarez@lbl.gov

odrigo Álvarez, G.P, Elmroth, E., Östberg, P.O., Ramakrishnan, L. **Enabling workflow aware scheduling on HPC ystems.** 26th International Symposium on High-Performance Parallel and Distributed Computing (HPDC 2017)

WoAS patch for Slurm and ScSF (simulator) are open source and **available at:**
http://frieda.lbl.gov/download

Supported by U.S. Department of Energy, Office of Science, Office of Advanced Scienti□c Computing Research (ASCR). the National Energy Research Scienti□c Computing Center, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Financial support has been provided in part by the Swedish Government's strategic effort eSSENCE and the Swedish Research Council (VR) under contract number C0590801 (Cloud Control).

UMEÅ UNIVERSITET

BERKELEY LAB