Imbalance Optimization in Scientific Workflows

Weiwei Chen[1], Ewa Deelman[1], Rizos Sakellariou[2]
[1] Information Sciences Institute, University of Southern California, {wchen,deelman}@isi.edu
[2] School of Computer Science, University of Manchester, rizos@cs.man.ac.uk

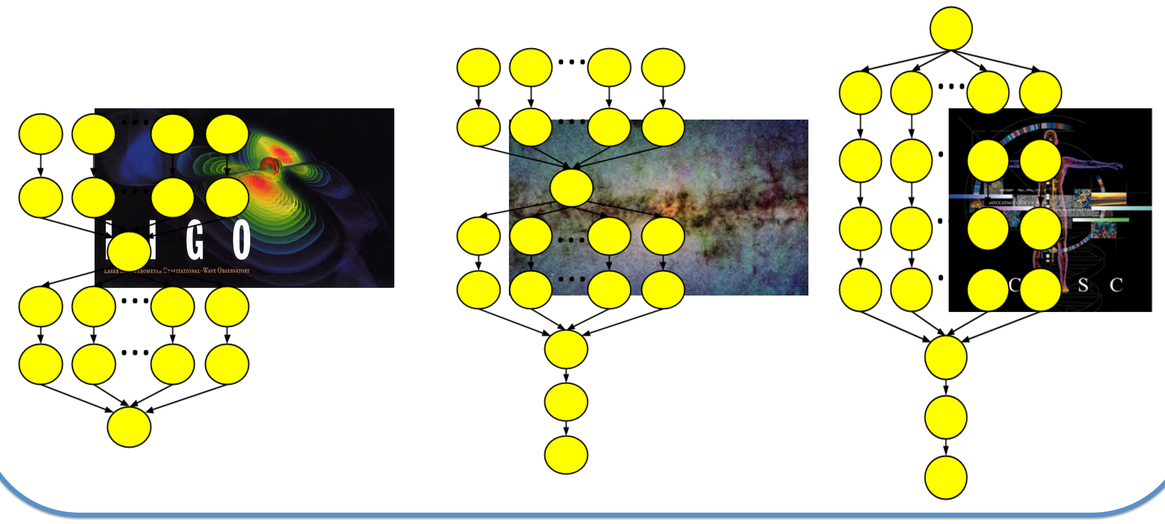
1 Introduction

Computational scientists develop and use large-scale, loosely-coupled applications that are often structured as scientific workflows (earthquake science, astronomy, biology)

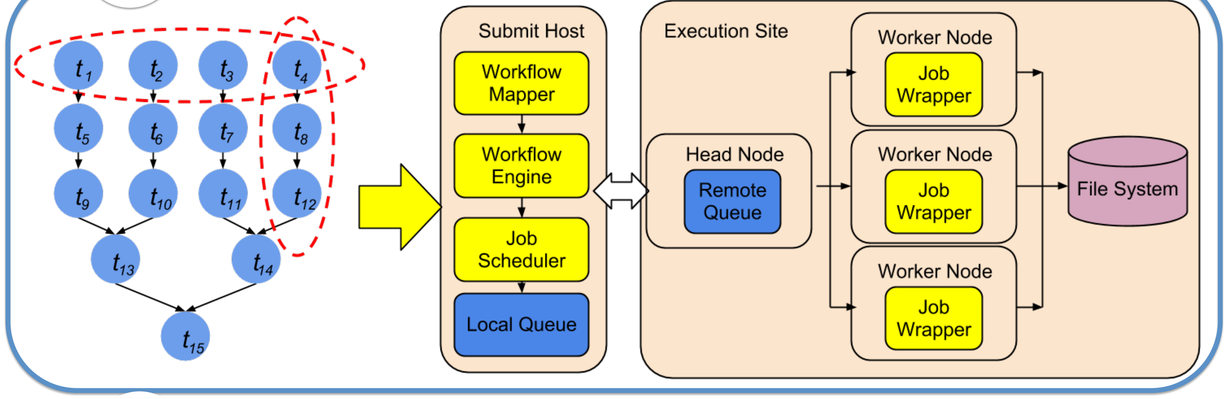
- Executing these applications on Grids or Clouds has significant system overheads
- Many applications involve millions of tasks processing massive data analysis on distributed resources
- Task Clustering merges small granularity tasks into large jobs so as to reduce overheads
- However, the balancing of runtime and dependency is not yet addressed. Particularly measuring dependency imbalance quantitatively is a big challenge

2 Scientific Workflows

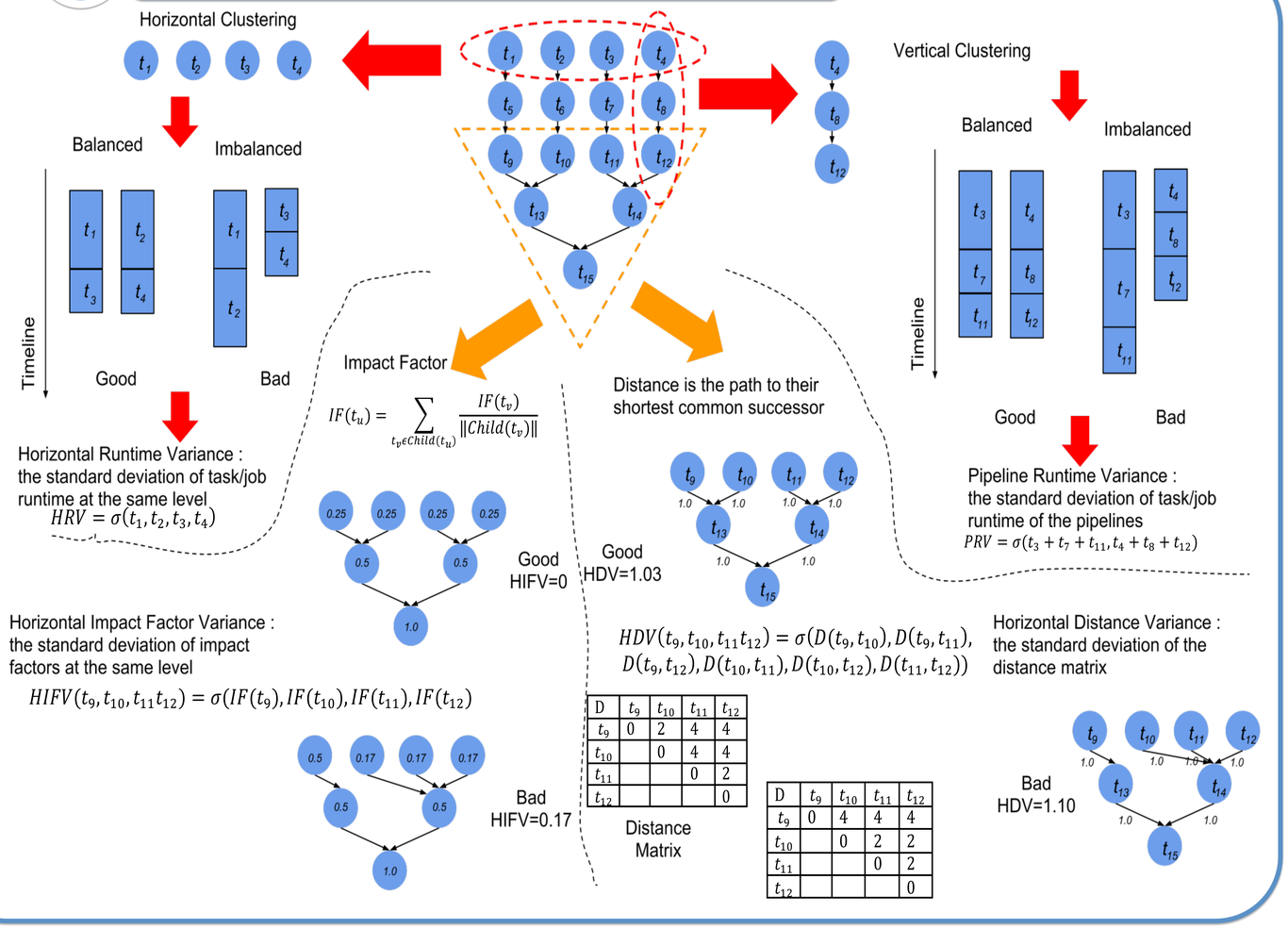
LIGO Inspiral | Montage | Epigenomics



3 Task Clustering



4 Imbalance Measurement

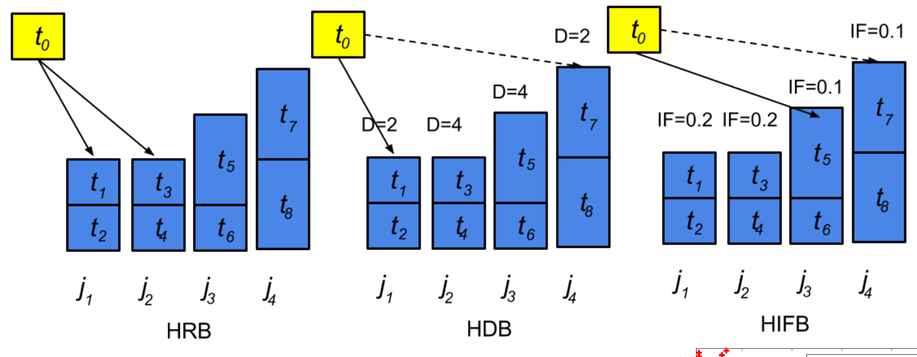


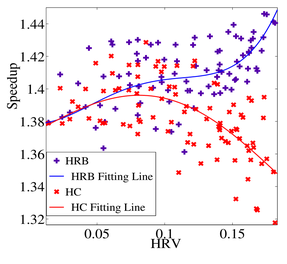
D	t_9	t_{10}	t_{11}	t_{12}
t_9	0	2	4	4
t_{10}		0	4	4
t_{11}			0	2
t_{12}				0

D	t_9	t_{10}	t_{11}	t_{12}
t_9	0	4	4	4
t_{10}		0	2	2
t_{11}			0	2
t_{12}				0

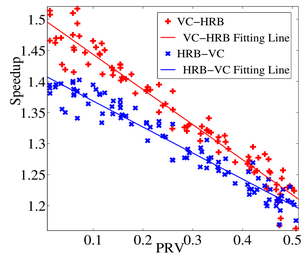
5 Imbalance Optimization

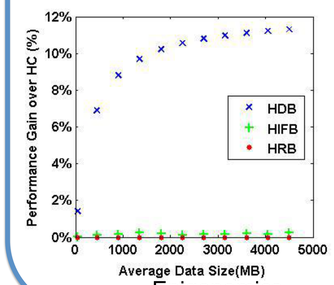
- **Horizontal Runtime Balancing (HRB):** add the longest task to the shortest job
- **Horizontal Distance Balancing (HDB):** sort tasks based on distance, then HRB
- **Horizontal Impact Factor Balancing (HIFB):** sort tasks based on IFs then HRB



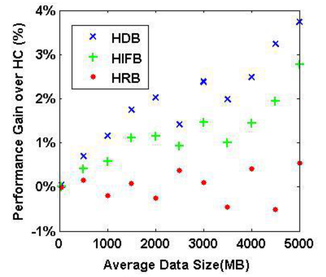


- HRB performs better than HC with the increase of HRV and PRV
- HDB HIFB perform better than HRB when data size increases but worse when HRV increases

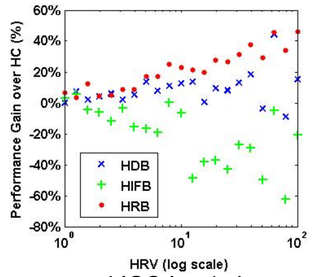




Epigenomics



LIGO Inspiral



LIGO Inspiral

This work is supported by NFS under grant number IIS-0905032.