HPC System Lifetime Story: Workload Characterization and Evolutionary Analyses on NERSC Systems

Gonzalo P. Rodrigo[1], P-O Östberg[1], Erik Elmroth[1], Katie Antypas[2], Richard Gerber[2], Lavanya Ramakrishnan[2]

1: Umeå University, Umeå - Sweden

2: Lawrence Berkeley National Lab, Berkeley, California - USA

Goal

Understand the characteristics of present workloads to design future systems

- Analyses of the evolution and trends in the workloads of a high performance cluster and a supercomputer.
- Infer how current workloads differ from older ones: Classical workload managers were built to schedule tightly coupled parallel jobs. If the workload evolves enough, schedulers may not be able to produce efficient job management decisions.

Background

Infrastructure evolution: Exascale

- New hardware bring new scheduling challenges: burst buffer allocation.
- Data movement costs are increasing due to power and performance requirements.
- Memory scaling and I/O bandwidth are limited compared to compute capacity.

Application Evolution

- Data intensive applications are a growing trend in science.
- Diverse and complex workloads.
- Stream processing: coordinate life events with compute resources without advance reservations.

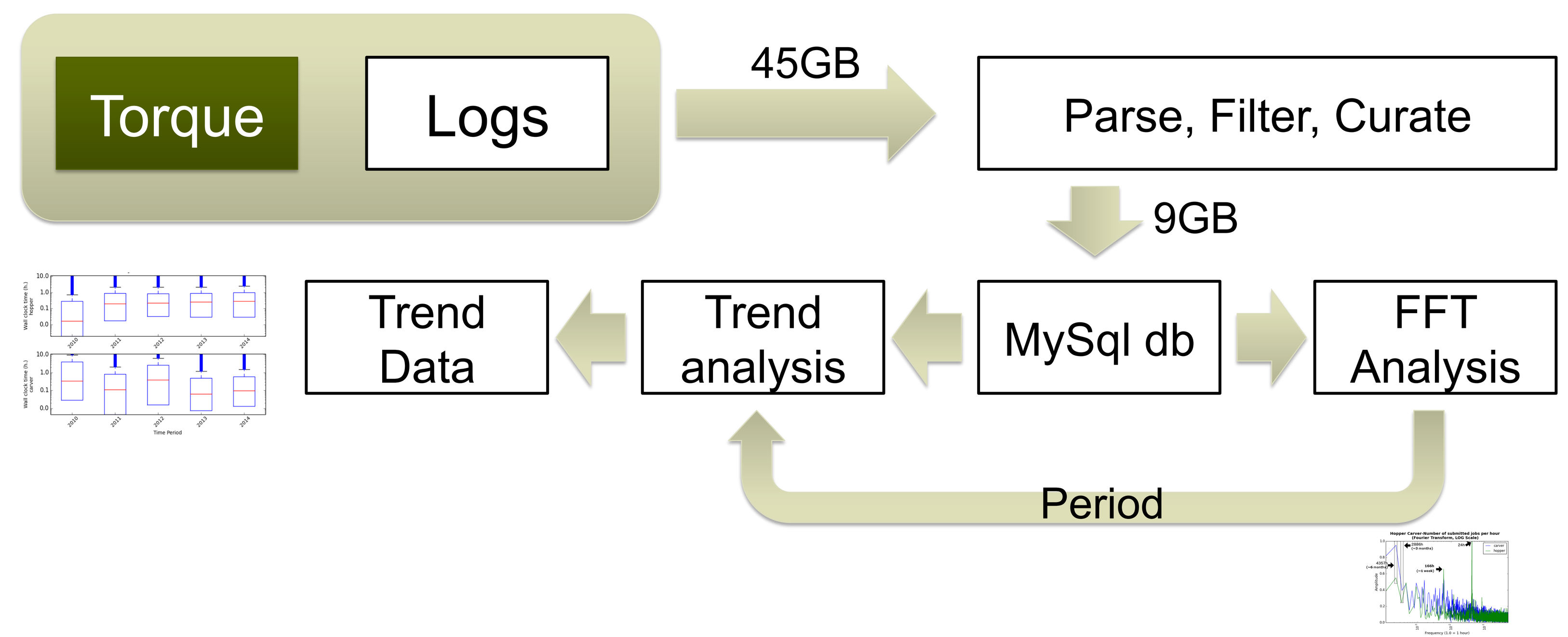
The Systems

Hopper	Carver
Deployed January 2010	Deployed 2010
Cray XE6	IBM iDataPlex
Gemini Network	Infiniband (fat-tree)
6,384 Nodes, 24 cores/node 154,216 cores	1,120 Nodes, 8/12/32 cores/node, 9,984 cores
1.28 Pflops/s	106.5 Tflops
Torque + Moab	Torque + Moab

Methodology

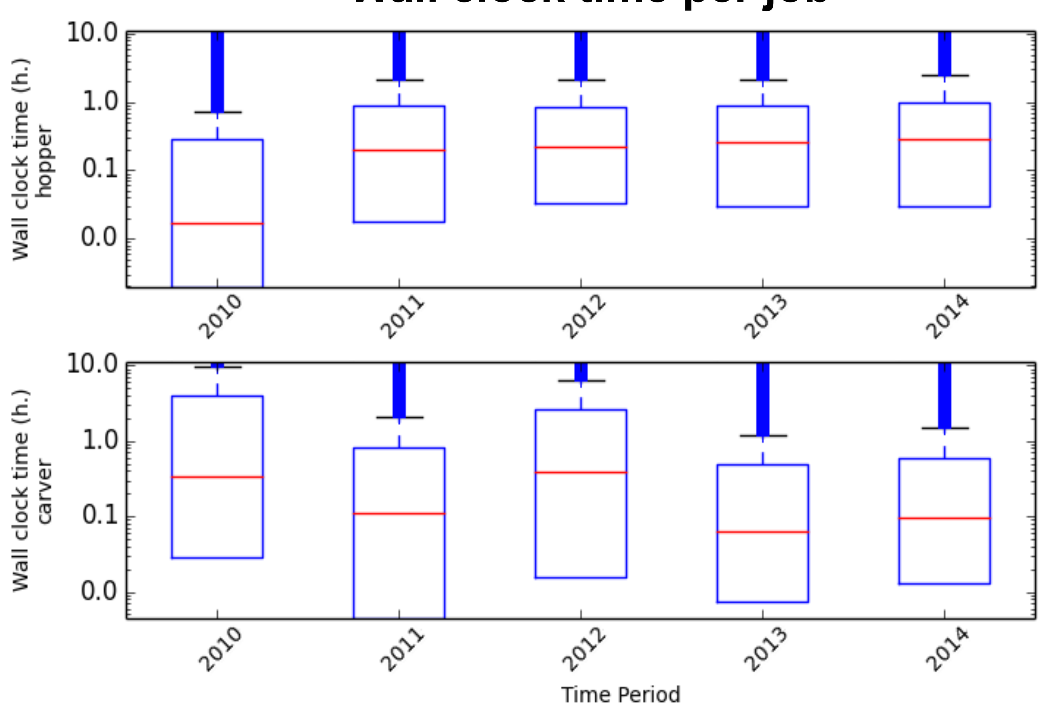
Studying the changes of the workload through time should provide insights on future workloads. This analysis is performed at job level.

- **Job variables:** Wall clock, number of cores (allocated), compute time (core hours allocated), wait time, wall clock time estimation accuracy, and time patterns.
- **Dataset**: Jobs 2010 – 2014: 4.5M (Hopper) and 9.3M (Carver) jobs. 45 GB of raw data filtered into 9 GB of useful data.
- **Data source**: Torque log events, including job's wall clock (requested & actual), number of cores, submission/start/completion time.
- **Time pattern analysis**: Yearly pattern reinforced by NERSC allocation year.

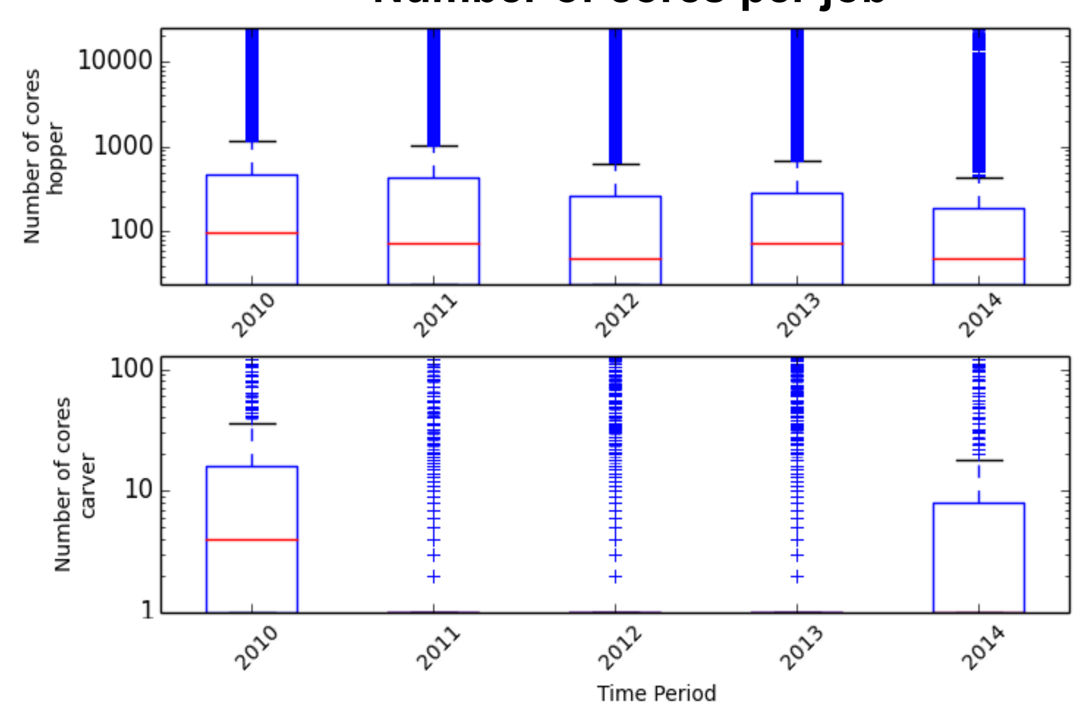


Results

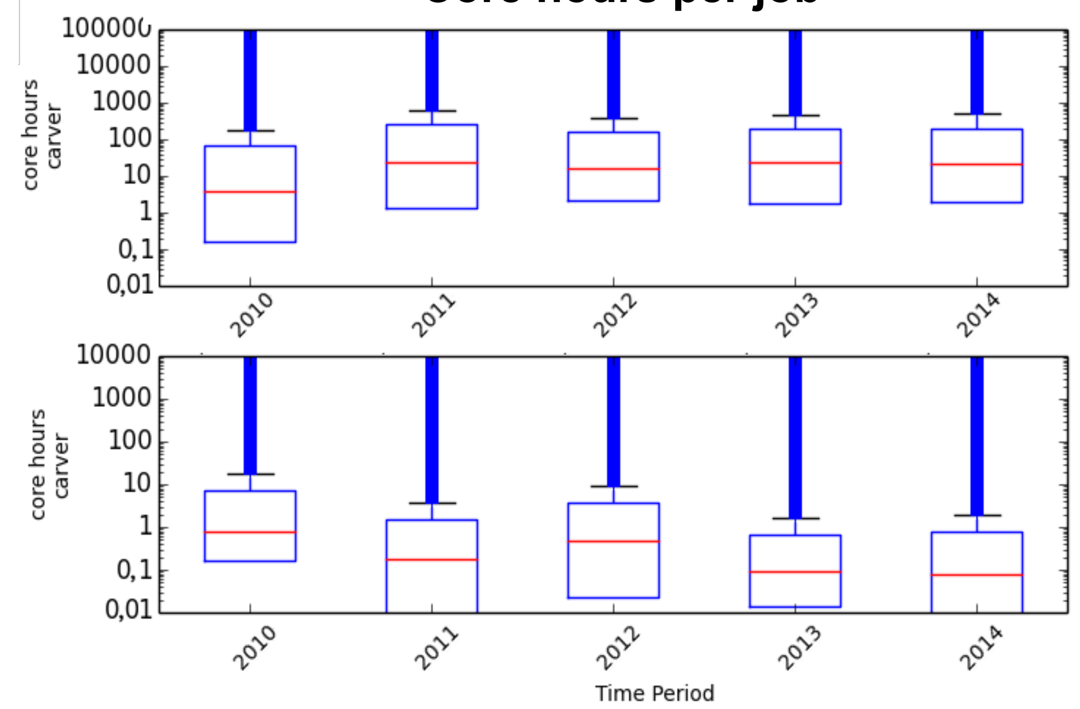
Wall clock time per job



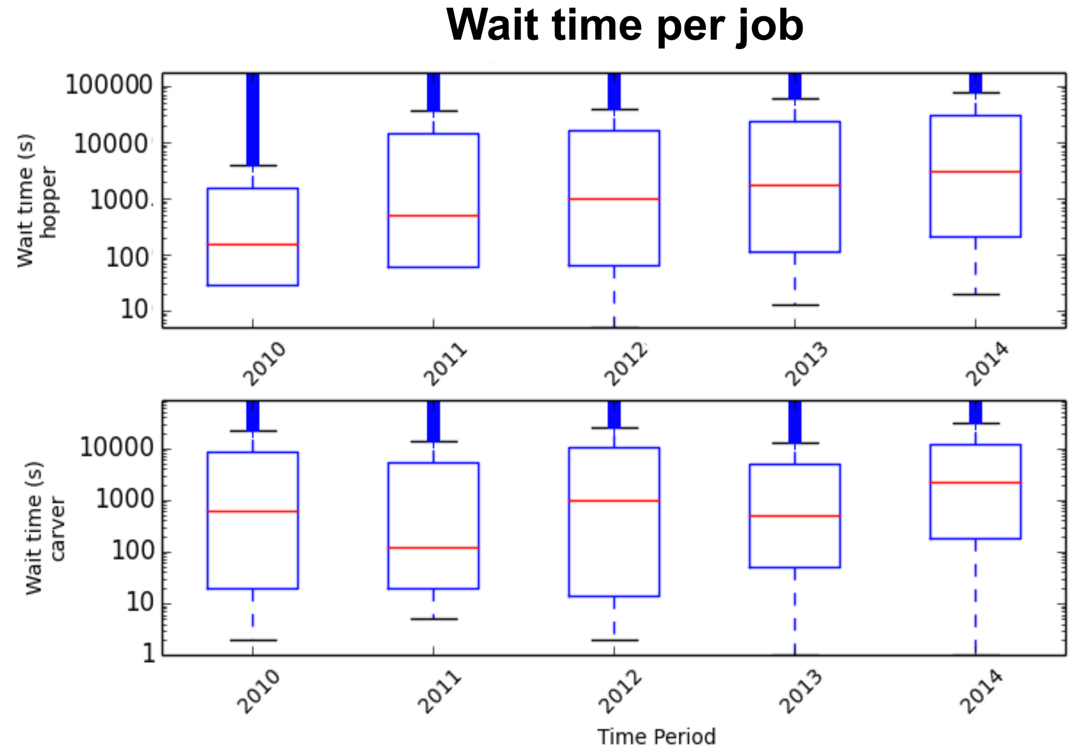
Number of cores per job



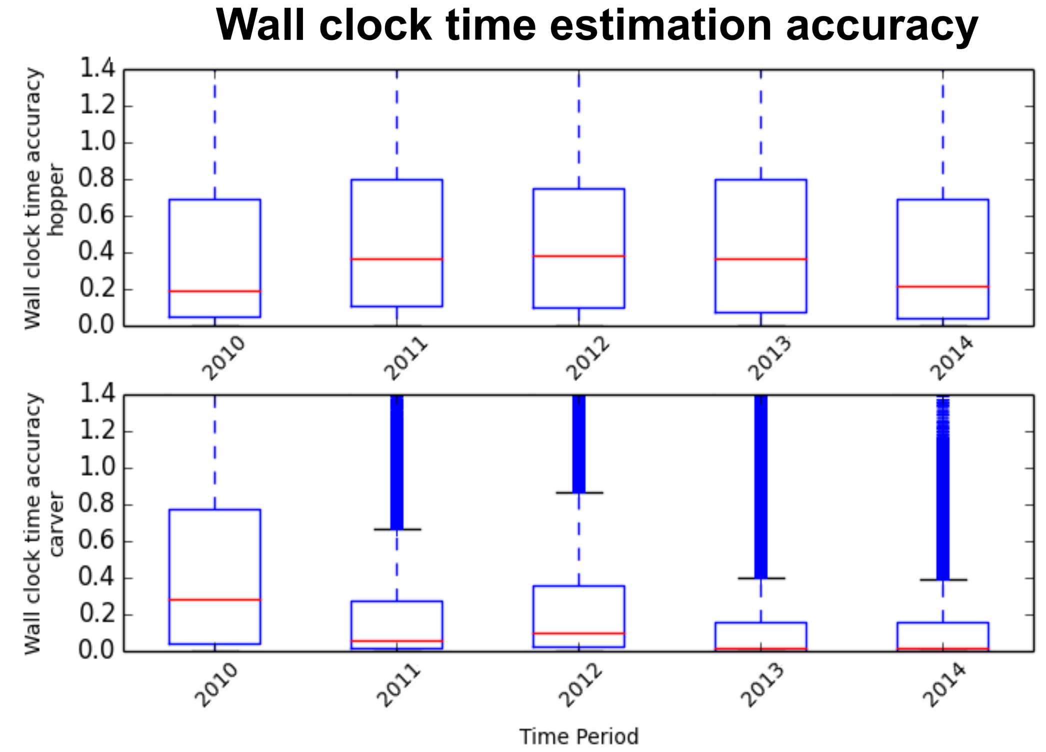
Core hours per job



Wait time per job



Wall clock time estimation accuracy



Two different systems with, initially, very different workloads become more similar as time advances. In general, Hopper was built to execute classical tightly coupled large parallel jobs. Carver is a high performance cluster that executes large parallel and serial jobs (1 core, long wall clock time). Each type of jobs is executed in a different resource partition with different configuration: node sharing between applications is enabled in the serial partition.

	2010	Job evolution from 2010 to 2014	
	Hopper vs. Carver	Hopper	Carver
Wall Clock	Longer jobs in Carver	Longer	Shorter
Number of Cores	Wider jobs in Hopper	Less cores	Slightly more (Median 1 core)
Core Hours	Bigger jobs in Hopper	No changes	Smaller jobs (But more jobs)
Wait time	Longer waits in Carver	Increase bigger than Carver	Increases
Wall clock accuracy	Lower accuracy in Carver	Peak in 2012	Decreases even more

Acknowledgements

- Office of Science of the U.S. Department of Energy, contract No. DE- AC02-05CH11231.
- Swedish Government's strategic effort eSSENCE.
- Seventh Framework Programme, grant agreement 610711 (CACTOS).
- Swedish Research Council (VR), contract number C0590801 (Cloud Control).