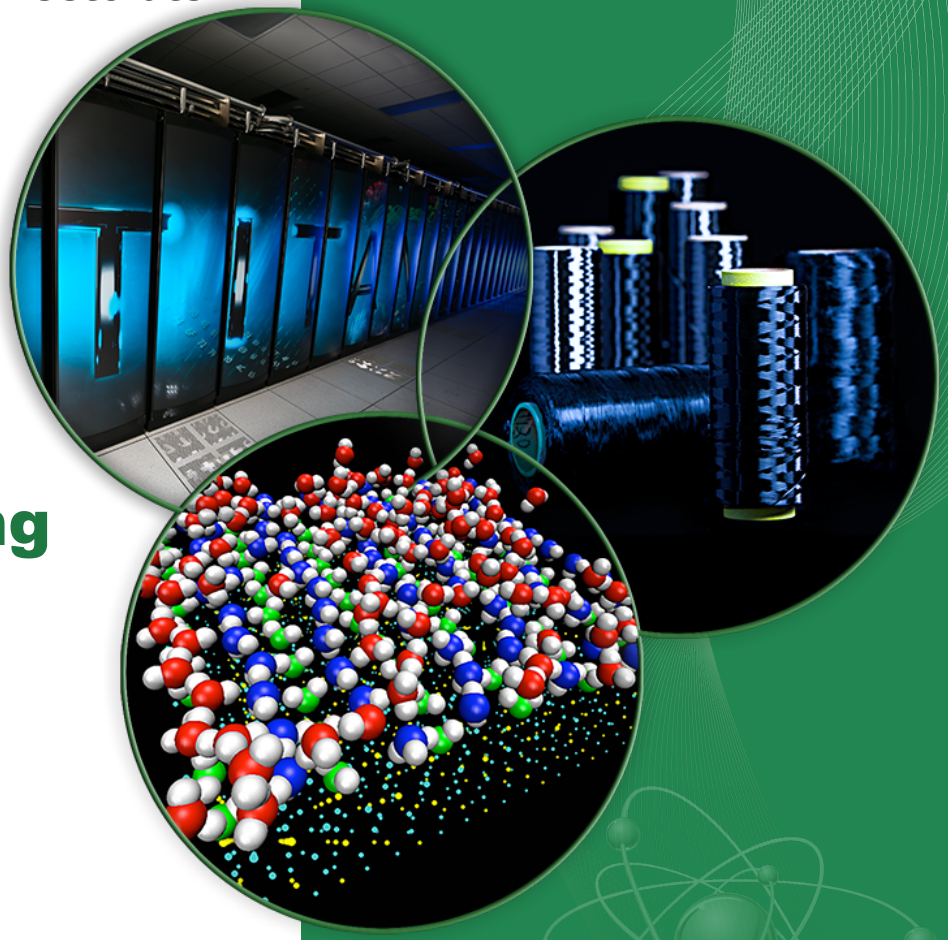
Oak Ridge National Laboratory

Computing and Computational Sciences Directorate



Intelligent Job Scheduling

Gregory A. Koenig (koenig@ornl.gov)
Oak Ridge National Laboratory

September 22, 2015

ORNL is managed by UT-Battelle
for the US Department of Energy

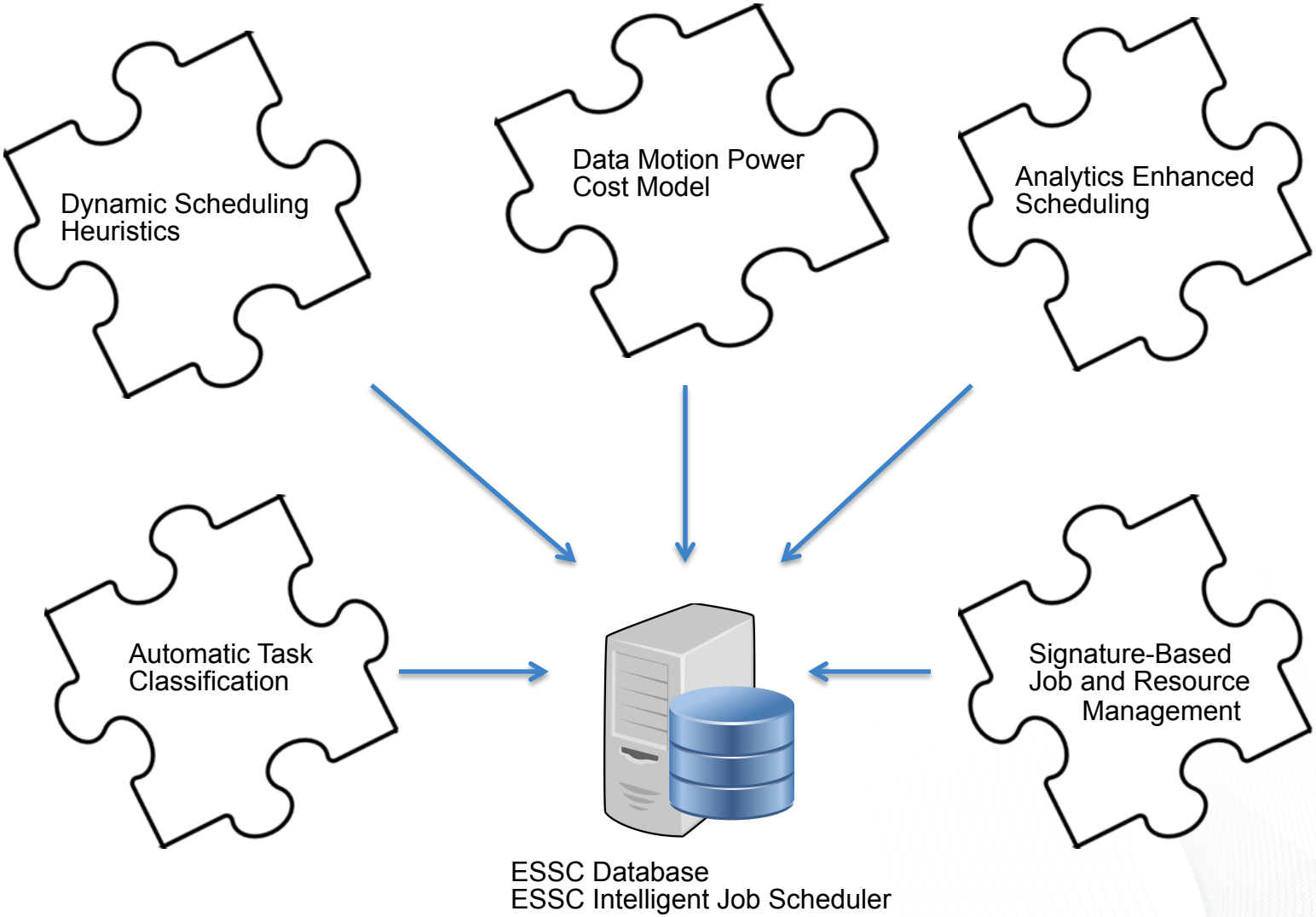
OAK RIDGE National Laboratory

Project Objective

- Advance the state of the art in system scheduling and resource management for extreme scale geographically distributed computing environments to make best use of limited resources
  - Computing resources: map jobs onto machines in such a way that maximizes the overall system performance measure, giving the most time-critical jobs high priority
  - Energy resources: map jobs onto machines in such a way that the overall system operates in the most energy efficient way possible
- Assumptions
  - The environment is oversubscribed, with significantly more work than computing resources
  - The computing environment consists of heterogeneous resources

OAK RIDGE National Laboratory

Project Components



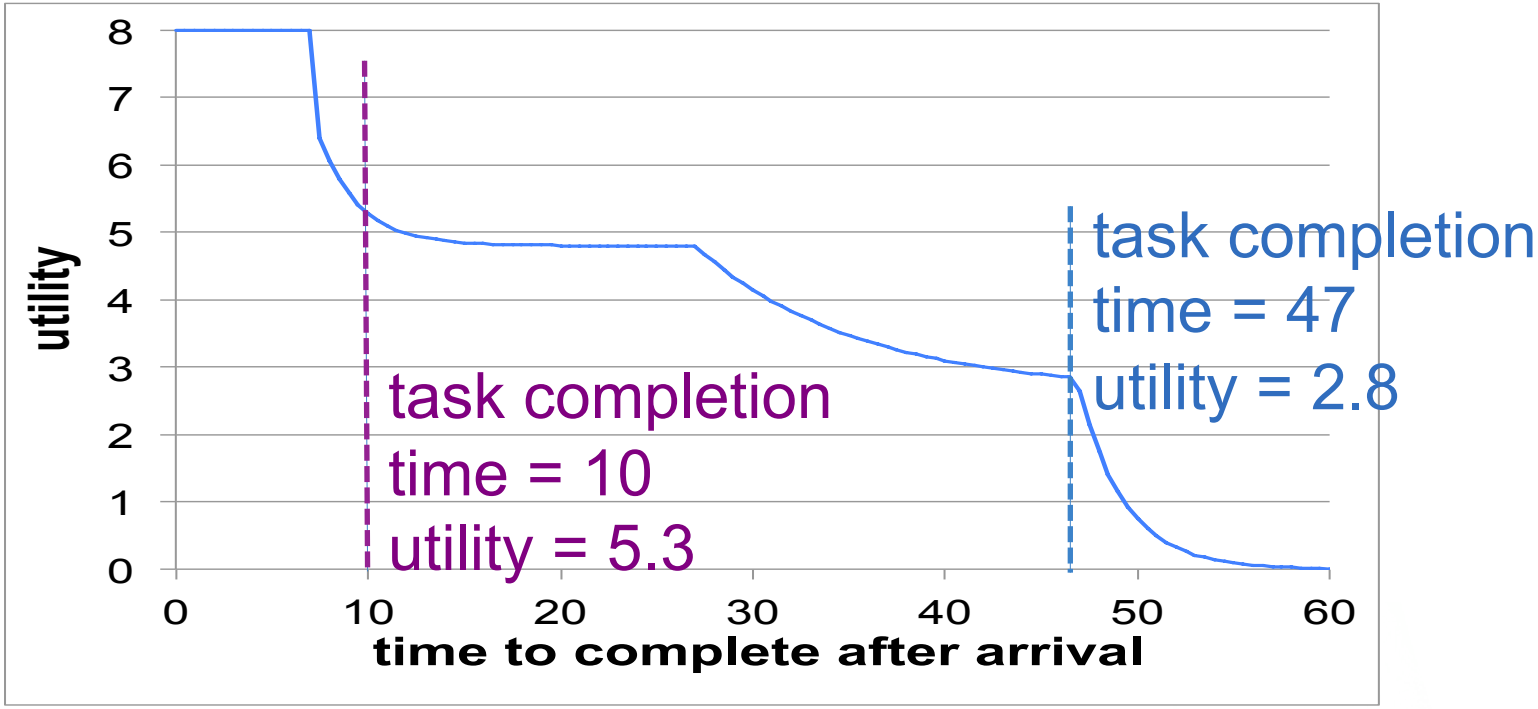
OAK RIDGE National Laboratory

Dynamic Scheduling Heuristics Research

- **Problem Statement:** A workload of diverse serial and parallel tasks arrive dynamically into an oversubscribed heterogeneous computing environment
- The resource manager has knowledge of
  - Estimates of execution time and average power consumption for each possible task-to-machine mapping
  - A monotonically decreasing utility function giving the time-decaying value for each task
- Goal of the resource manager
  - Maximize total utility earned from tasks during the day
  - May be constrained by energy budget for the day
- The scheduling problem is NP-hard
  - Cannot find optimal solution in reasonable time
  - Use heuristics
- Can use bi-objective optimization to study trade-offs between utility earned and energy consumed

OAK RIDGE National Laboratory

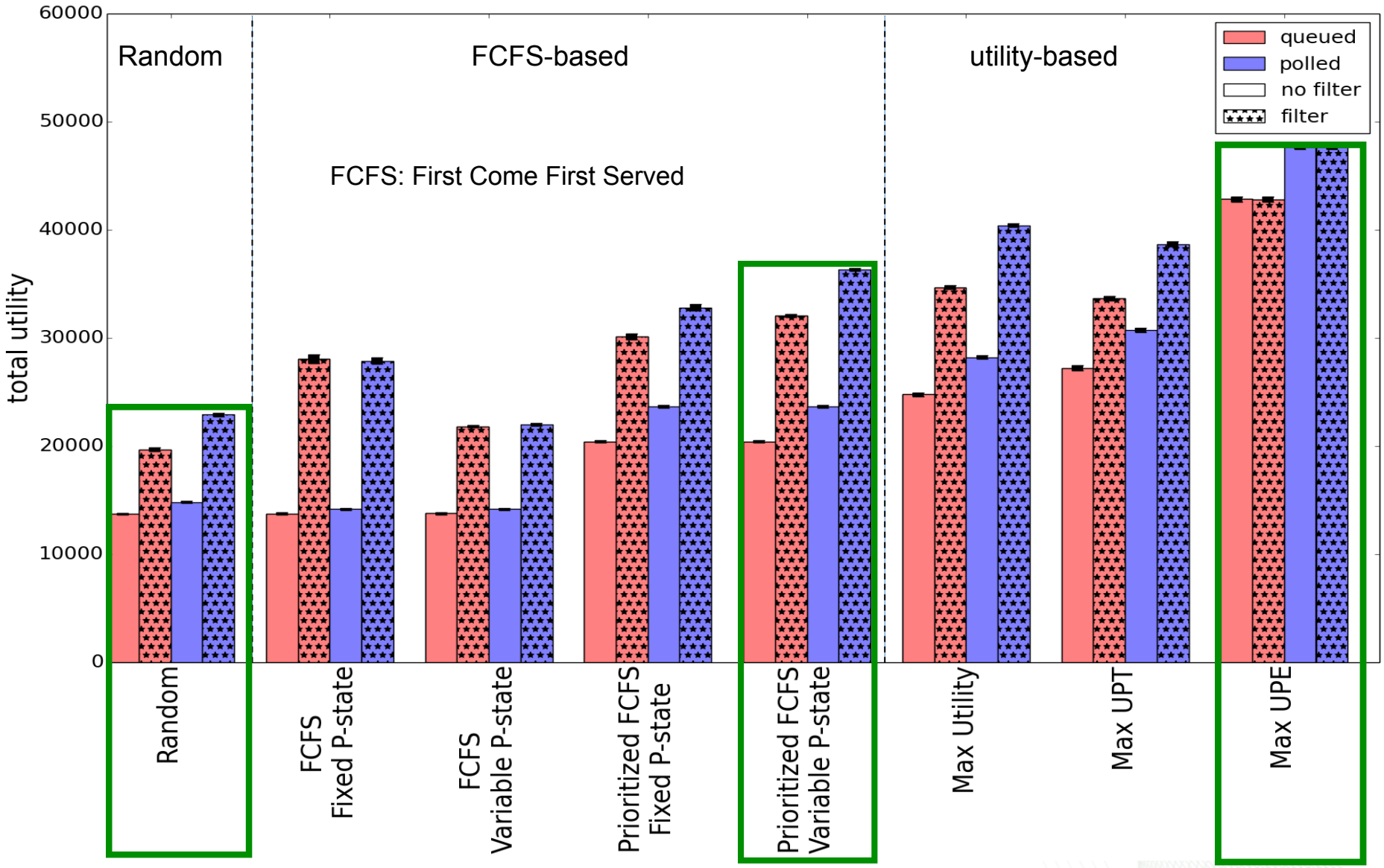
Measuring a Job's Relative Value: Utility



- Each job is characterized in terms of
  - Priority: the initial value of the job
  - Urgency: the rate of decay of the job's value
- Each potential task/machine pairing has values that describe the Estimated Time to Compute (ETC) and Average Power Consumption (APC)

Oak Ridge National Laboratory

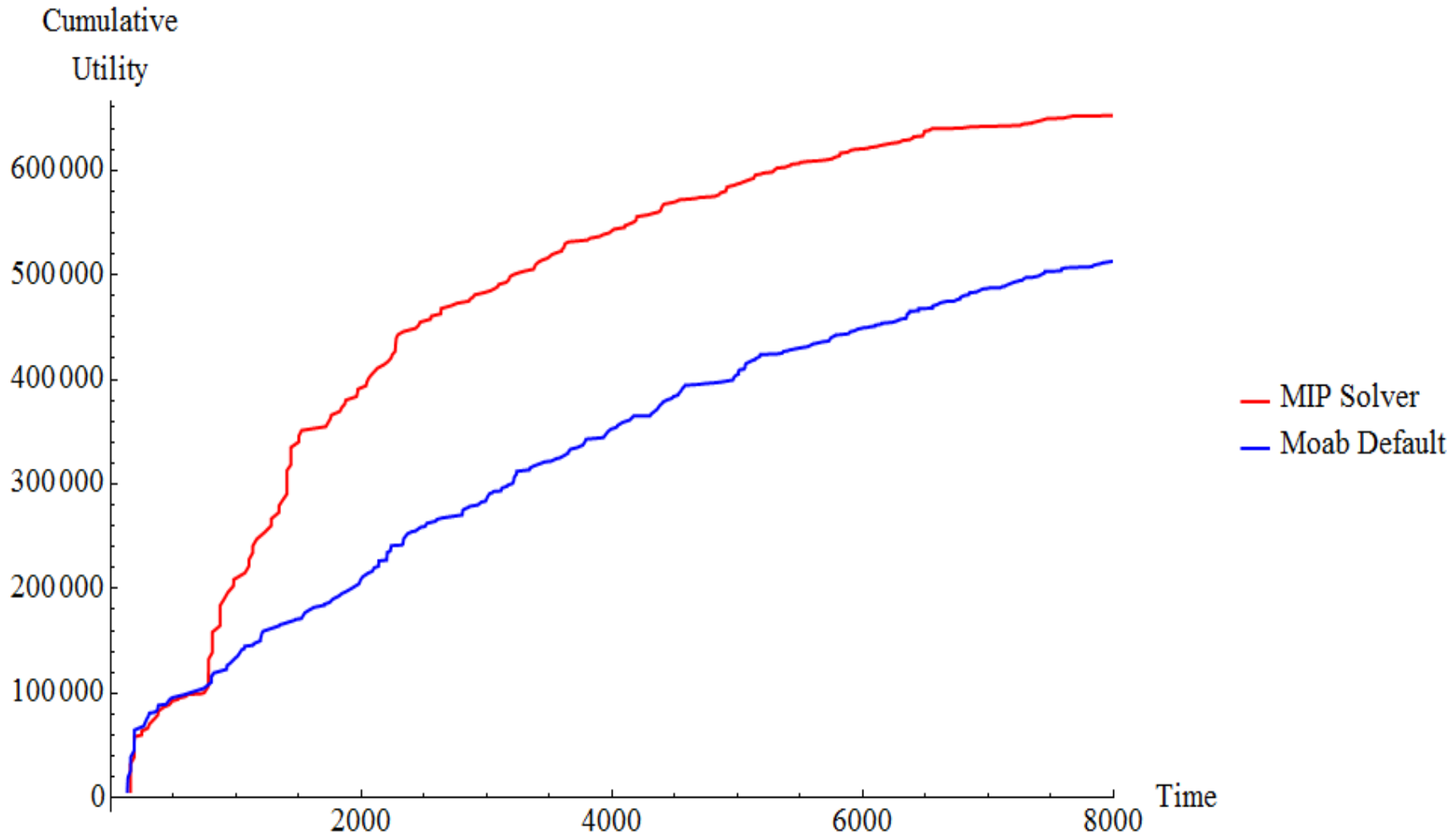
Example Simulation Study



- 36,000 tasks, 100 machines, averaged over 48 trials

OAK RIDGE National Laboratory

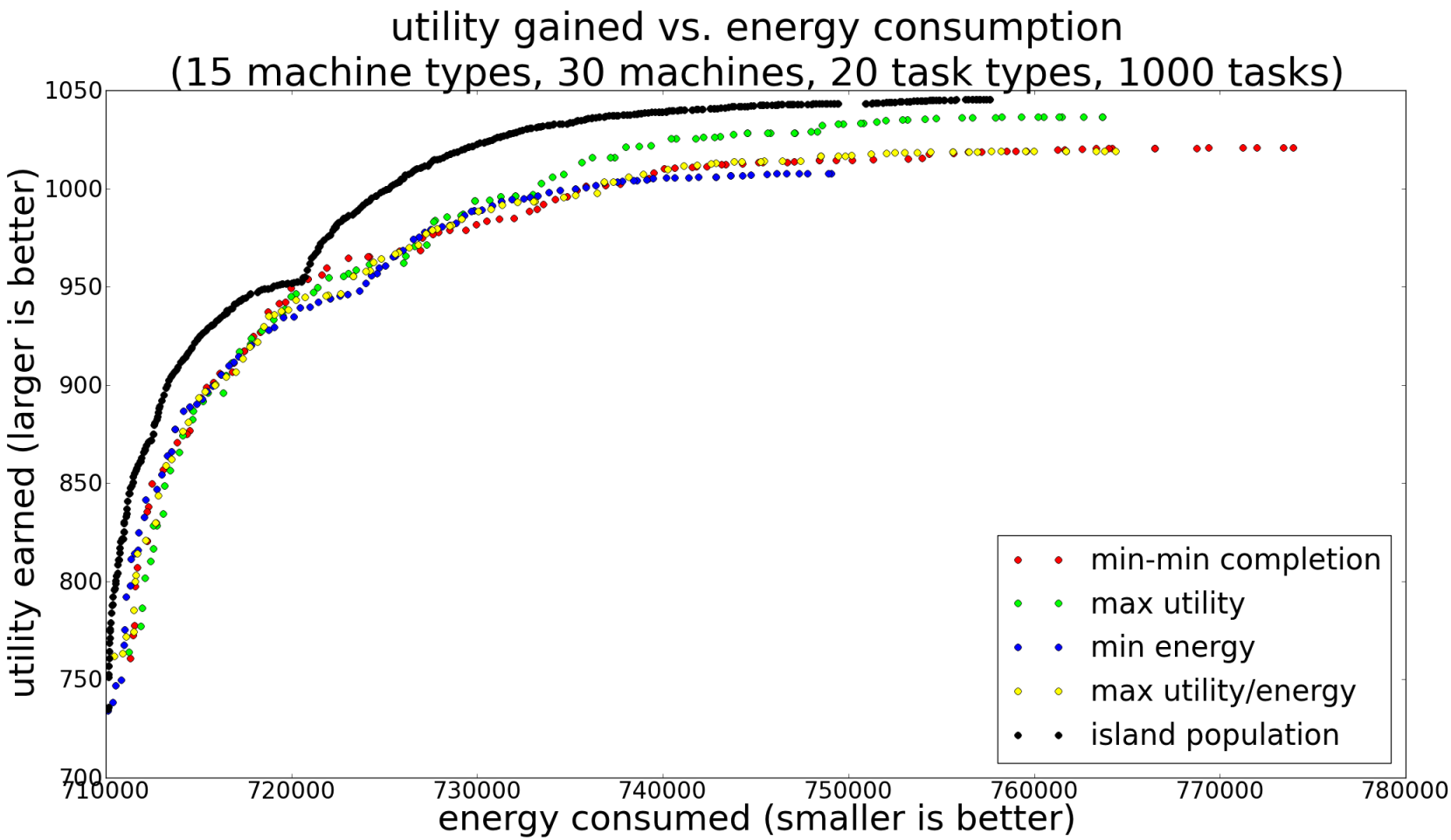
Comparison to Moab in ESSC Testbed



OAK RIDGE
National Laboratory

Source: H.J. Siegel, Colorado State University

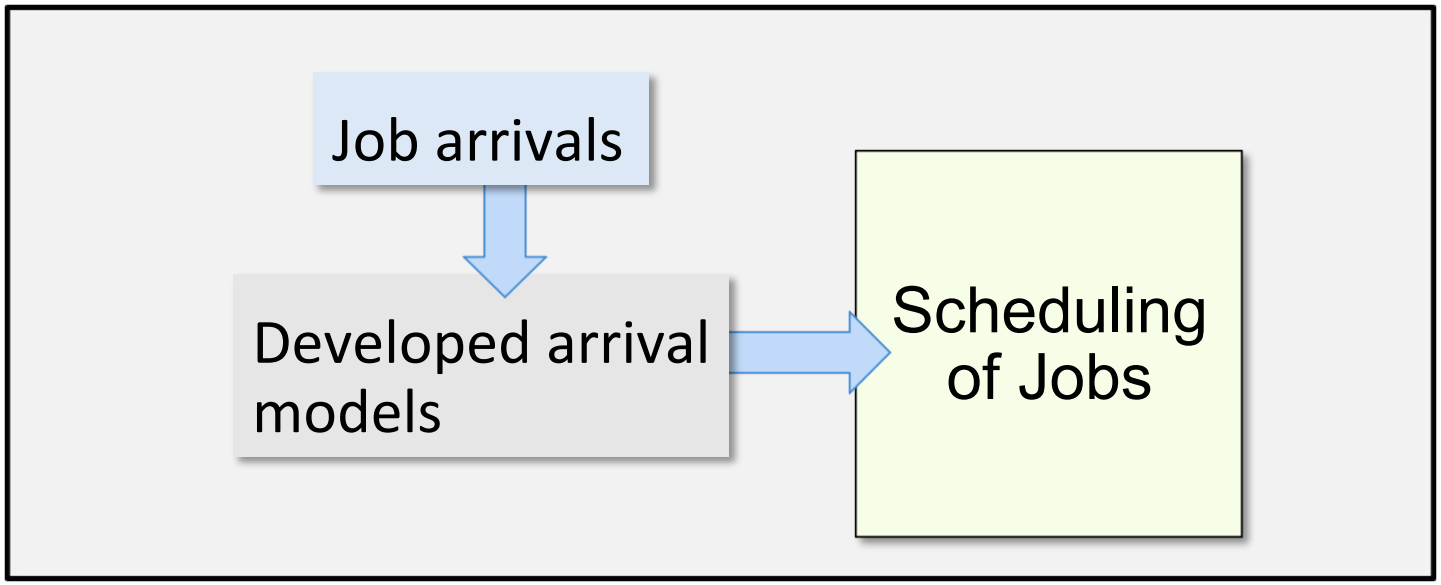
Trade-Offs Between Utility and Energy



- A Pareto front contains all the resource allocations <u>not</u> “dominated” by any other resource allocation

OAK RIDGE National Laboratory

Can we improve scheduling with predictive analytics?

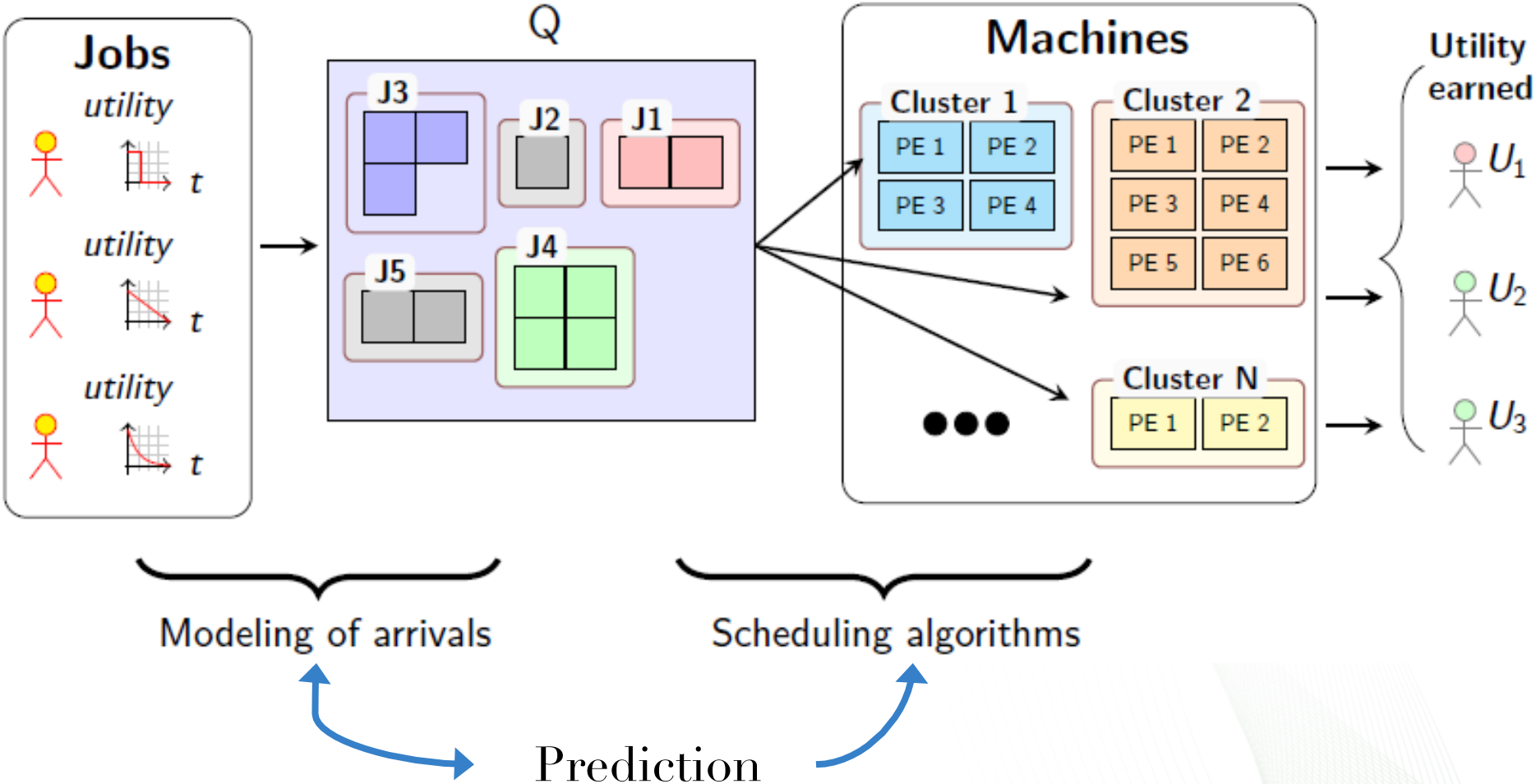


Hypothesis:

Having some knowledge of "what is coming" enables better scheduling

OAK RIDGE National Laboratory

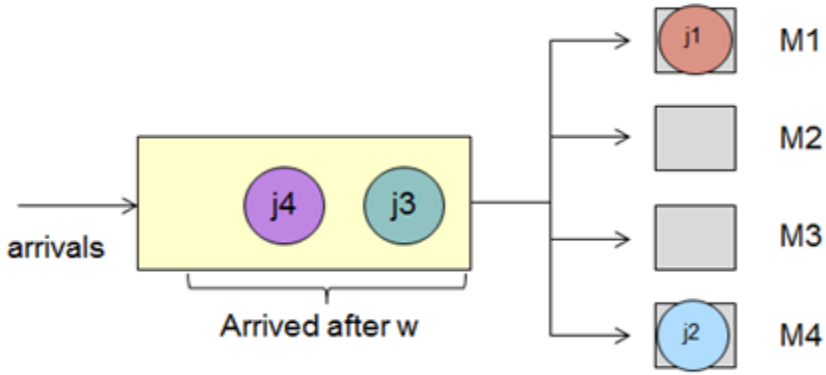
Problem can be formulated as a Mixed Integer Program



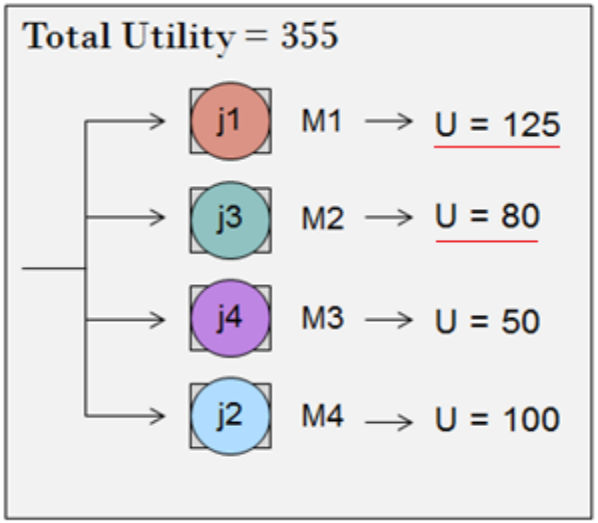
OAK RIDGE National Laboratory

A simple prediction illustration

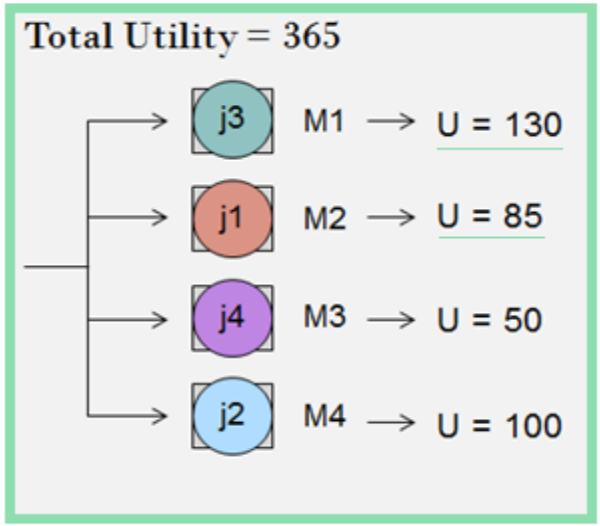
- Scheduling occurs at fixed intervals **w**
- A job may arrive just after the scheduling time **w**



No prediction



With Prediction



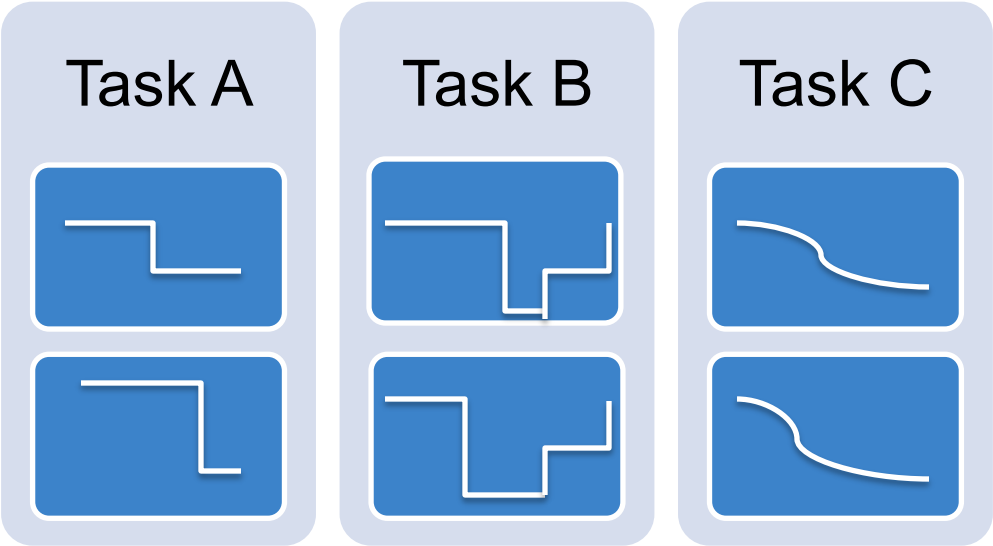
Take-away

Predictive analytics is beneficial (e.g., more effective machine utilization, higher utility)

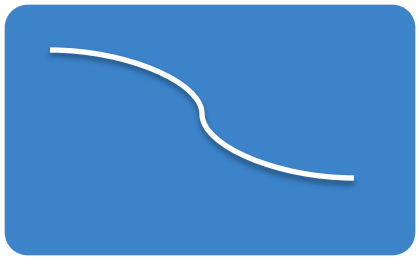
Oak Ridge National Laboratory

Classifying power traces of known tasks

Known power traces



Unknown trace



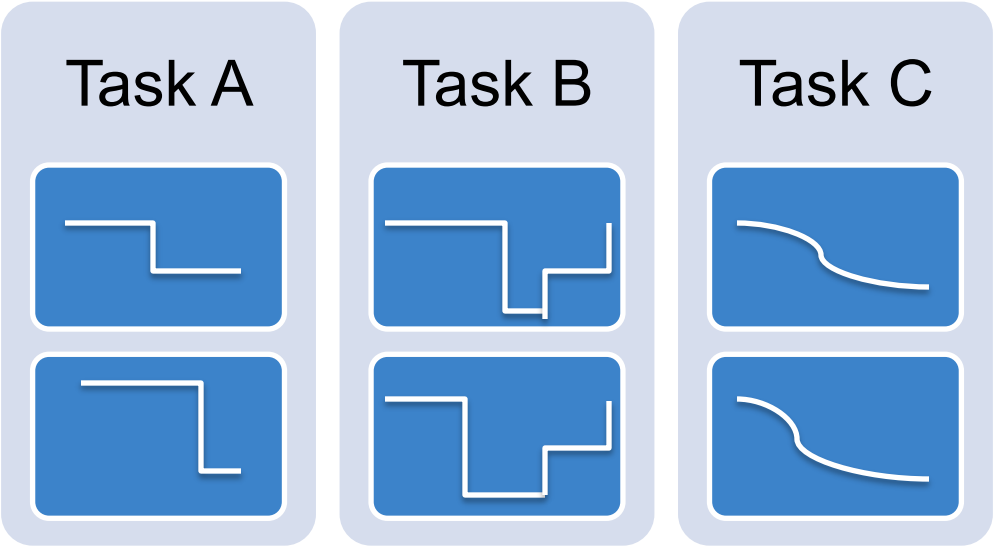
1. Compress each trace into a **signature**
   [DCSkewness, DCKurtosis, DCNonlinearity, DCSerialCorrelation, Hurst, kurtosis, lyapunov, max, mean, median, min, nonlinearity, skewness, standard_deviation, serial_correlation, trend]
2. Create a **classifier**
3. Guess which task generated each unknown trace

Accuracy for our ~220 traces: 85-95% from very simple models

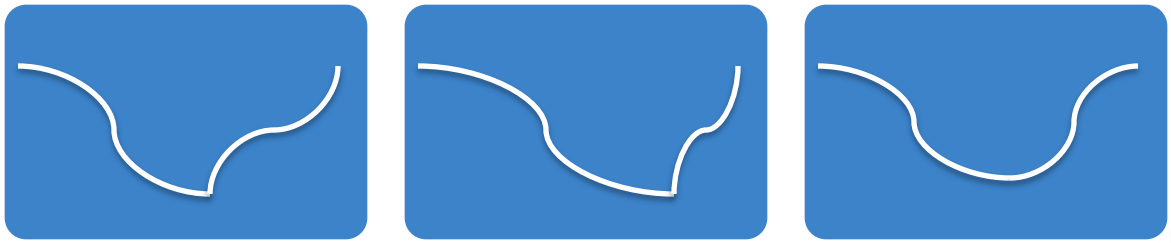
Oak Ridge National Laboratory

Adapting to new task patterns

Known power traces



Recent unknown traces



1. Quantify **certainty** of task identification
   - To inform scheduling decisions
   - To recognize new tasks

Result: Certainty values track prediction accuracy.

2. Detect **novelties**: have I ever seen this task before?

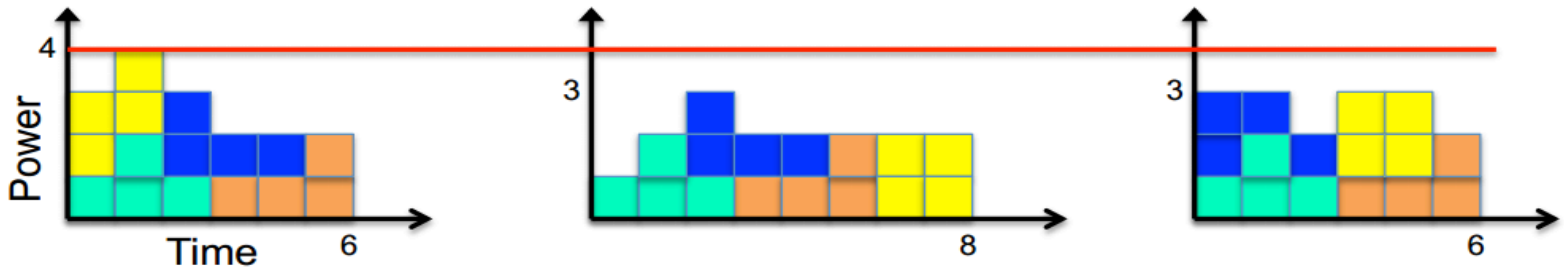
Result: Detect 80% of novelties (even with similar CPU-bound workloads), with 4% false positives

3. Dynamically adjust training set: group related novelty traces into new tasks, so they can be recognized in the future

OAK RIDGE National Laboratory

Signature-Based Resource Management

- Viewing job scheduling as a Tetris game

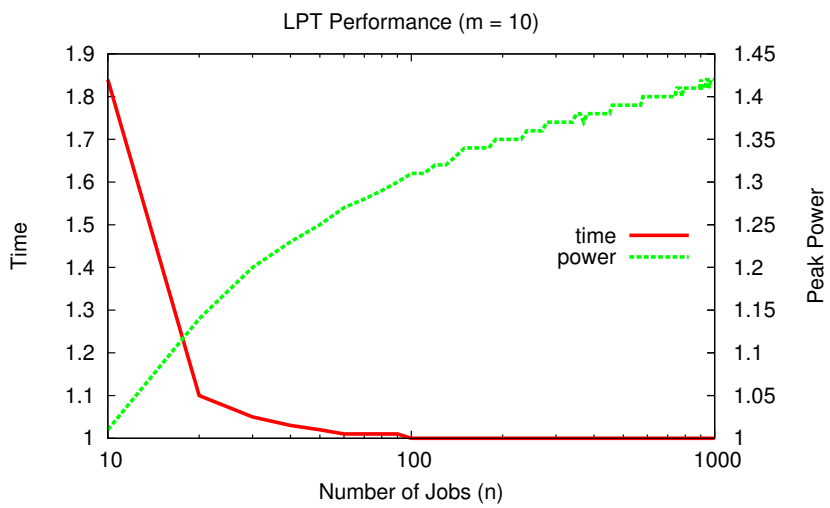


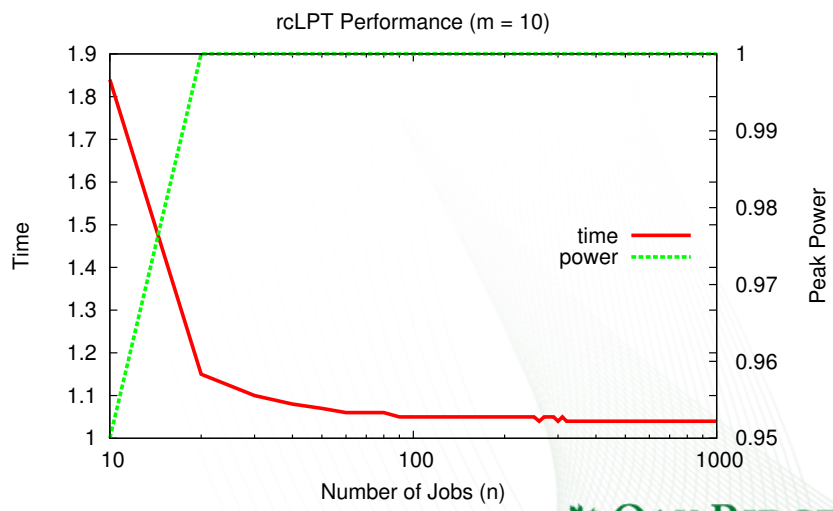
- Driven by power usage patterns. Can we:
  - Associate a pattern with each application?
    - Do applications exhibit consistent power behavior?
    - How can we represent an application's power behavior?
    - How similar is that behavior across different runs, inputs, configurations, and hardware?
  - Enhance scheduler with pattern information?

OAK RIDGE National Laboratory

Signature-Based Optimization

- Computing optimal schedules is very difficult.
- Traditional heuristics are good at minimizing makespan but not peak power demand.
- Modern heuristics constrain peak power too hard, resulting in longer makespan.
- Appropriate problem formulation is the key!

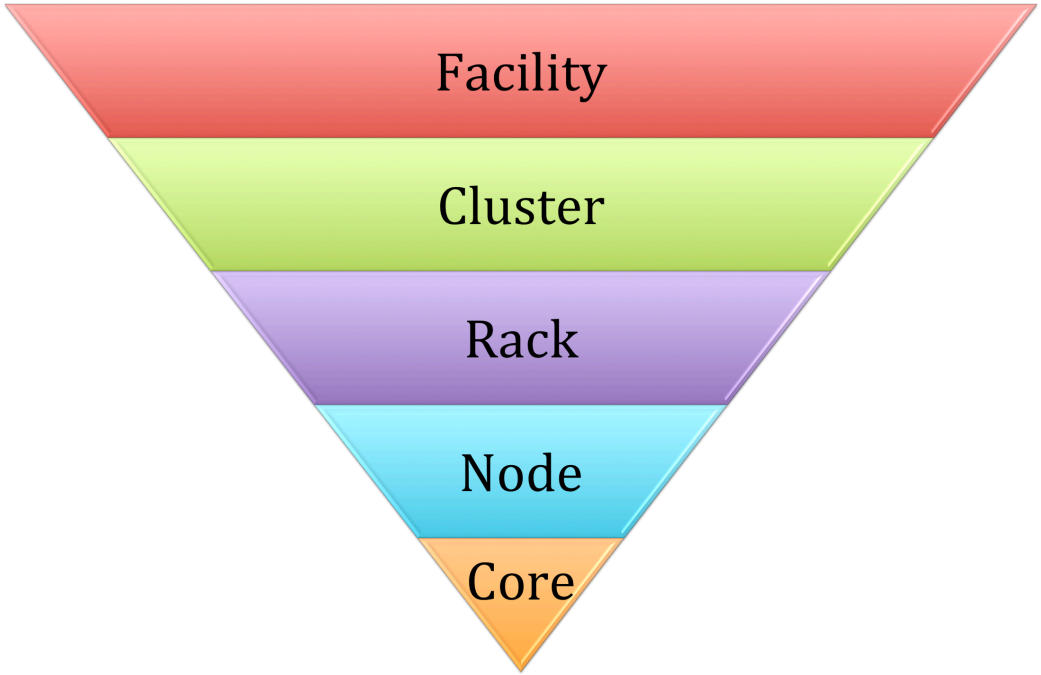




Oak Ridge National Laboratory

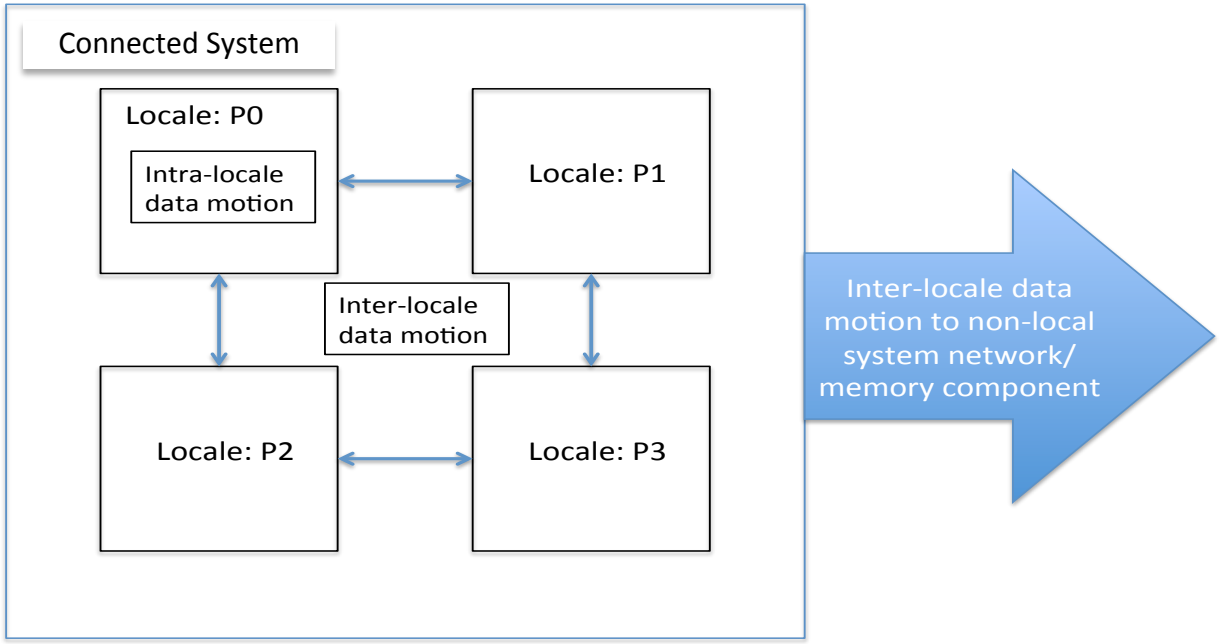
Data Motion Power Cost Models

- Incremental derivation based on hierarchical representation of data motion



Approach

- Express each level of hierarchy as a system of connected components
- Functionally describe power consumption for components essential to data movement



Possible Future Directions (1)

- Allow the Intelligent Job Scheduler to preempt jobs
  - Currently, once a job has been placed onto resources, it remains on those resources until it completes
  - Preemption allows jobs to be suspended in favor of other more valuable jobs that arrive into the work queue
  - Important to constrain the conditions under which preemption happens so the schedules are solvable
- Introduce demand-response capabilities
  - Demand-response allows an HPC facility to cooperate with its Energy Service Provider in consuming energy
  - Exchange knowledge of when energy demand might be high or when supply might be constrained
  - Understanding local demand can be useful in determining whether to move a job to a geographically different location

OAK RIDGE National Laboratory

Possible Future Directions (2)

- Address issues related to trapped capacity
  - Pack jobs onto resources with awareness of both utility and trapped capacity, trying to reduce trapped capacity
  - Can we improve overall trapped capacity by moving waiting jobs to sites that have a high measure of trapped capacity?
- Allow topology-aware placement of tasks
  - In addition to scheduling jobs on the basis of utility and energy, schedule and/or place jobs onto resources to take into consideration the network topology connecting the resources
  - Topology-aware placement can reduce latency in parallel jobs, improving the time required to complete a given job

OAK RIDGE National Laboratory

Acknowledgements

This work was supported by the United States Department of Defense (DoD) and used resources of the DoD-HPC Program at Oak Ridge National Laboratory.