Optimization Algorithms for Large-scale, Thermal-aware Storage Systems

Amol Deshpande
Samir Khuller
Ankur Srivastava

UNIVERSITY OF MARYLAND
Supported by NSF HECURA 2009-2012

Energy Usage in Data Centers

- As machines/cores/clock rate increase, we are using more and more energy.
- As load increases, the machines get hotter and we expend significant energy ($) in cooling the data center.
- Can we reduce energy costs by selectively shutting off machines (sleep state) OR by clever load balancing (data replication)?

Different aspects of our project

- Optimize Energy Usage for large scale data processing applications.
- Consider load on processors, disks, and thermal effects due to adjacent processors.
- Total Energy used is a function of load as well as cooling energy required (a significant part of the total energy used).

Different Aspects of the Project

- Machine activation: some components may be shut off completely to save energy.
- Data Layout: exploiting locality of data access.
- Load rebalancing: power consumed directly depends on workload, this can be reduced by rebalancing workload -- consider thermal effects from adjacent processors while defining load distribution policy.

Machine Activation

- Traditional scheduling theory deals with a given set of jobs, machines and constraints on jobs (deadlines, release times etc).
- We add a new dimension with “activation cost” for each machine – given an activation cost budget, which subset of machines should we activate so that we can find a good schedule on the set of active machines?

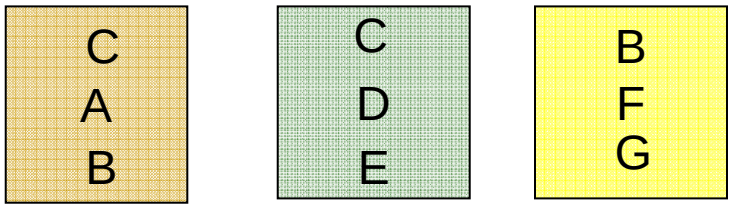
Machine Activation Results

- Many scheduling problems are NP-hard – we develop and analyze heuristics for these problems, mainly focusing on approximation algorithms.
- Ideally, we would like to be able to do this in an ONLINE setting – where jobs are arriving and departing over time.
- Workload is typically cyclical, with infrequent peaks and deep valleys.

Data Layout Optimization

- We assume a very general framework
  - We have a large collection of data items and need to process a collection of queries.
  - Each query Qi needs to access a subset Si of the data items.
  - The cost for each query is the number of storage units (disks, memory banks etc.) that it needs to access to run the query.

Data Layout Optimization



- Queries are as follows:

Q1={A,B,D,C} cost(Q1) =2

Q2={B,F,G} cost(Q2)=2

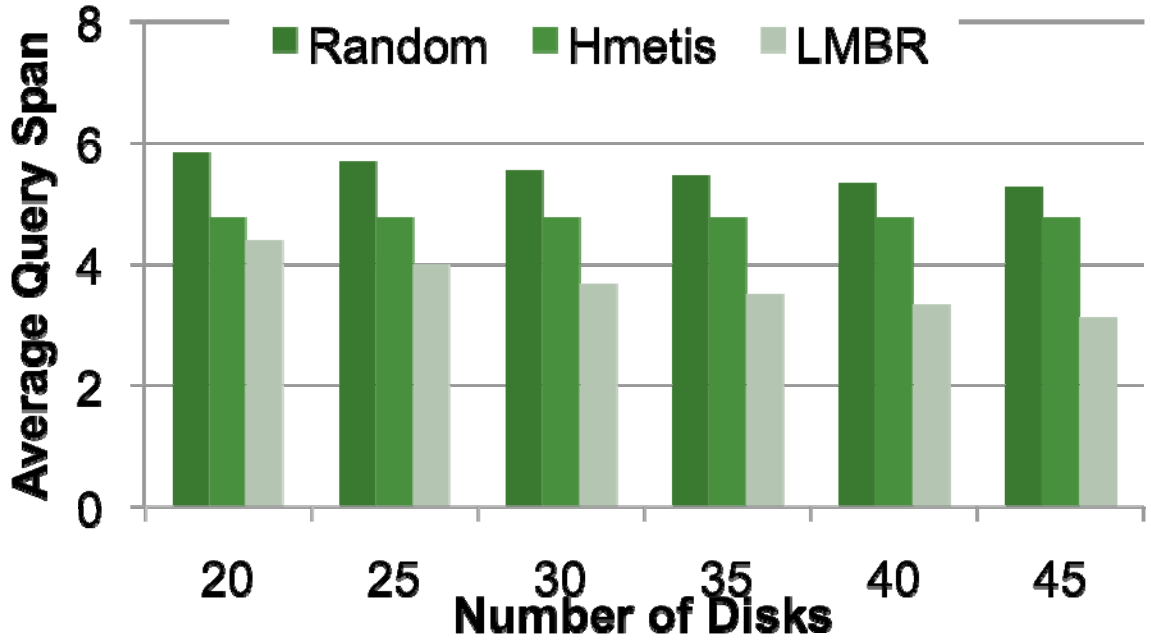
Q3={C,D,E} cost(Q3)=1

Q4={F,G} cost(Q4)=1

Q5={A,B,F,G} cost(Q5)=2

A DIFFERENT LAYOUT MAY HAVE A DIFFERENT COST

Preliminary Results for Data Placement

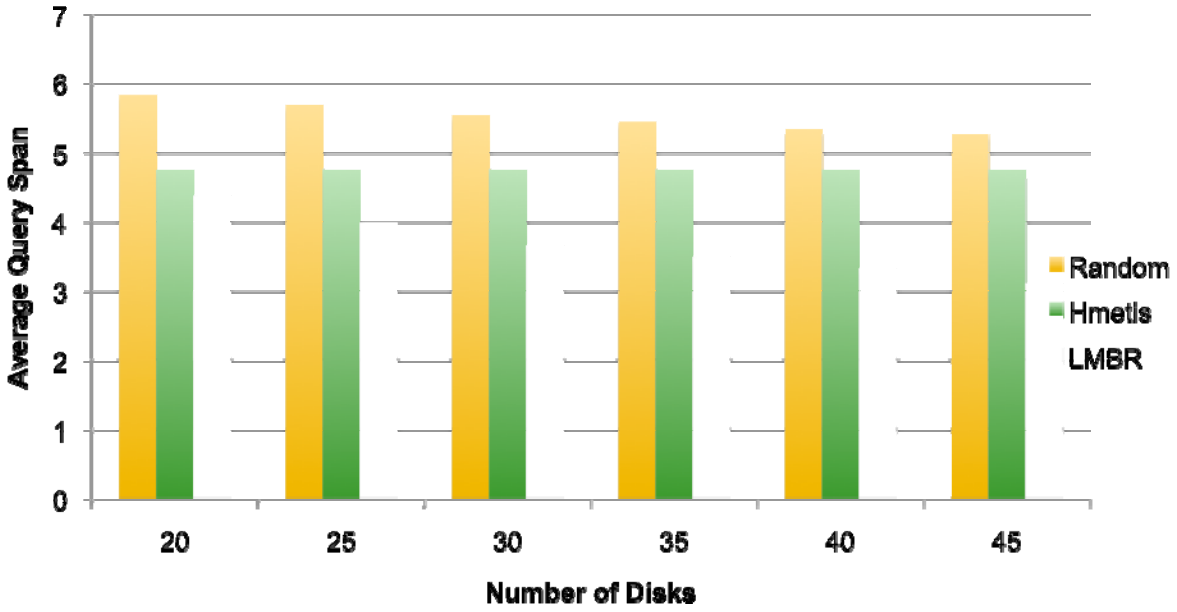


HMETIS: State-of-the-art Hypergraph Partitioning Algorithm

LMBR: A greedy algorithm that does sophisticated local moves

Significant energy savings by doing workload-driven optimization

Preliminary Results for Data Placement



HMETIS: State-of-the-art Hypergraph Partitioning Algorithm

LMBR: A greedy algorithm that does sophisticated local moves

Significant energy savings by doing workload-driven optimization

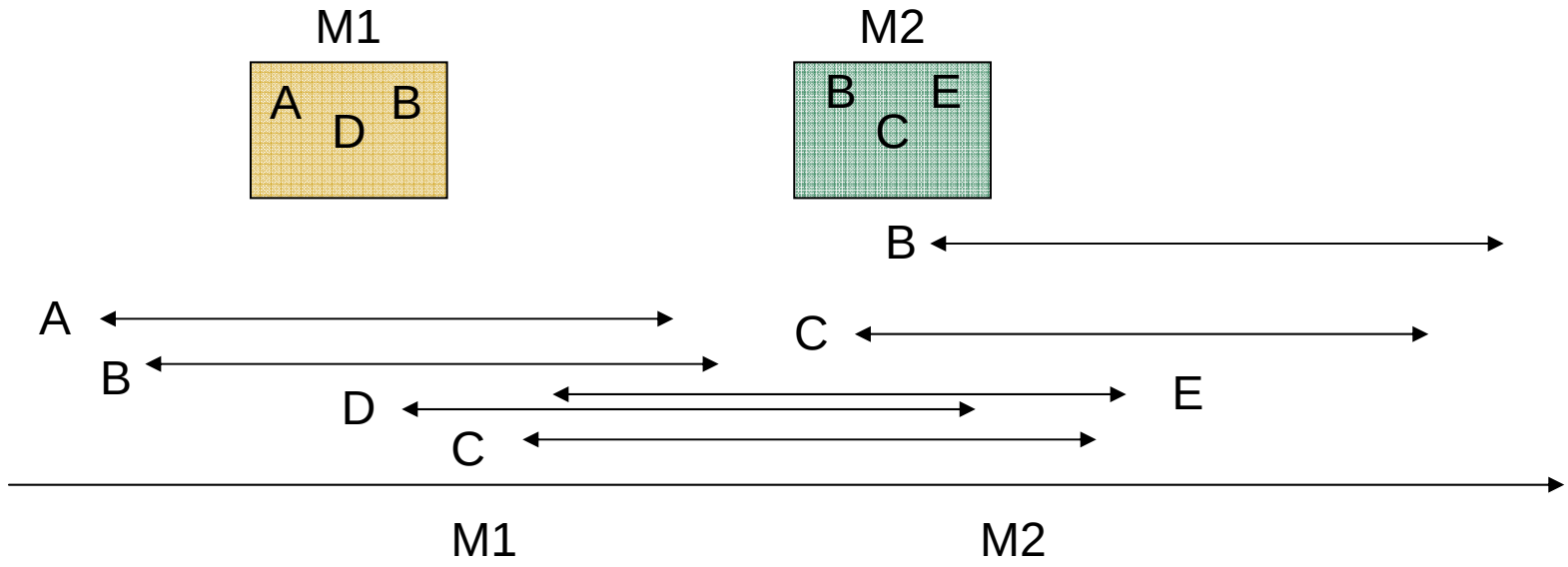
Model it as a Hyper-graph problem

Scheduling Optimization (fixed layout)

- Queries have a release time and a deadline and unit processing time.
- We need to process all the queries, and when we schedule a query we need to activate the memory banks that contain the data that the query needs to run.
- At each time step we can run upto P queries.

How can we run the queries using the least amount of energy?

Scheduling Queries

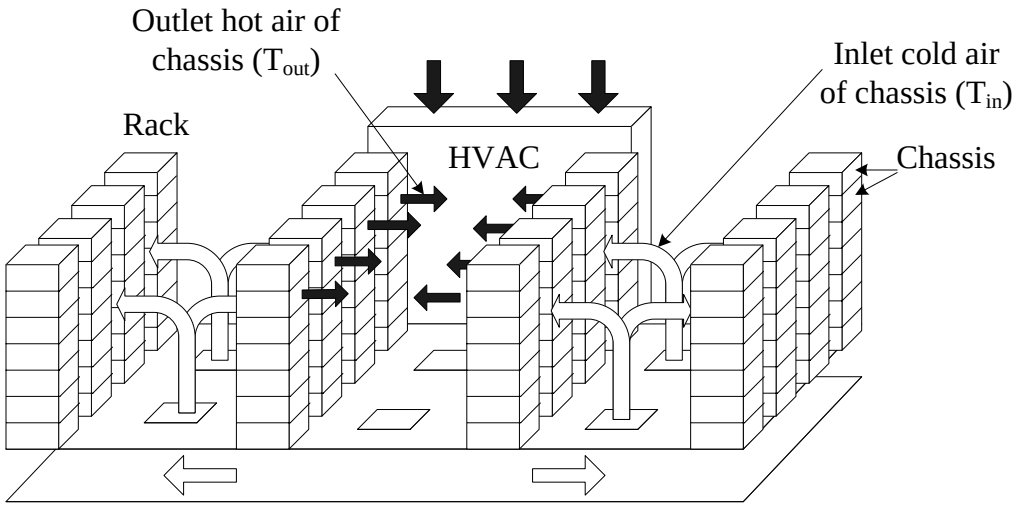


- Goal is to activate memory to answer the queries and to minimize memory activation.
- Memory has two states – sleep state and active state.
- Again the resulting problem is NP-hard, we develop a new heuristic for the problem.

Thermal effects

- Power consumed depends on workload and the primary goal is to ensure that the max temperature is not very high.
- However temperature is a function of local load and load on close machines.
- Clever load balancing lets us raise the temperature of the incoming air from the cooling system and reduces AC cost significantly (saving energy).
- We study workload balancing by exploiting the fact that several copies of the data are present.

Datacenter thermal model



Relationship between power consumption of a data node on chassis, and its inlet, outlet air temperature (Tsup: temperature of the cooling air produced by A/C):

$$K_i T_{out}^i = \sum_{j=1}^{n} a_{ji} K_j T_{out}^j + (K_i - \sum_{j=1}^{n} a_{ji} K_j) T_{sup} + P_i$$

Disk power consumption on data nodes (Pi is constant since we assume disks work on same speed): $P_i \propto N\omega^{2.8} D^{4.6}$

A/C power consumption: $P_{AC} = \frac{\sum_{\forall i} P_i}{COP}$ $COP = 0.0068 T_{sup}^2 + 0.0008 T_{sup} + 0.458$

Total power consumption:

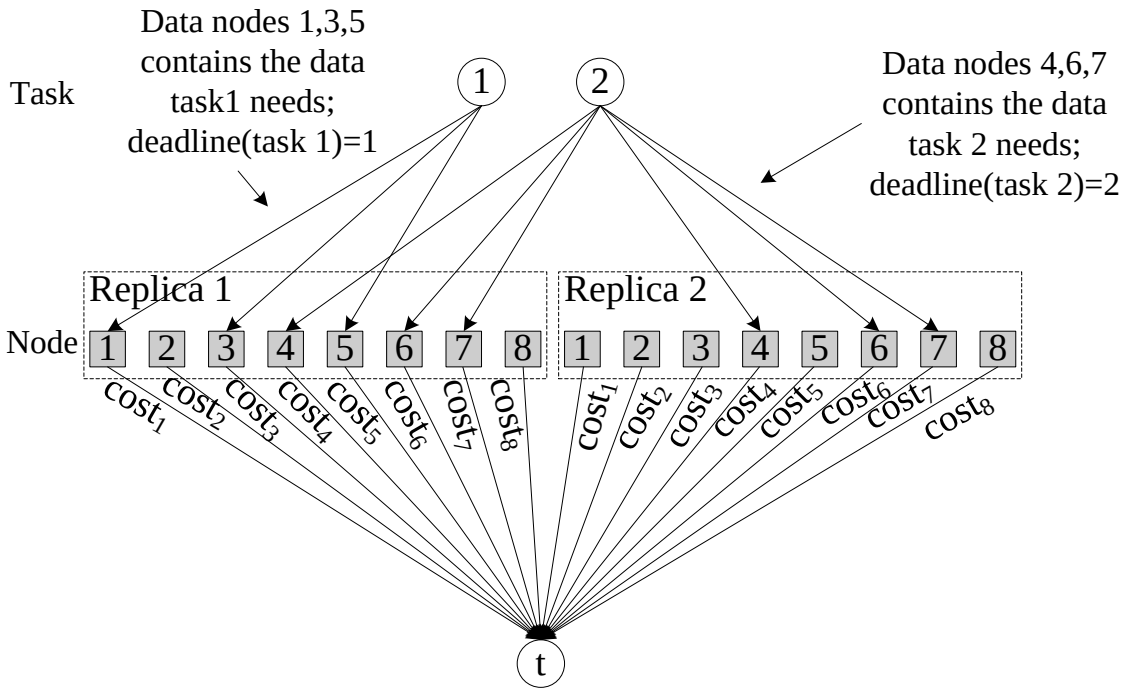
$$P_{total} = P_{AC} + \sum_{\forall i} P_i = (1 + \frac{1}{COP}) \sum_{\forall i} P_i$$

Load Balancing

- Given: *n* data nodes, *d* data intensive tasks, each task takes 1 unit of time to execute, the deadlines for the tasks range from 1 to *q:*

- We would like to assign each data access task to a data node containing the data it needs, so that each task can read its required data locally.
- We would like to minimize the power consumption of the system, while ensuring the temperature of data nodes to be within the maximum constraint.

Solve ILP with minimum cost flow



- source nodes ①② represent tasks (flow available = 1).
- transshipment nodes 1 - 8 represent data nodes in datacenter. Each data node is replicated *q* times (*q*=2 in this figure, data nodes in replica 1 or 2 indicate the data nodes at time 1 or 2.)
- sink node (t) (demand = number of tasks)

- Edges:
- task → data node storing the data this task needs and whose corresponding deadline is not larger than the deadline of this task (capacity 1, cost 0).
- data node i→ sink node (capacity 1, cost: cost(i))

$$cost(i) = K_i \sum_{j=1}^{n} a_{ij} - (K_i - \sum_{j=1}^{n} a_{ji} K_j)$$

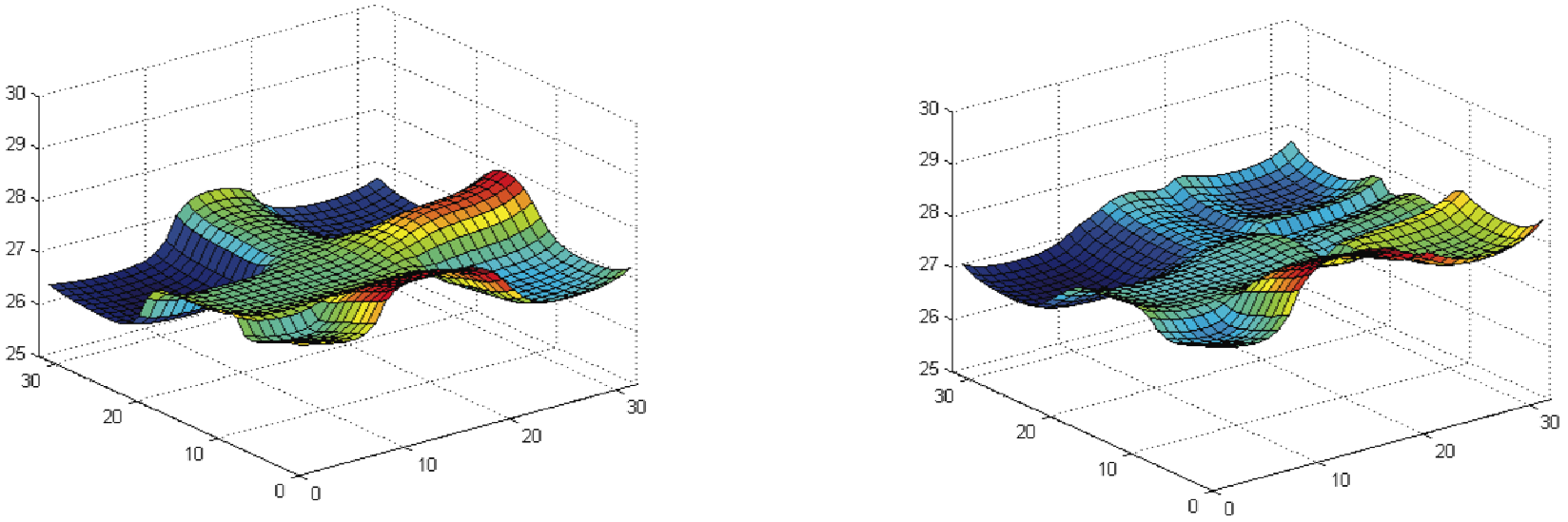
Results

Table I
COMPARISON OF OUR METHOD AND *random* METHOD WHEN TASKS HAVE DIFFERENT DEADLINES

Task number	Algorithm	Rescheduling	Runtime(s)	Total power (W)	$T_{sup}(K)$	$mean(T_{out})$(K)	$stdev(T_{out})$
5000	Our method	No	5.949	5.6178e4	305.6005	317.2310	1.1657
	Our method	Yes	35.947	5.5737e4	306.9471	317.8895	0.7927
	Random		5.477	5.6593e4	304.4631	316.6095	2.0848
10000	Our method	No	10.347	1.3071e5	293.3775	313.7472	5.7069
	Our method	Yes	36.984	1.2396e5	296.4670	314.8426	5.7005
	Random		23.972	1.4091e5	290.1834	314.4067	5.5720
15000	Our method	No	21.7324	3.2904e5	280.5644	313.0872	5.5704
	Our method	Yes	62.544	2.5748e5	284.8338	313.3713	4.5626
	Random		24.349	4.3561e5	276.2348	311.6551	5.1763

- Mean(Tout) is the average outlet air temperature of chassis on different racks; stdev(Tout) is the standard deviation of outlet air temperature of chassis on different racks
- The system power consumption in our method is lower than *random* method by 28.13% on average.

Thermal Profiles in a Data Center

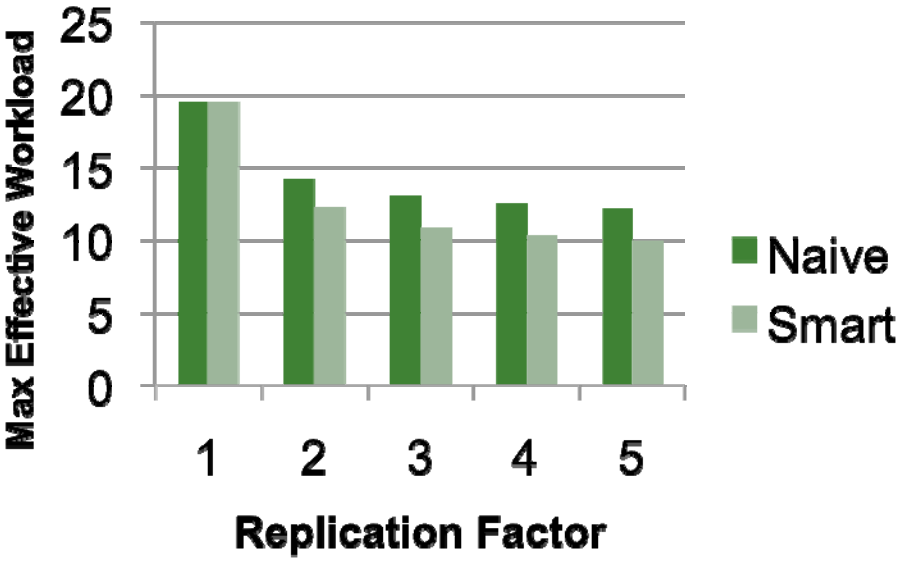


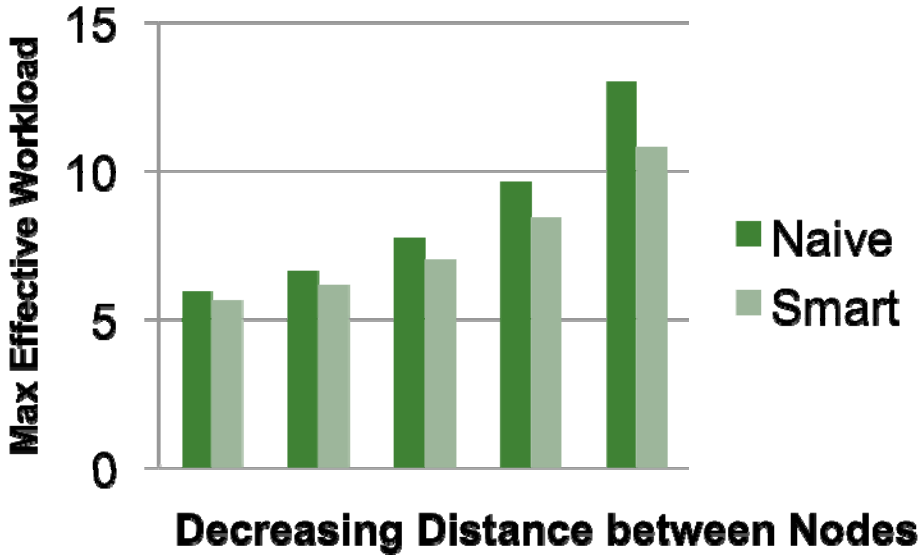
Simulated thermal profiles *5 minutes apart*

- Processing units located at the grid points
- Standard power dissipation model assumed

Large, rapidly changing temperature variations

Spatially-aware Optimization

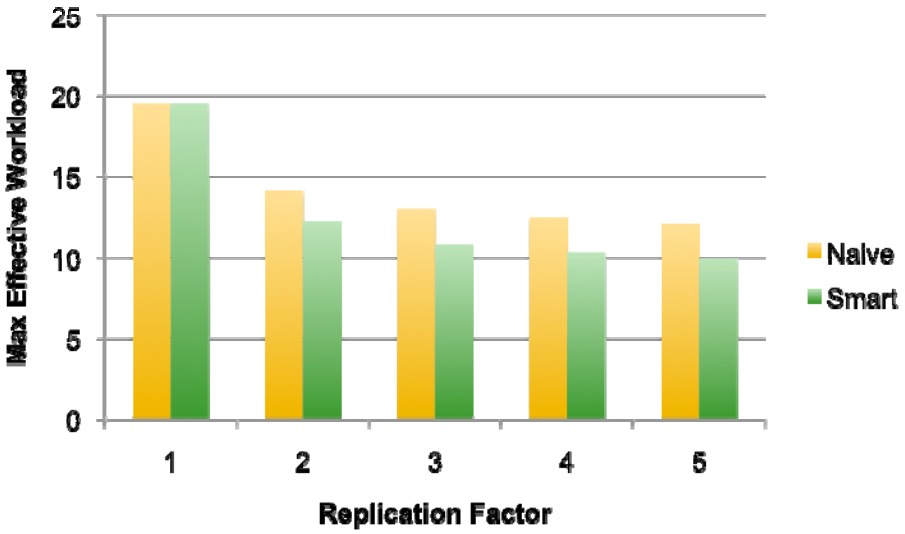


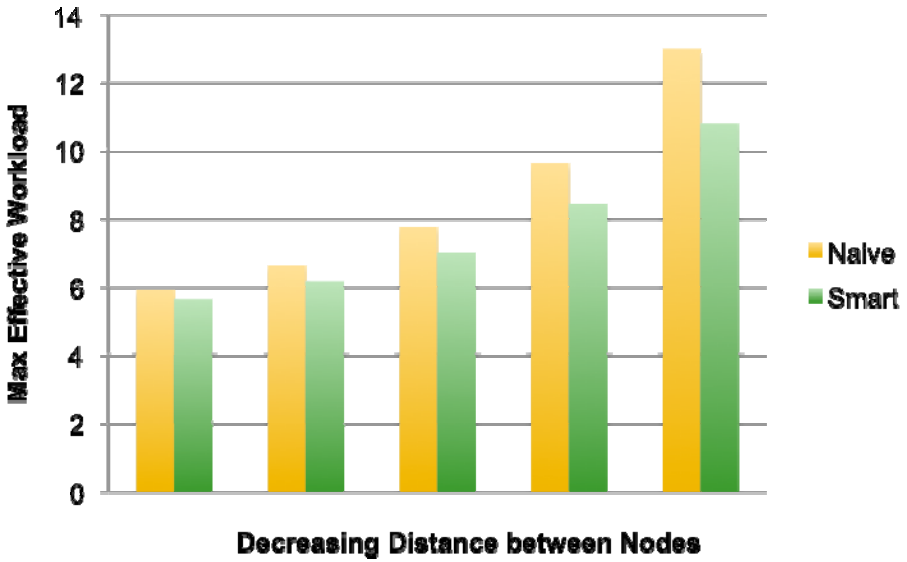


Comparing the performance of a greedy heuristic for task assignment: NAÏVE ignores spatial locality, SMART doesn't

Max effective workload roughly correlated with the node temperature

Spatially-aware Optimization





Comparing the performance of a greedy heuristic for task assignment: NAÏVE ignores spatial locality, SMART doesn't

Max effective workload roughly correlated with the node temperature

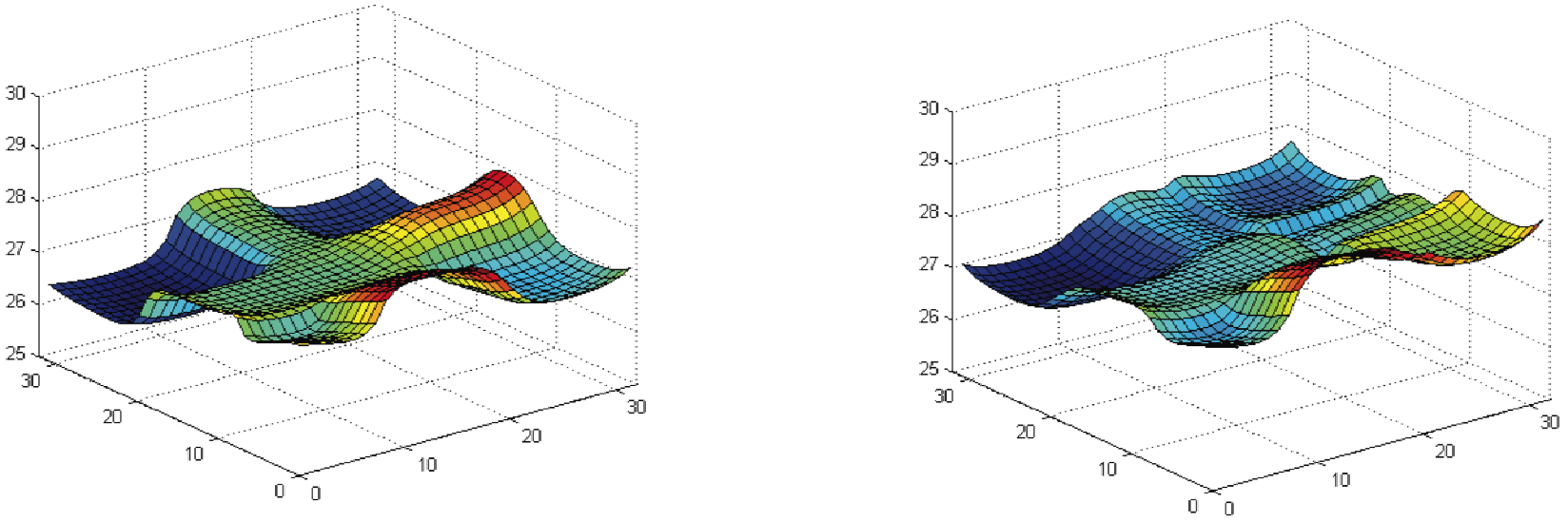
Work done so far

- Khuller, Li, Saha “Machine activation problems” SODA 2010.
- Shi, Srivastava “Thermal and Power aware task scheduling” Green Computing 2010.
- Deshpande, Khuller, Srivastava “A Case for Spatially-aware Optimization in Data Centers” manuscript.
- Deshpande, Khuller, Kumar, Saha “Data layout optimization for energy efficient query processing” in progress.

Next Steps

- Since the algorithms give good performance on simulated data, the next step would be to develop robust code that would go into software for managing workload on data centers.
- Develop a better understanding of this new class of optimization problems.

Thermal Profiles in a Data Center

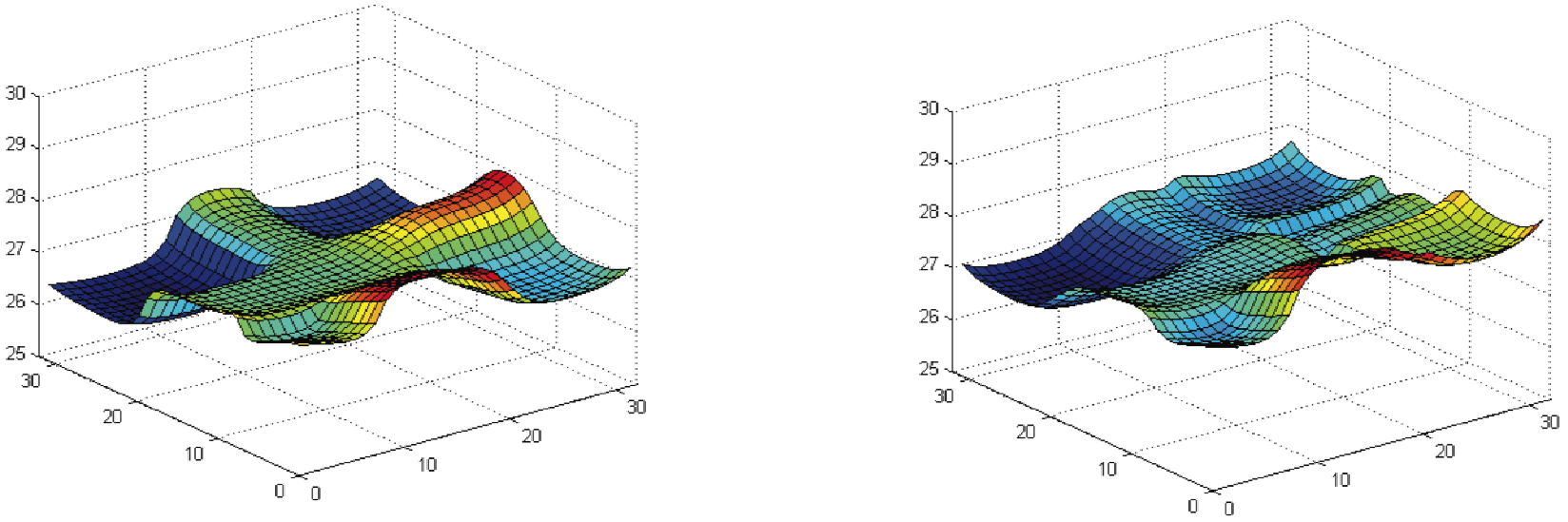


Simulated thermal profiles *5 minutes apart*

- Processing units located at the grid points
- Standard power dissipation model assumed

Large, rapidly changing temperature variations

Thermal Profiles in a Data Center

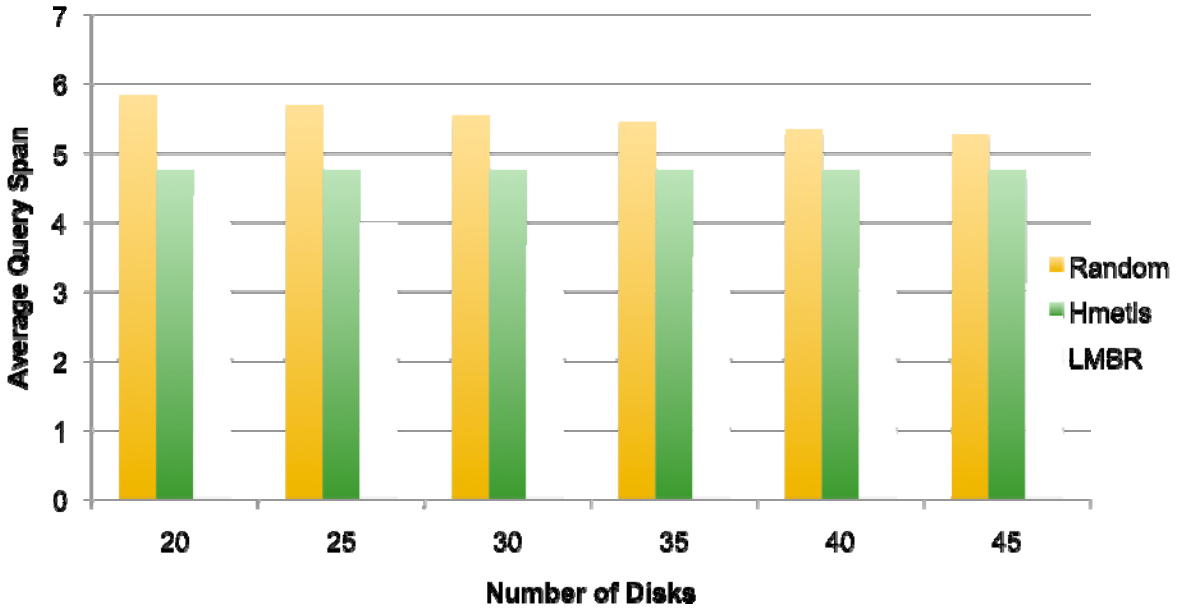


Simulated thermal profiles *5 minutes apart*

- Processing units located at the grid points
- Standard power dissipation model assumed

Large, rapidly changing temperature variations

Preliminary Results for Data Placement



HMETIS: State-of-the-art Hypergraph Partitioning Algorithm

LMBR: A greedy algorithm that does sophisticated local moves

Significant energy savings by doing workload-driven optimization