Autonomic Power & Performance Management of Large-scale Data Centers

Bithika Khargharia[1], **Salim Hariri[1]**, Ferenc Szidarovszky[1],
Manal Houri[2], Hesham El-Rewini[2],
Samee Khan[3], Ishfaq Ahmad[3] and Mazin S. Yousif [4]

[1]University of Arizona, [2]Southern Methodist University,
[3]University of Texas at Arlington, [4] Intel Corp.

NSF NGS Workshop, March 2007

Outline of This Talk

- Motivation
  - The Basic Problem
  - Objectives
  - Our Approach: *Hierarchical autonomic power and performance management*
- Research Approach and Results
  - Component-level Management (Optimization)
  - Preliminary Results
  - Cluster-level Management (Game Theory)
  - Preliminary Results
- Future Research Directions

Autonomic Middleware for Large Scale Scientific and Engineering Applications

- Physics Aware Programming Paradigm
- Autonomic Runtime Manager
  - Automatic detection of application execution phases and properties
  - Select the appropriate algorithm, solver at runtime
  - Select the appropriate resources and libraries
- Anomaly Based Management Framework
  - Performance
  - Fault
  - Security
  - Configuration

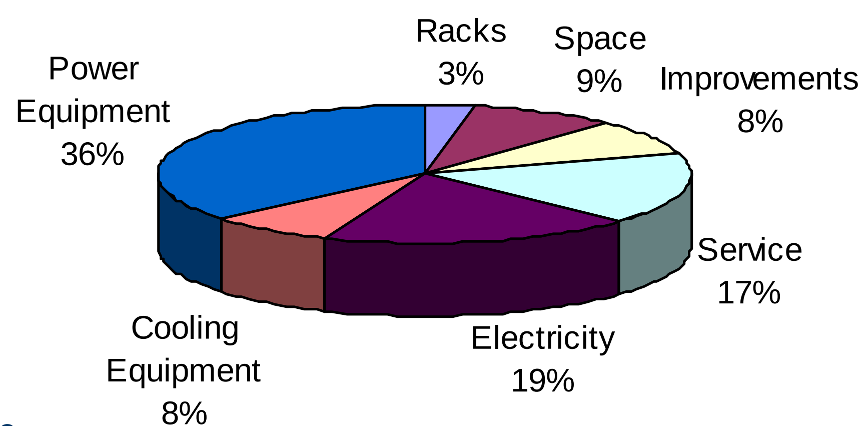
The Basic Problem

The Environmental Problem
- High CO_2 emission

The Financial Problem
- High TCO

The Technical Problem
- Dynamic & Heterogeneous
- Rack power distribution vS. space utilization

Current Density Levels 200 watts/sq. foot
Energy Needs 80 TWh/year
Energy Costs $8B/year @ 100$/MWh
CO2 release 50 tons/year

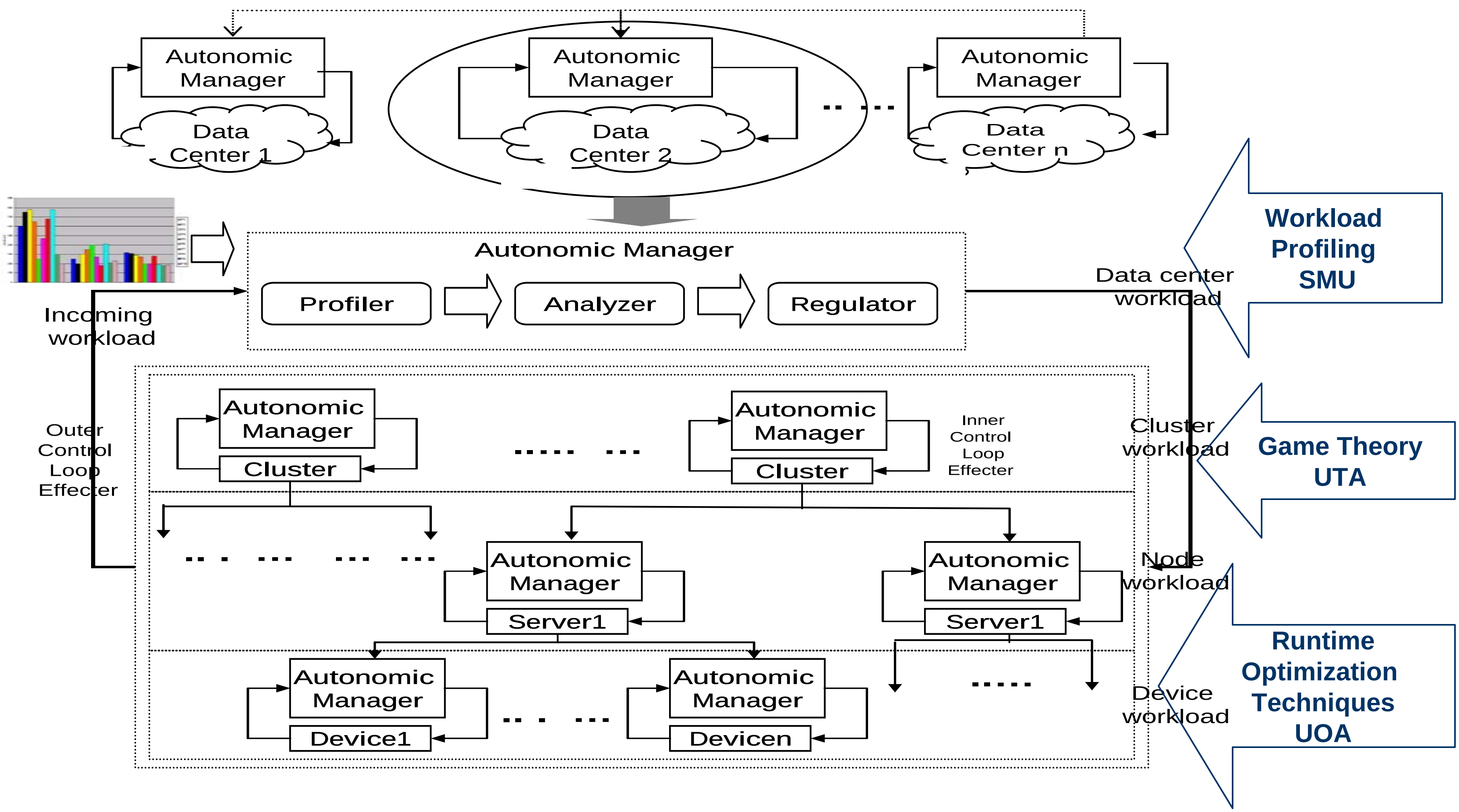


Objective of This Work

Methodology for autonomic power & performance management

- multi-layer management with bi-directional interactions among all layers (data center level, cluster level, component level)
- online monitoring & analysis
- adaptive learning & profiling strategy for data center workloads
- dynamically reconfigure CPU, Memory, I/O
- Use data mining and statistical techniques to implement real-time identified management strategies

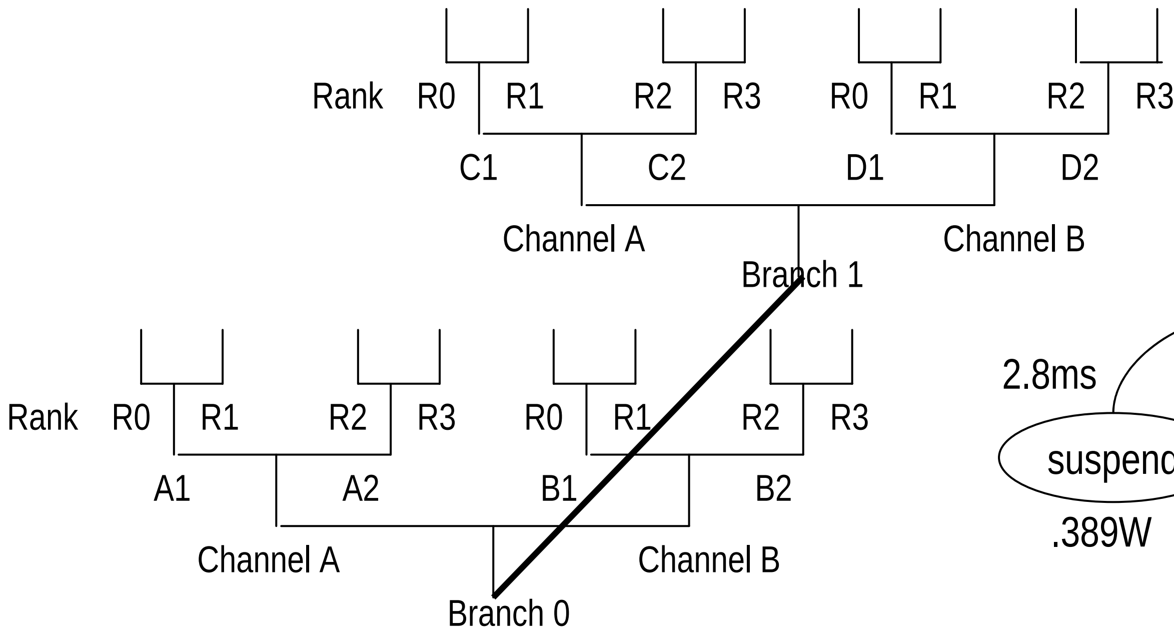
Hierarchical Autonomic Power & Performance Management



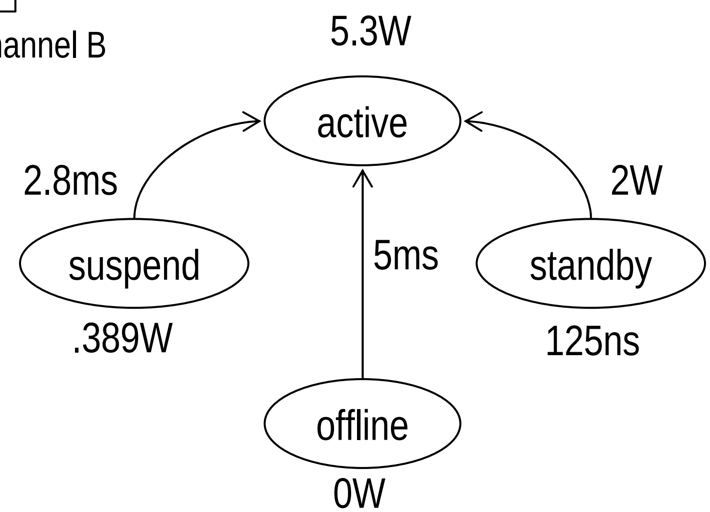
Hierarchical Autonomic Power & Performance Management

- *Top-level AM:* Distributes data center workload based on workload profiling and analysis
- *Cluster-level AM:* Uses game theory to devise a power & performance aware mapping of tasks to machines
- *Component-level AM:* Optimizes task working-set data placement on fully-interleaved memory modules.

Power and Performance Managed Memory System



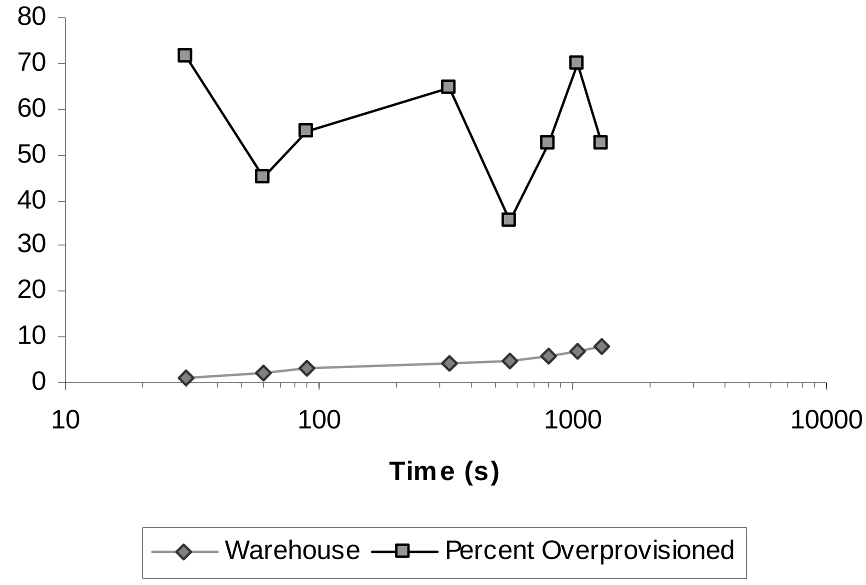
Memory Architecture



FBDIMM(DDR2) Power States

Power and Performance Managed Memory System

- Dynamically predict application memory requirements
- Determine the smallest memory configuration required by the application
- Transition the remaining modules to low-power states



SPECjbb 2005 heap over provisioning

Problem Formulation

Maximize *performance/watt*

$$ppw_i = \frac{1}{\tau_{a_k} * e_i}$$

$$e_i = \sum_{k=0}^{N} (c_{jk} * t_{trans_{jk}} + p * n_k * t_{obs}) * x_{jk}$$

such that

$$1. n_k * size / Rank >= N_{opt} * pageSize$$

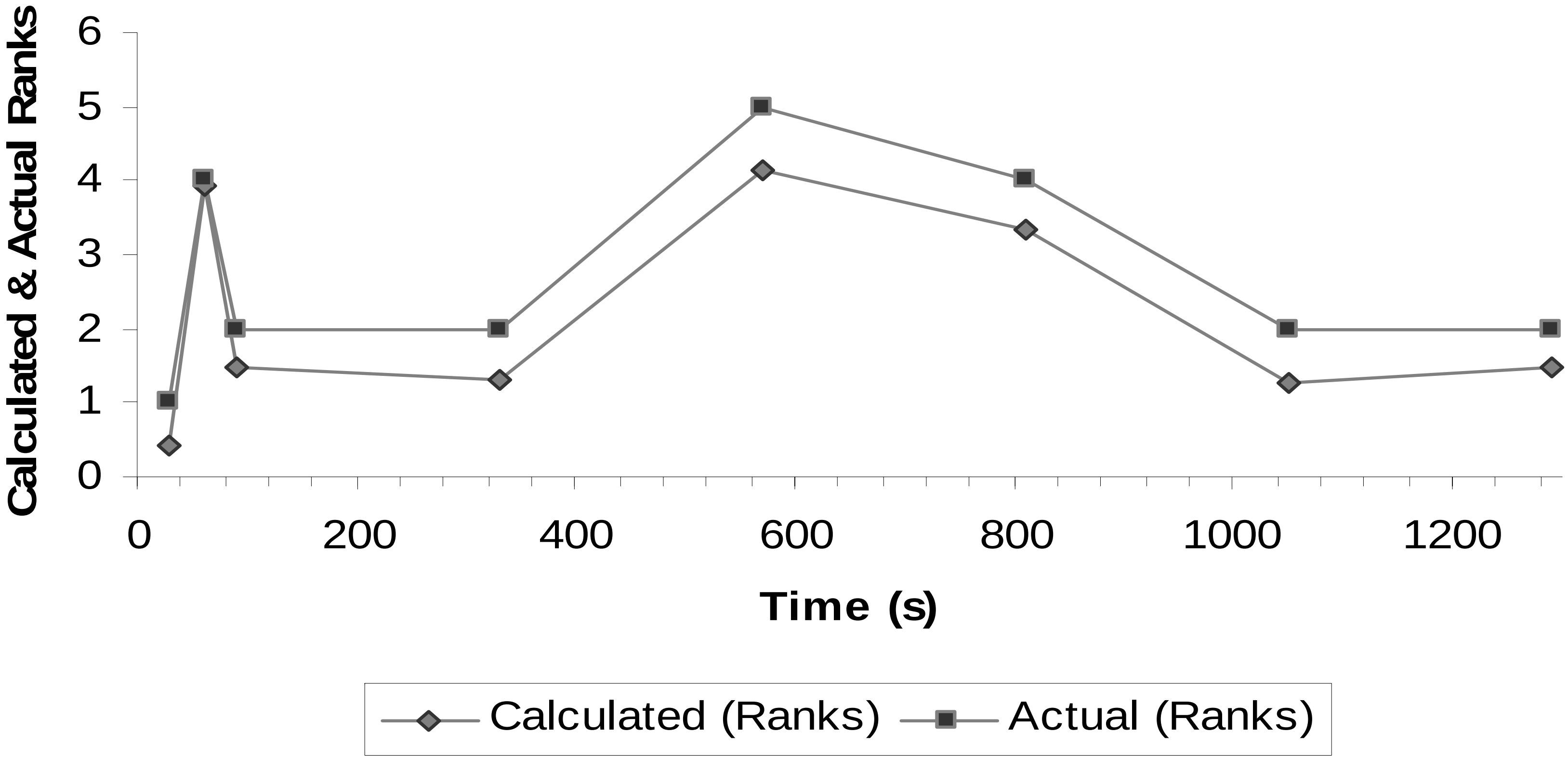
$$2. Max(\bigvee_{ch:} \rho_k) <= \rho_{th}$$

$$3. Min(\vee_r \tau_{a\,k}) >= \tau_{a_{th}}$$

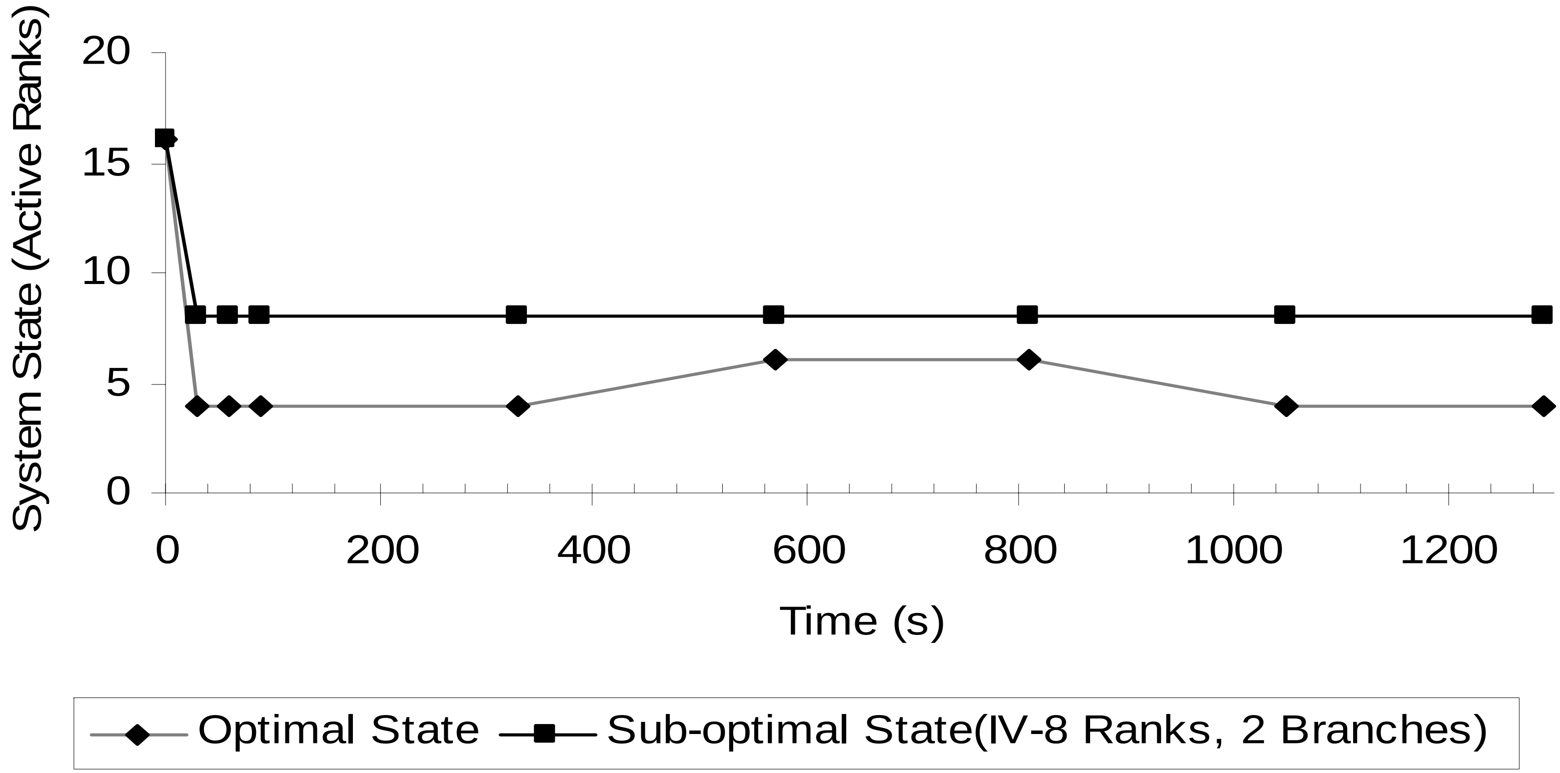
$$4. \sum_{k=0}^{n} x_{jk} = 1$$

$$5. \vee\, x_{jk} = 0 | 1$$

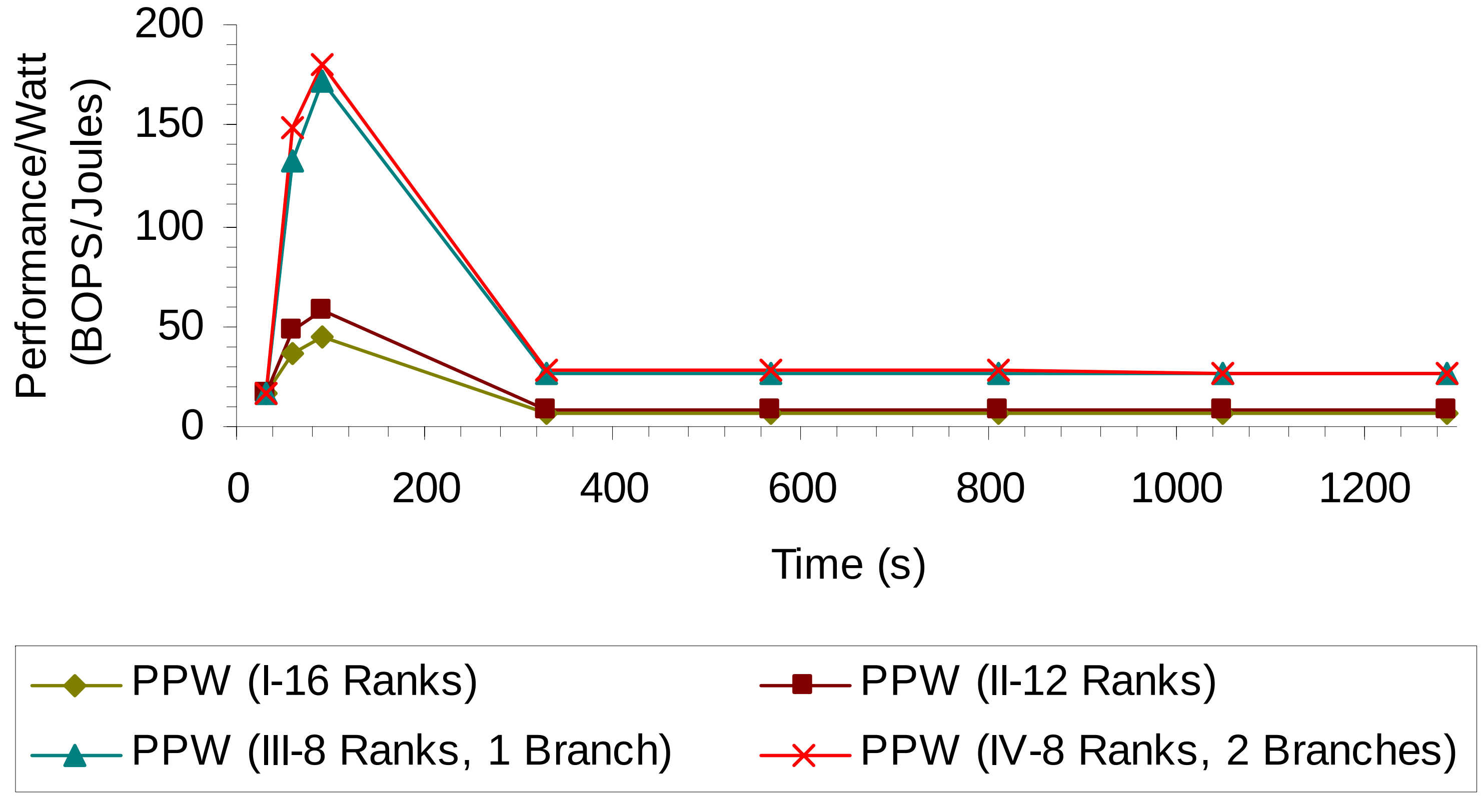
SPECjbb 2005 Working Set Pages



Performance/Watt Analysis



Performance/Watt Analysis



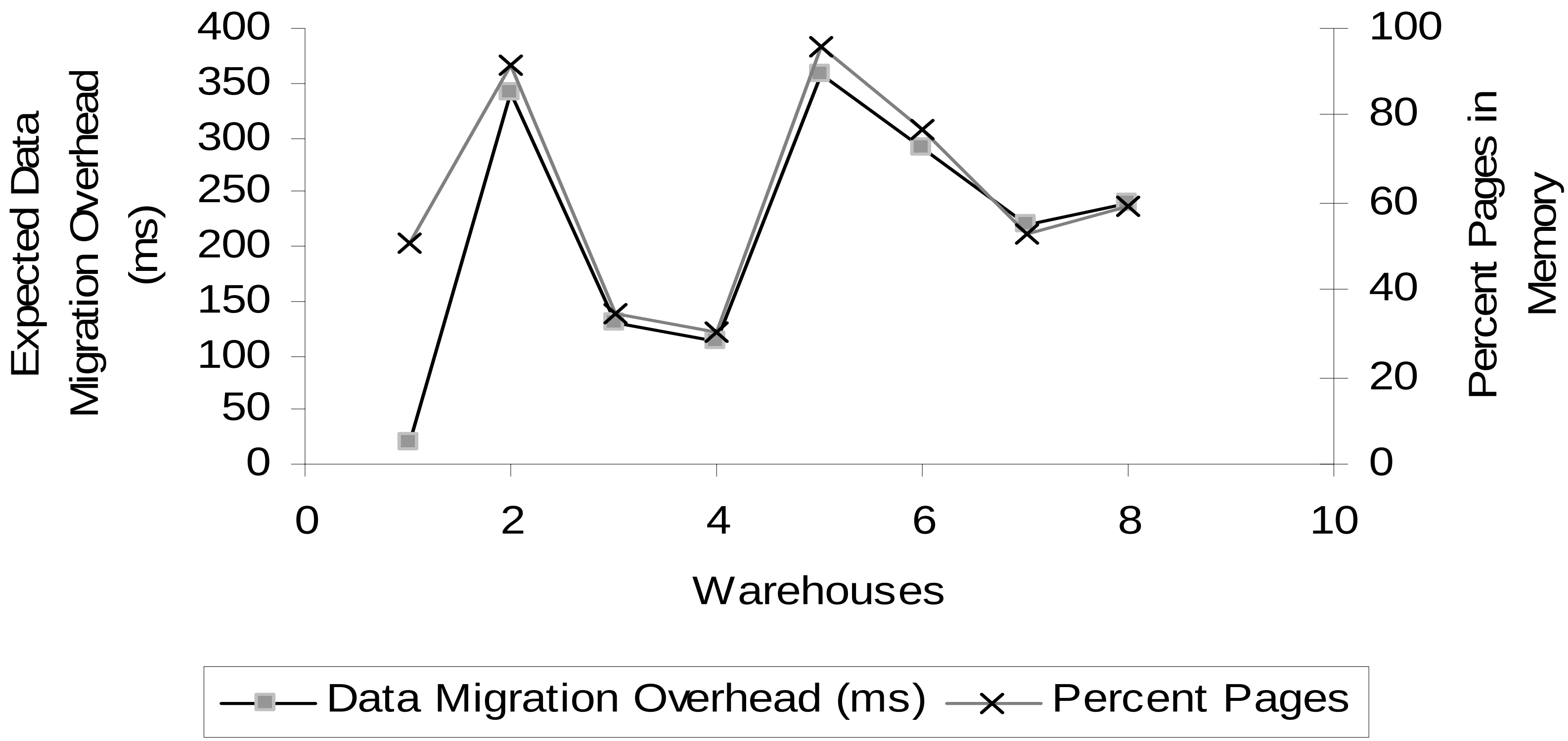
Maximum performance/watt improvement of 88.48%

Energy saving of about 48.8 % (26.7 kJ)

Migration Strategies

Performance drop for random migration measured at 5.72% for SPECjbb2005

Migration Overhead



Transition overhead of about 18.6 ms (16 ranks to 8 ranks) at warehouse 1

Power and Performance Management Cluster Level

- Co-operative Game Theory
  - Centrally allocate tasks to machines
  - Co-operate to achieve system-wide power and performance improvement
- Non-cooperative Game Theory
  - Distributed mapping of tasks to machines
  - Machine competes with each other to improve its own power and performance.

Problem Formulation

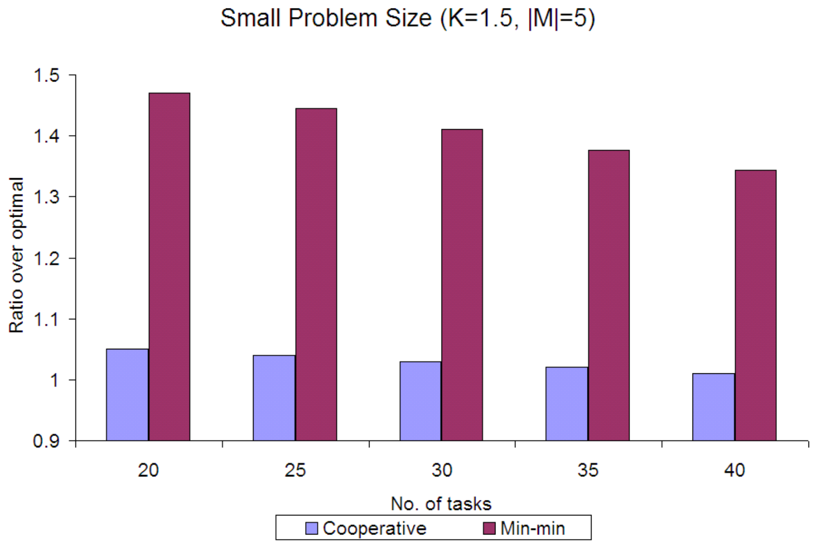
$$\min \sum_{i=1}^{n} \sum_{j=i}^{m} p_{ij} x_{ij} \text{ such that } \min \max_{1 \le j \le m} \sum_{i=1}^{n} t_{ij} x_{ij} \text{ subject to}$$

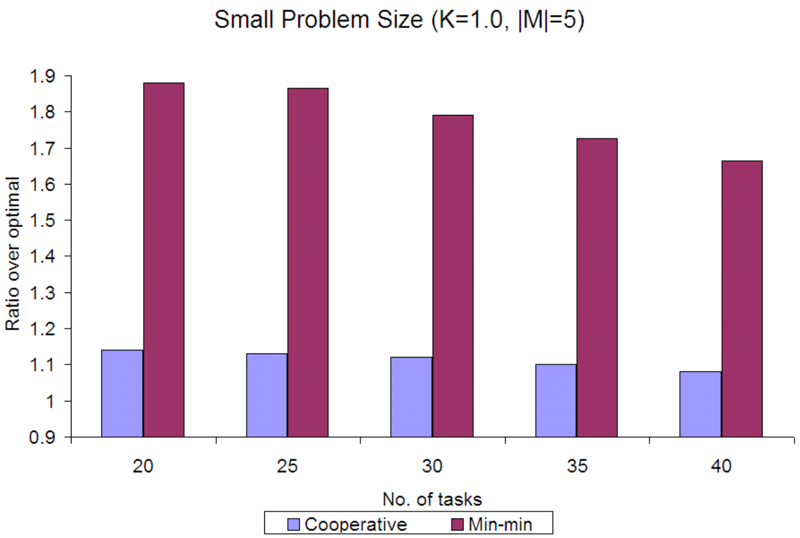
1. $x_{ij} \in \{0,1\}, i = 1, 2, ..., n; j = 1, 2, ...m.$
2. if $t_i \to m_j, \forall i, \forall j,$ such that $A(t_i) = A(m_j),$ then $x_{ij} = 1$
3. $t_{ij} x_{ij} \le d_i, \forall i, \forall j, x_{ij} = 1$
4. $(t_{ij} x_{ij} \le d_i) \in \{0,1\}$
5. $\prod_{i=1}^{n} (t_{ij} x_{ij} \le d_i) = 1, \forall i, \forall j, x_{ij} = 1$

Experiments

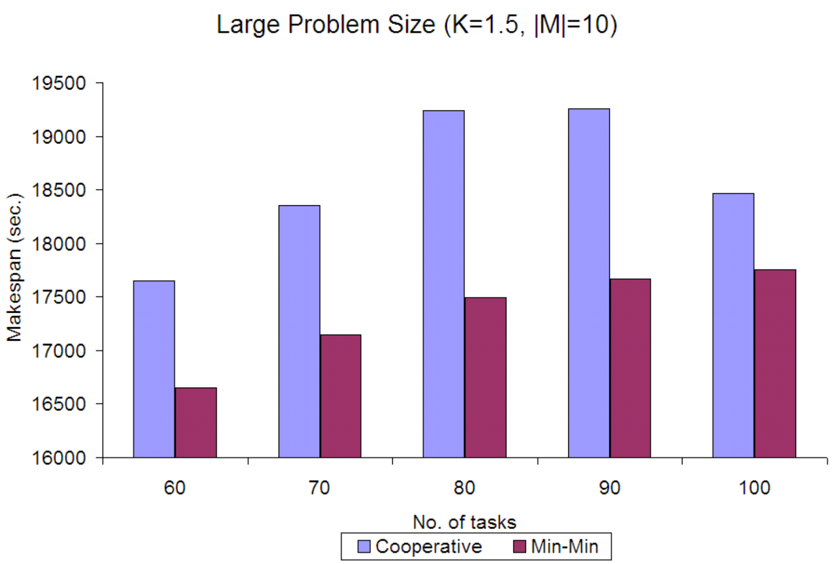
- Comparisons against LINDO and min-min heuristic.
- System heterogeneity is captured using the Gamma method.
- di is calculated as $K \times wi \times X$, where K is a pre-specified positive value for adjusting the relative deadlines of tasks.

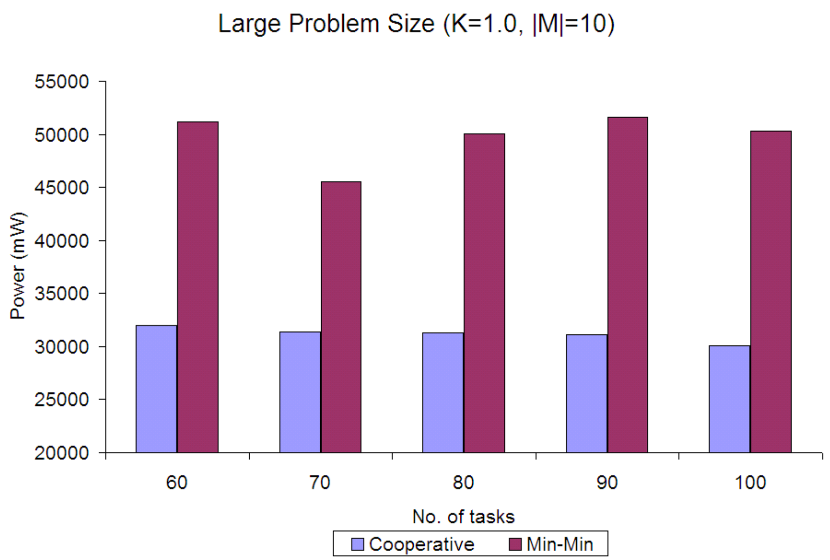
Comparison against Optimal



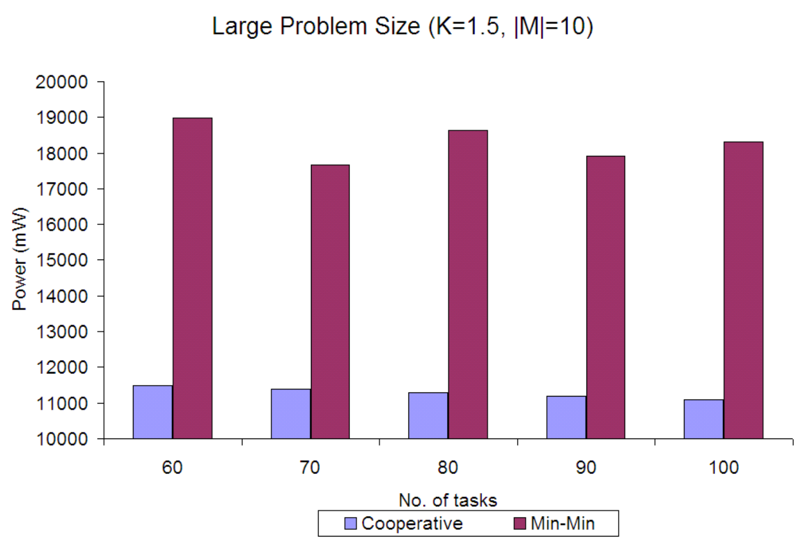


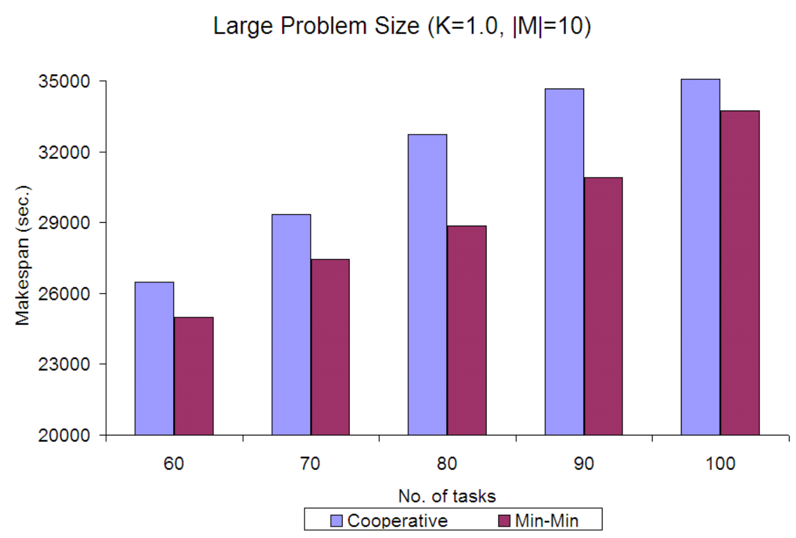
Power Savings





Makespan Comparision





Future Research Directions

- Define accurate workload profiling parameters that can be used at multiple level of the hierarchy
- Dynamic non-co-operative game theory
  - Dynamic behaviors
  - Feedback from all layers
- Autonomic PP management of I/O, Processor, memory
  - Autonomic Interleaved Memory System
- Data mining and statistical techniques to implement AM control & management