IBM Research

A multi-agent view of data center energy management

Jeff Kephart (kephart@us.ibm.com)
IBM Thomas J Watson Research Center
Hawthorne, NY

Multi-agent systems and autonomic data centers

- **I envision data centers of the future as an ecosystem of interacting semi-autonomous entities – an *autonomic*, *multi-agent* system**
  - Autonomic elements ~ Agents
  - Autonomic systems ~ Multi-agent systems
- Agents will
  - represent, or be embedded in, different products from different vendors
  - reside at (and control) many levels of the management stack
  - collectively manage the data center to specified objectives and constraints relating to Power, Performance, Availability, Reliability, Security …
- This vision is a natural extrapolation of present-day facts and trends
  - Industry and academia are developing a multitude of control knobs and automated techniques to save energy
  - These will be incorporated into a multitude of management products from different vendors
  - They will operate simultaneously within and across multiple levels of the stack
  - These products must cooperative effectively

Data Center as a Multi-Agent System

I want to increase the fan speed on that CRAC unit

I want to turn that server off to save power

I want to perform valuable computational services

Power Manager

I want to process workload on that server

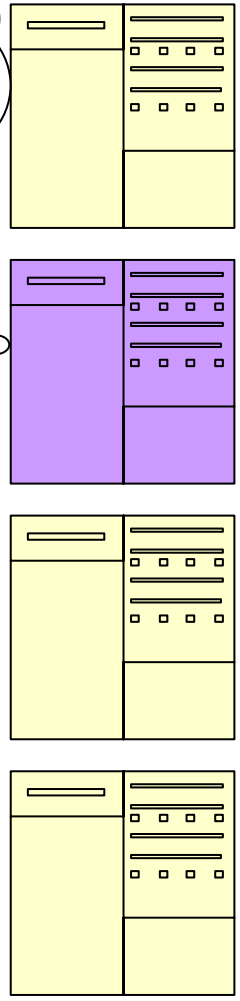
Performance Mgr

I want to use that server to achieve 3-fold redundancy

Availability Manager

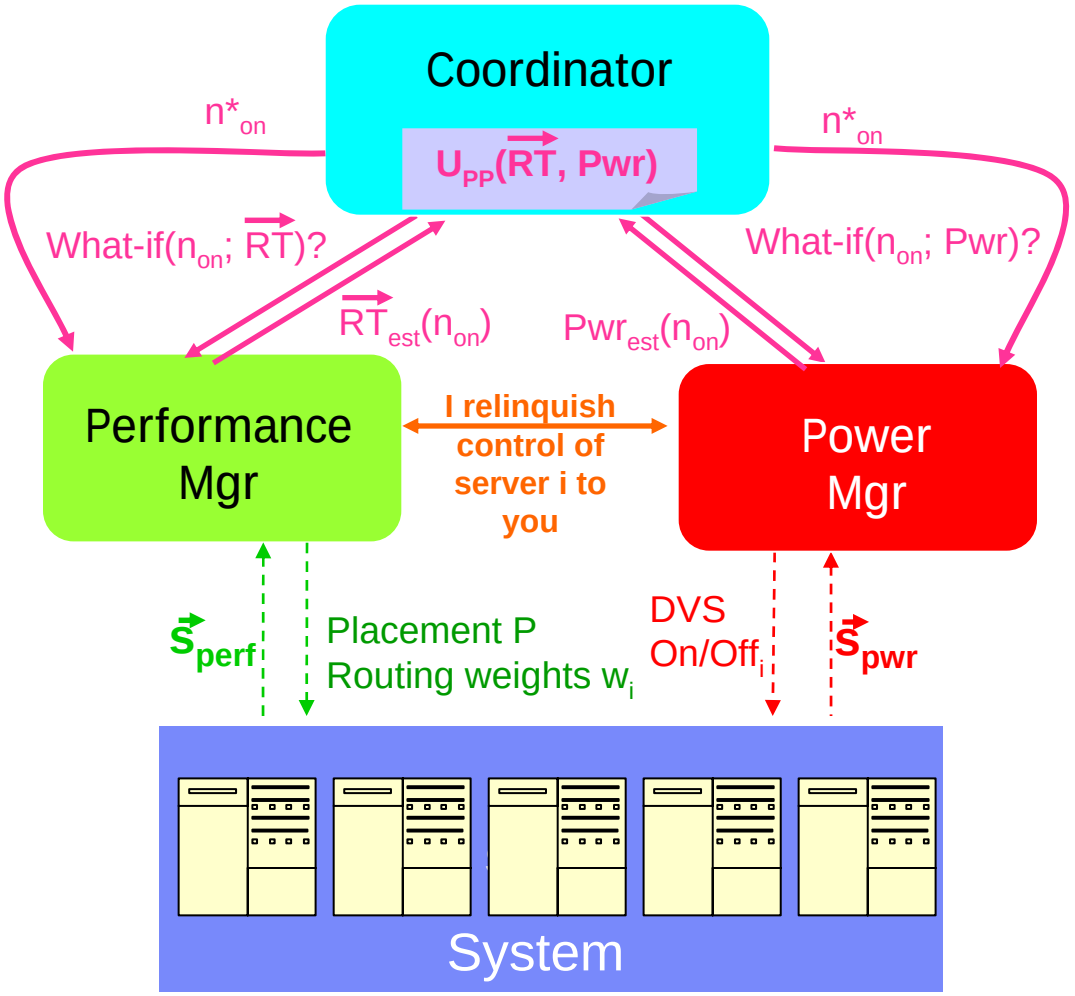
I want to leave that server in its present power state to reduce thermal stress

Reliability Manager



Example: Interaction between power and performance agents

- How might semi-autonomous power and performance agents interact?
  - Mediated through coordinator
    Direct bi/multi-lateral interactions
- Scenario (with mediation)
  - Performance manager observes subset $\mathbf{s}_{perf}$ of system state, and controls application placement and load balancing weights
  - Power manager observes subset of $\mathbf{s}_{pwr}$ of system state, and controls on/off state of servers
  - Coordinator understands overall power-performance tradeoffs as expressed in a joint utility function, and queries performance and power agents for likely impact when n servers are turned on, finding optimal number n*



Kephart, Chan, Das, Levine, Tesauro, Rawson, Lefurgy. *Coordinating Multiple Autonomic Managers to Achieve Specified Power-Performance Tradeoffs*. **ICAC 2007.**
(Emergent phenomena can occur when autonomic managers don't communicate effectively.)

Research Challenges

- Architectural questions
  - What is the scope/boundary of the agents? Where situated?
    - By management discipline? By layer of stack? …
  - What and how do they communicate?
  - Is multi-agent approach viable, using negotiating agents and mediators to manage performance, power, …?
  - Are markets and auctions effective coordination mechanisms when there are numerous agents and "goods"?
    - What are goods in this case (e.g. 1 core in multi-core system?)
    - Hierarchical markets extending across multiple data centers
    - Consequences of coupling data center markets to global economy?
- Algorithmic (and other) challenges
  - Learning what-if models on the fly despite noise, complexity
  - Avoiding undesirable emergent phenomena
  - Eliciting preferences (tradeoffs among power, performance, ..)
- Beyond IT
  - Agents will represent PDUs, CRACs, chillers, etc., vastly increasing the size and variety of the MAS
  - Coordinate workload placement, load balancing, etc. with facilities management
  - E.g. co-manage cooling and workload migration

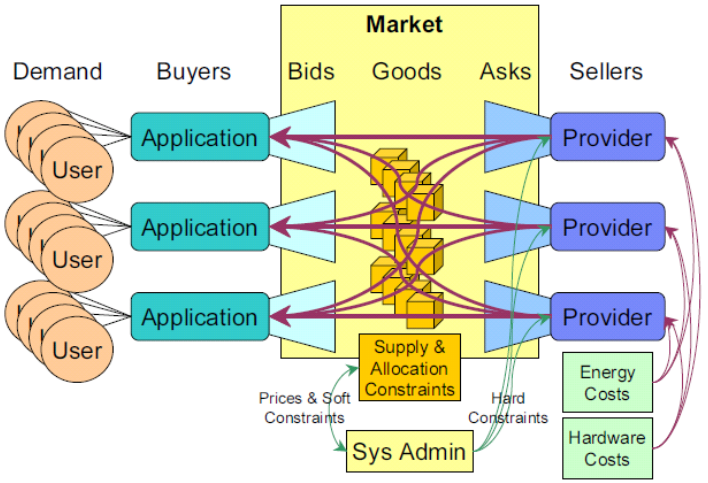


Figure 1: The Data Center Market Model

Lubin, Kephart, Das and Parkes. *Expressive Power-Based Resource Allocation for Data Centers*. **IJCAI 2009.** (Exploring market-based resource allocation for data centers.)

One more challenge

- ICAC needs a renewed emphasis on the holistic vision of autonomic computing

- Technologies that support inter-element interactions, such as service-level agreements, negotiation protocols and algorithms, and conversation support
- System-level technologies or services that entail interactions among two or more components of self-managing systems…
- Autonomic computing systems or prototype systems that exhibit self-configuration, self-optimization, self-healing, and/or self-protection
- Human interaction with autonomic systems

ICAC 2004 Call for Papers (excerpt)

Backup

Agents and AC: Definitions

- Autonomic computing definition
  - "Computing systems that manage themselves in accordance with high-level objectives from humans." Kephart and Chess, *A Vision of Autonomic Computing*, **IEEE Computer**, Jan. 2003
- Software agent definition
  - "An encapsulated computer system, situated in some environment, and capable of flexible, autonomous action in that environment in order to meet its design objectives." Jennings, Sycara and Wooldridge, *A Road Map of Agent Research and Development,* **JAAMAS 1998**
  - *Multi-agent systems*: collections of agents that interact with one another to achieve individual and/or system goals