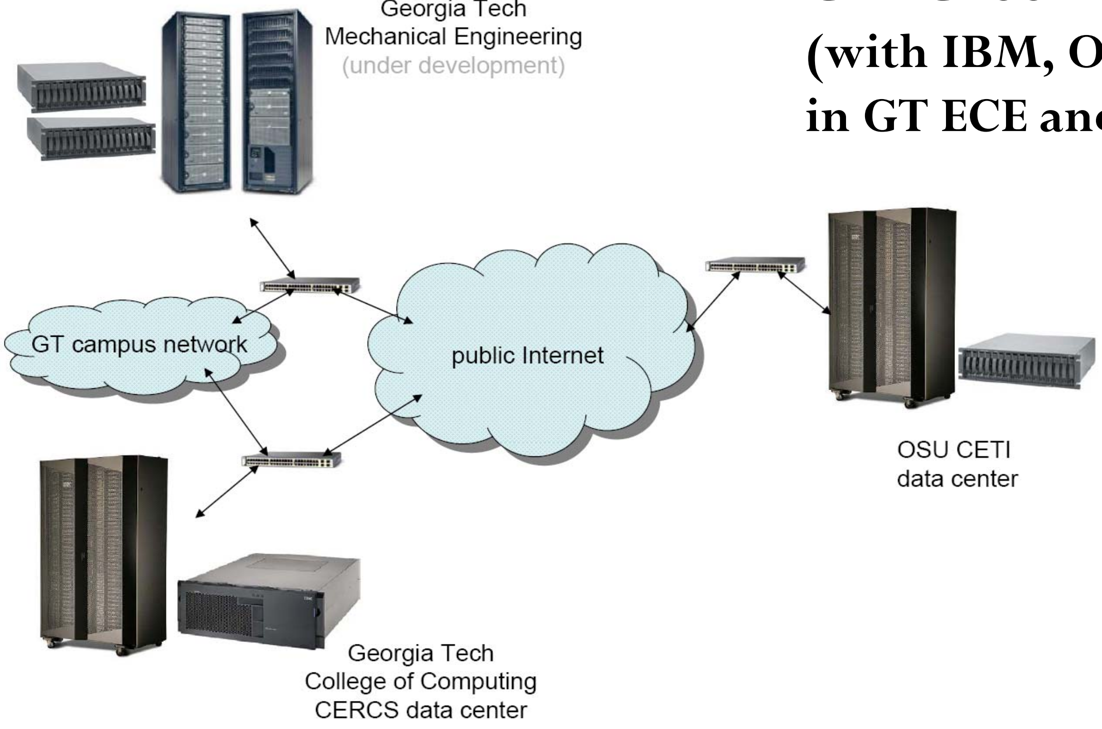
Power Management for Utility Clouds

Karsten Schwan (schwan@cc.gatech.edu)

GreenIT Cloud@GT - Context

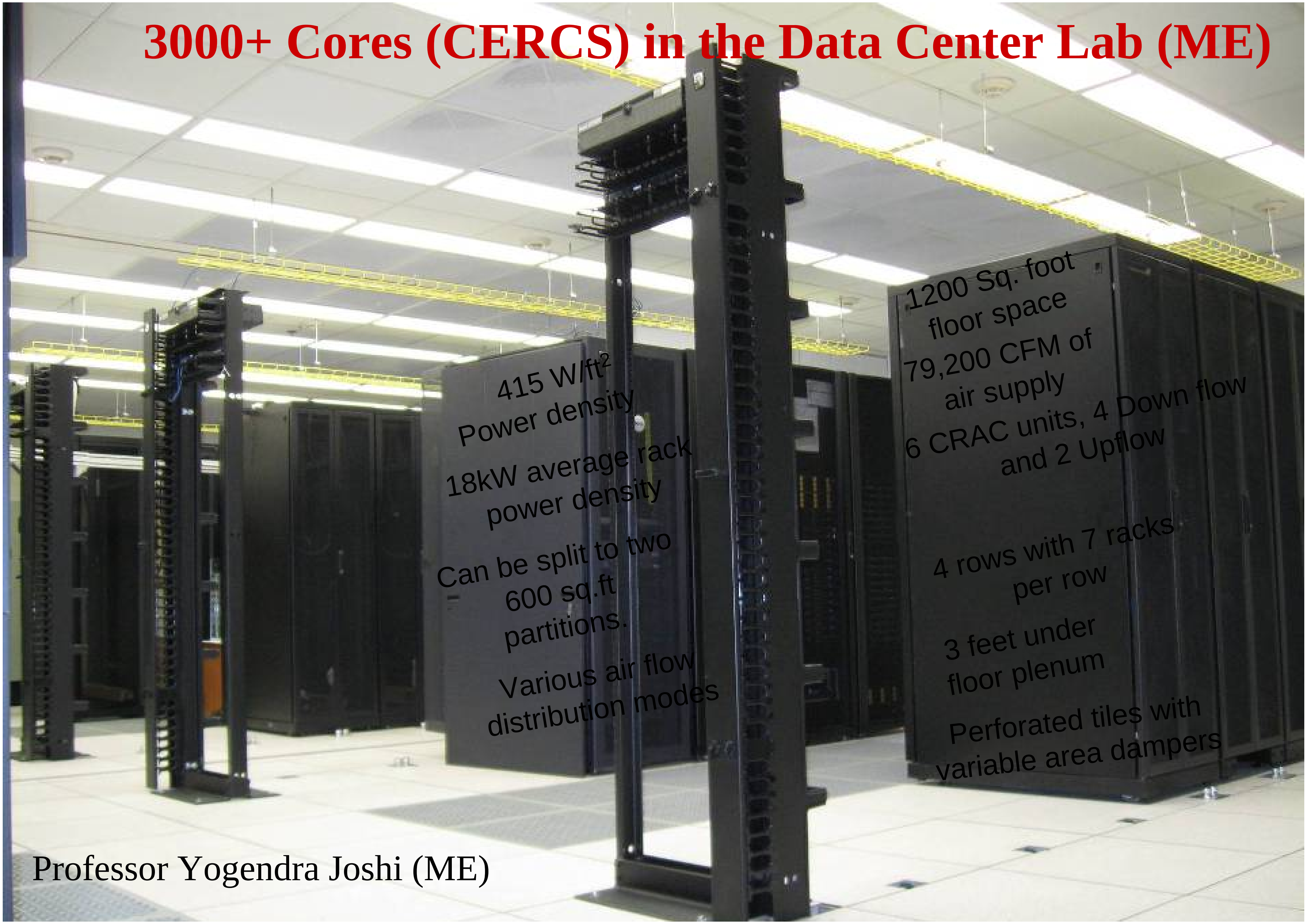


GT `GreenIT' Private Cloud

(with IBM, OSISoft, HP, VMWare, and PIs in GT ECE and ME; CSE Ohio State)

→ **OpenCirrus (HP, Intel, Yahoo)**

3000+ Cores (CERCS) in the Data Center Lab (ME)



415 W/ft^2
Power density

18kW average rack
power density

Can be split to two
600 sq.ft
partitions.

Various air flow
distribution modes

1200 Sq. foot
floor space

79,200 CFM of
air supply

6 CRAC units, 4 Down flow
and 2 Upflow

4 rows with 7 racks
per row

3 feet under
floor plenum

Perforated tiles with
variable area dampers

Professor Yogendra Joshi (ME)

International Workshop on ***Thermal Design and Management in Electronics,*** January 8th 2010, Mumbai, India

Managing Large-Scale Utility Clouds

Basics:

- Applications have meaningful requirements:
  - Critical enterprise applications have SLAs
  - Health applications have security constraints/roles in addition to performance/reliability requirements
  - Entertainment/sensor apps. have real-time constraints

  => Need for active management

- There are management silos:
  - HP's iLO, IBM's Director, IPMI and platform level management, IBM's Tivoli enterprise-level management, VMWare's Virtual Center

- *Silos and scale make integrated management infeasible*

=> I. Coordinated management – vManage architecture:

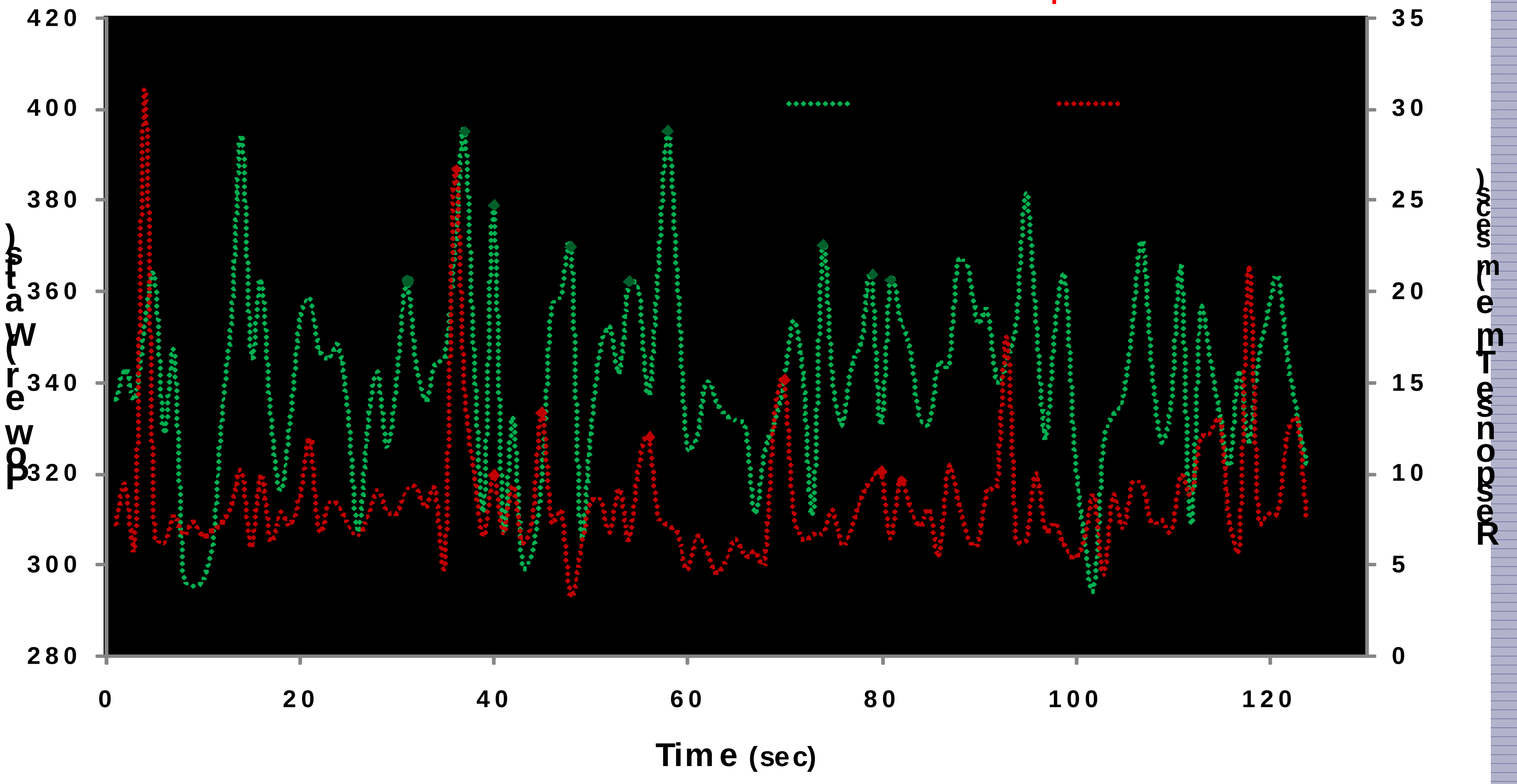
management at and across multiple levels of abstraction, subsystems, and machines (joint with HP Labs)

=> II. Need for new and scalable methods for mon. and mgt. – `Monalytics': combined monitoring and analysis

vManage: Problems with Silos

(using a mix of server applications)

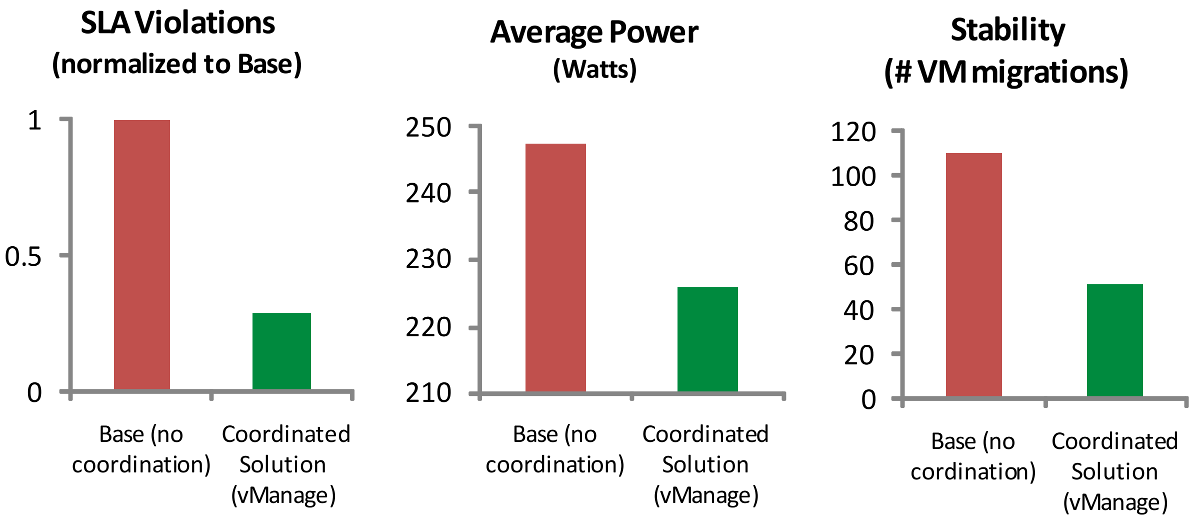
Oscillations among
SLA and power violations



Significant violations & instability due to oscillatory behavior

vManage: Coordination is Useful

- 28 VMs run over 20 hours on a 13-node testbed
  - 10 Nutch instances, 3 RUBiS instances, 6 static webserver instances



- Significantly better QoS (71%)
- Improved power savings (10%)
- Better stability (54%)

vManage: Loosely Coupled Platform and Virtualization Management in Data Centers [Slides]
Sanjay Kumar; Vanish Talwar; Vibhore Kumar; Partha Ranganathan; Karsten Schwan; 2009 International Conference on Autonomic Computing (ICAC)

vManage => Monalytics

Motivation: Monitoring to manage large-scale systems

Scale: #components and operation at multiple length and time scales:

e.g., length: datacenter health vs. subsystem state

=> scope in space (diverse data structures: agg. trees, DHTs, …)

e.g., time: high rate web requests, low rate VM migration => scope in time (window sizes, …)

Dynamics:

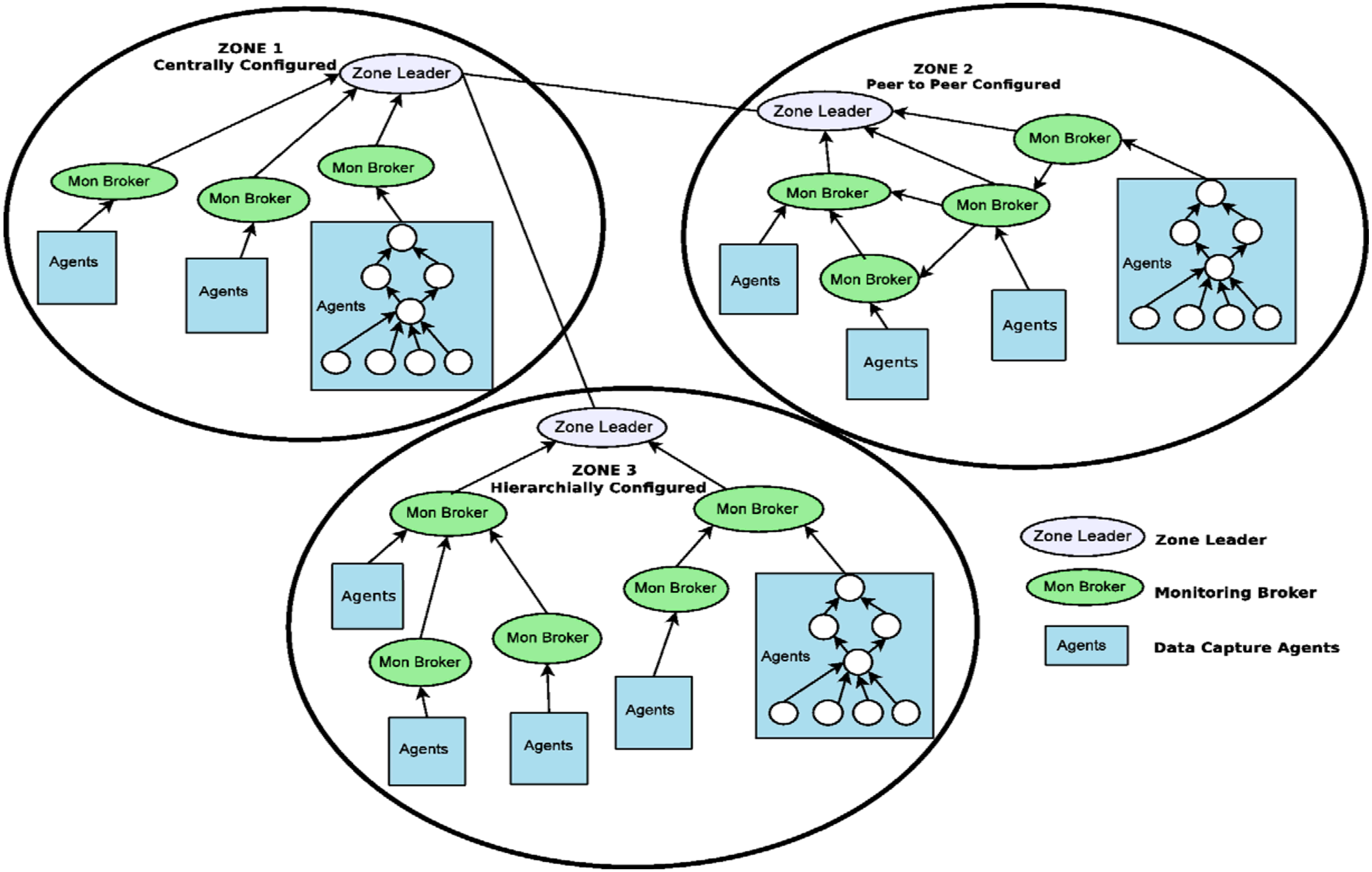
dyn. arrivals/departures, dyn. queries

Capturing and understanding data:

dyn. analysis

=> Monalytics – combined monitoring/analysis

Monalytics Architecture – Topologies



Dynamic Computational Communication Graphs

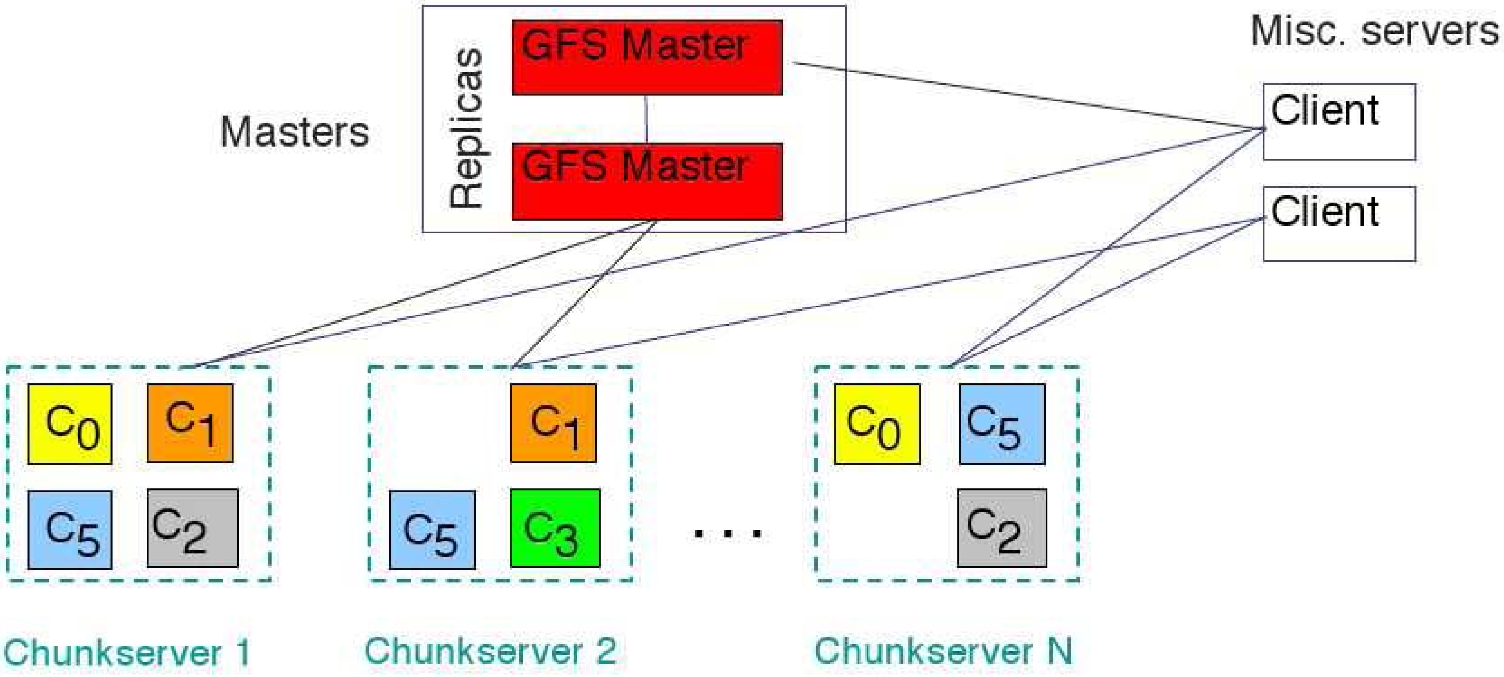
Monalytics – Hadoop Use Cases

- Ongoing work:
  - Hadoop – using HP's OpenCirrus cluster
  - PreData – using 'staging area' on Jaguar petascale machine

- Results to date (not using monalytics software):
  - powering off select machines in datacenter Hadoop computations – outlined next – joint work of our student Hrishikesh Amur with CMU's Storage Systems group – Greg Ganger

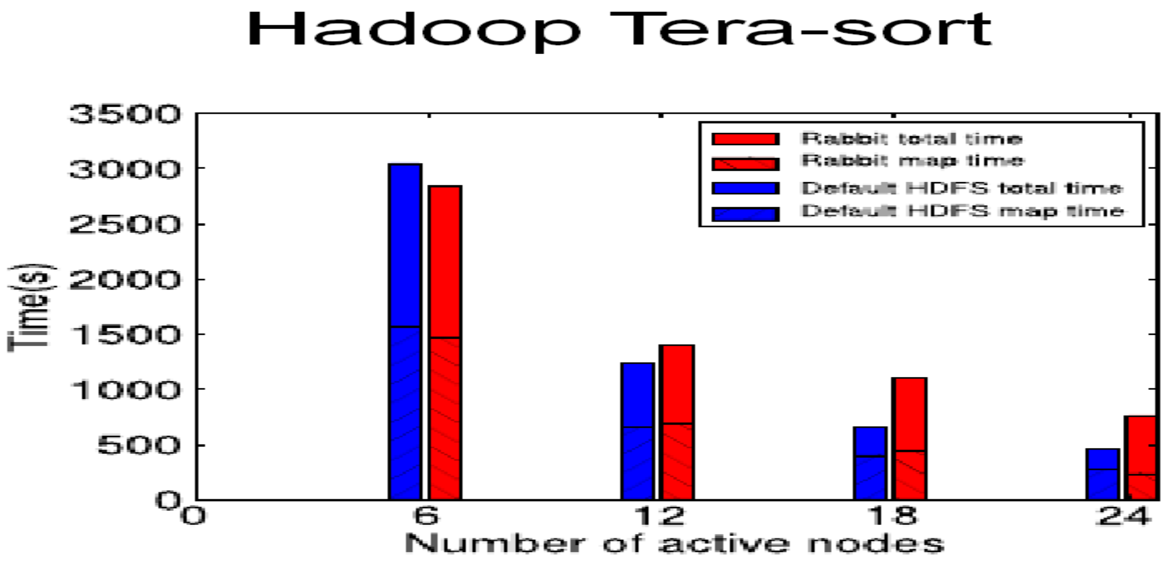
Problem Description

Turning Off Nodes Breaks Conventional DFS



Solution – Equal Work Policy

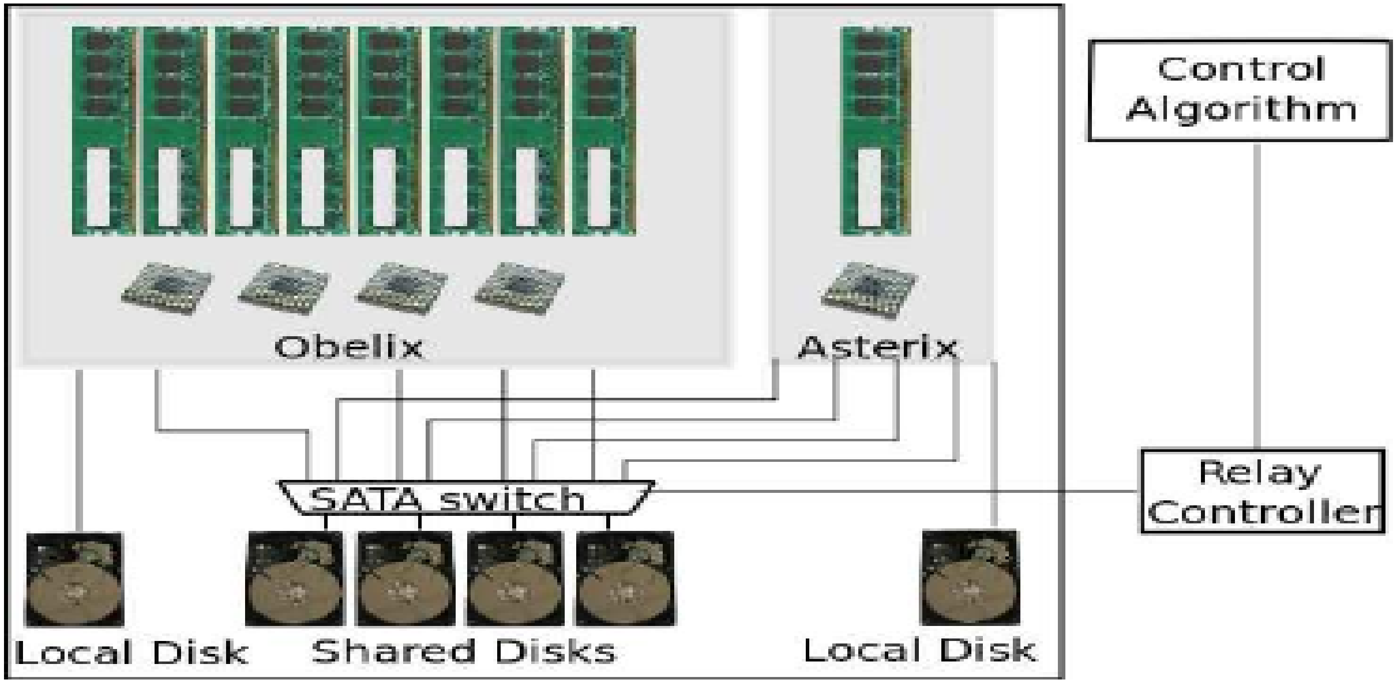
- Change Chunk -> Node allocation such that:
  - Low minimum power-performance setting
  - High maximum power-performance setting
  - Fine-grained scaling without data movement
- Outcome:
  - power-proportionality with almost identical read performance (write perf. needs addt'l. work)



Robust and Flexible Power-Proportional Storage: Hrishikesh Amur, James Cipar , Varun Gupta , Michael Kozuch , Gregory Ganger, Karsten Schwan, SOCC 2010.

Next Steps

- Issue: DFS complexity – data layout, writes, …
- Solution: emulate future heterogeneous multicore node with `small’ and ‘large’ cores: Asterix and Obelix



Initial Results

- Using Atom platform as Asterix, quad core IA platform as Obelix
- HDFS was configured on a single datanode and throughput was measured.

- *Asterix-II gives comparable performance to Obelix for a third of the power*.

Many Remaining Issues

- Power-proportional performance on datacenter machines:
  - Hadoop and DFS constitute one interesting, but specific, set of use cases
- vManage -> Monalytics address more general usage for explicit power/performance tradeoffs

- Monalytics -> CoolIT for combined facility/IT management:
- Effective management and coordinating across silos:
    - APIs and standards
    - coordination methods vs. local/global optimization: operating across different levels of scale and detail
    - constraints on management
    - optimization criteria/chargeback models/metrics

  - Scaling to Exascale:
    - dynamics: needed: scalable control, including:
      - automation in deployment and use (e.g., monalytics QoS)
    - ease of use: higher level abstractions, including
      - linking abnormal behavior detection to problem diagnosis and prevention
    - wide-area: distributed utility clouds

CoolIT in the The Energy Stack

CoolIT

General Theme: Coordinated energy management across all levels

Datacenter and Rack :

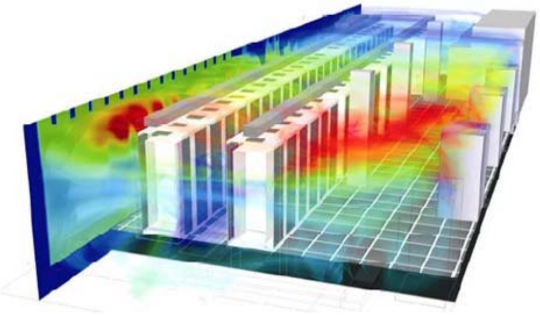
- Cooling, Management, power delivery(OIT, ME, CS)

•Thermal & airflow analysis, OS scheduling, cooling, (ME, CS)

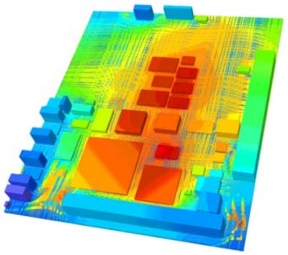
Board Level: Virtual power, OS scheduling, (ME, CS, ECE)

Chip/Package: power delivery, & management, thermal modeling, architectural support (ECE, ME, CS)

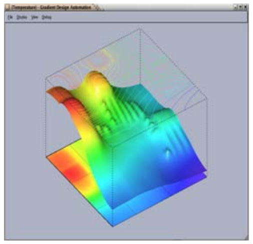
Circuit Level: power delivery, DVFS, clock gating, power states (ECE)



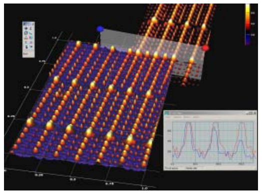
http://images.google.com/



http://www.scdsource.com/



http://chipdesignmag.com/



http://www.sciencegl.com/