Green Energy Prediction for Distributed Datacenters

Prof. Tajana Šimunić Rosing
Dept. of Computer Science

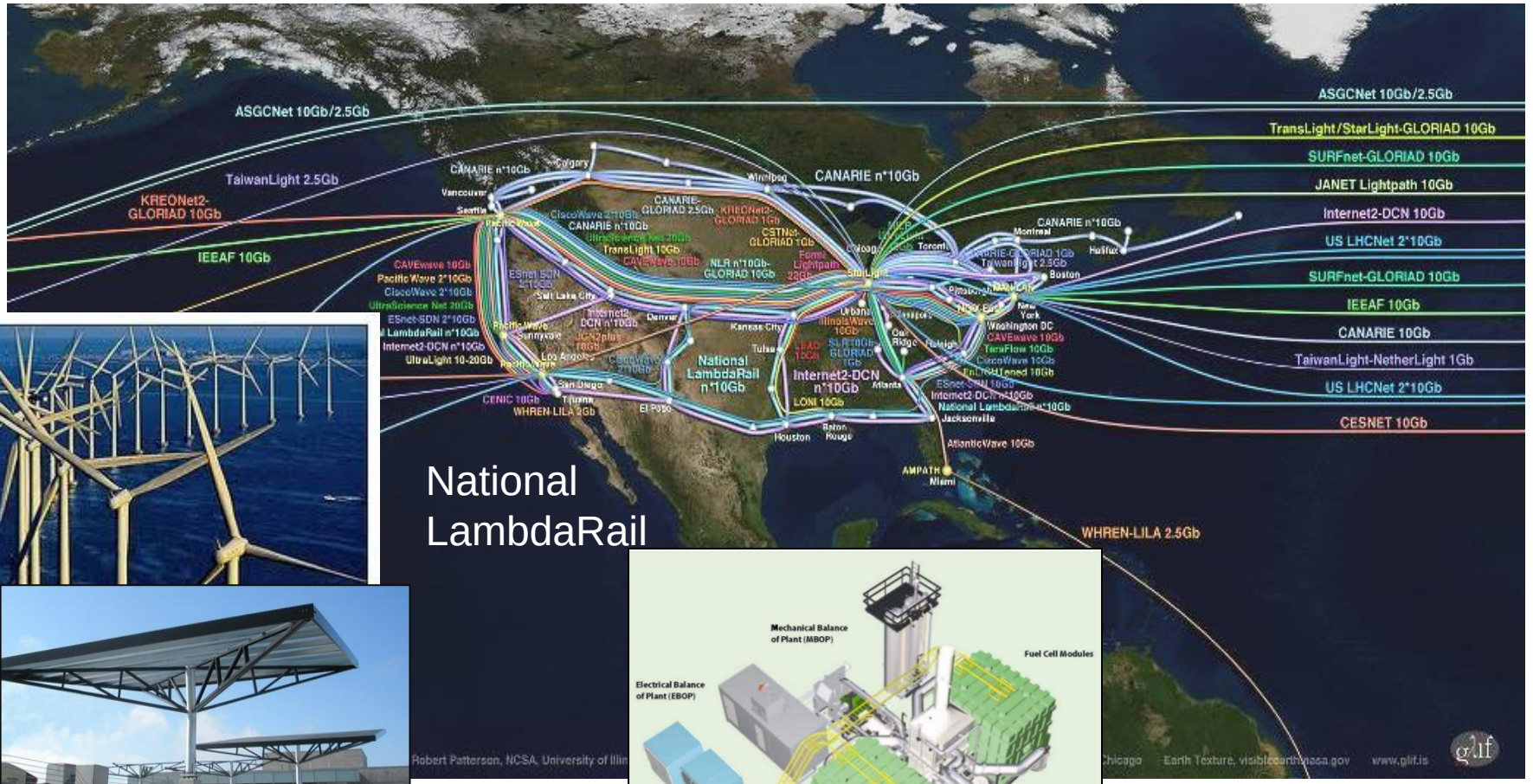
System Energy Efficiency Lab

seelab.ucsd.edu

Texas Instruments Inc.

Energy efficiency at global scale

NSF GreenLight & IRNC: TransLight/StarLight



National LambdaRail

Wind from NREL, 2MW Solar @UCSD

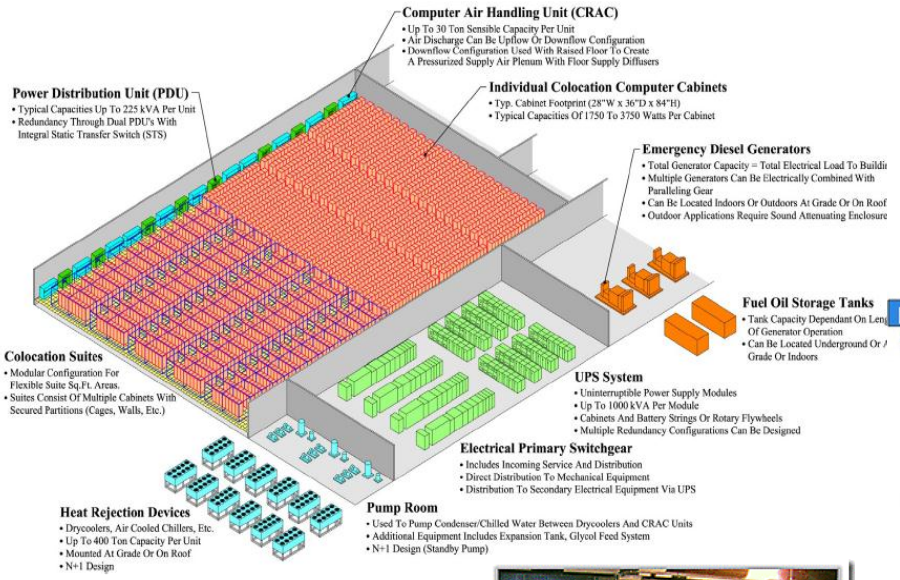
2.8MW Fuel Cell Power Plant , UCSD

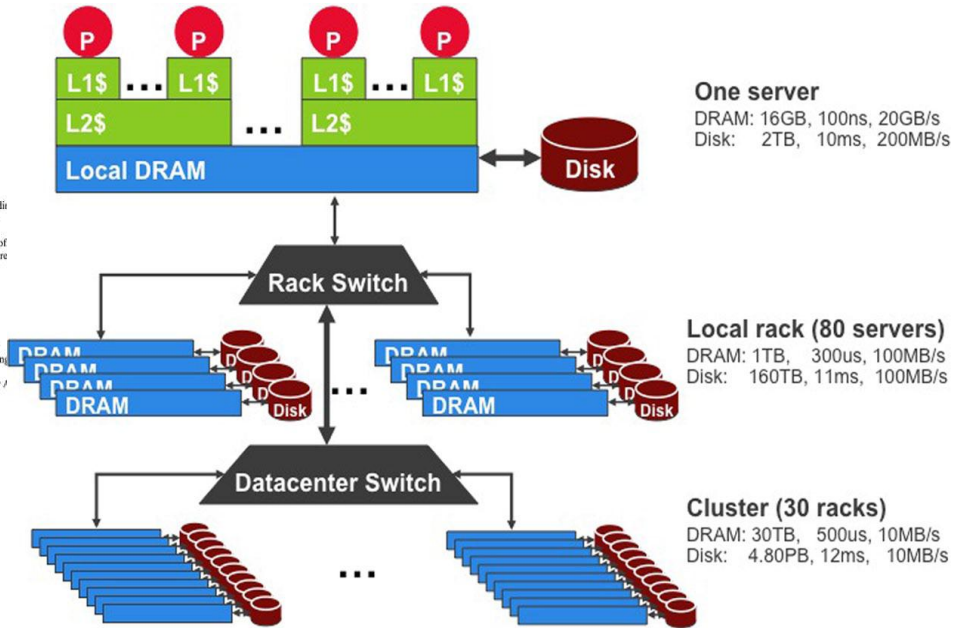
Key issues for distributed renewable-powered datacenters

- Green energy availability varies dramatically
  - Instantaneous use leads to significant energy efficiency losses
  - Prediction is needed
- Datacenter computing requires consistent performance
  - Infrastructure that monitors and manages computation in datacenters has to be aware of performance costs
    - Service response times are around 100ms, Max 10% batch job throughput hit
- Energy costs of datacenters are typically higher than green energy availability
  - Brown energy needs to be present to both supplement green and as “insurance” to meet performance constraints
  - Improvements in computation & networking infrastructure energy efficiency are necessary (power, thermal and cooling management)

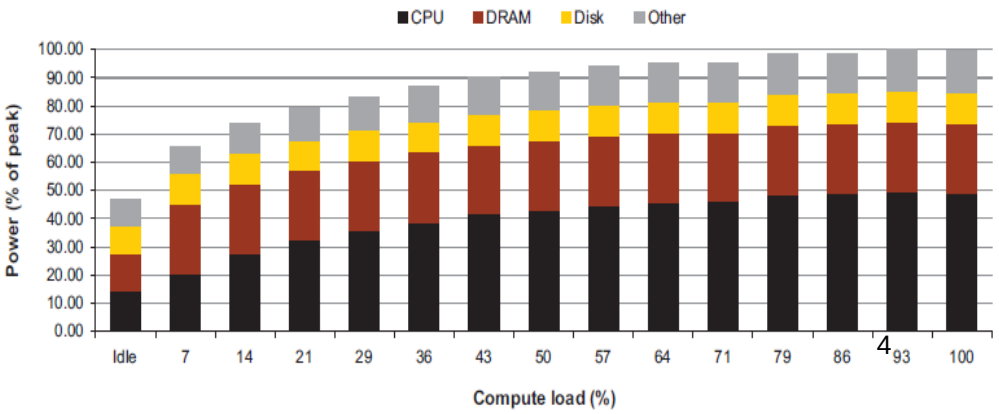
Datacenter energy efficiency

$$\text{Efficiency} = \frac{\text{Computation}}{\text{Total Energy}} = \left(\frac{1}{\text{PUE}}\right) \times \left(\frac{1}{\text{SPUE}}\right) \times \left(\frac{\text{Computation}}{\text{Total Energy to Electronic Components}}\right)$$





Barroso & Hölzle, 2009



Energy efficiency of the infrastructure

NSF Project GreenLight

- Green cyber-infrastructure in energy-efficient mobile facilities
- Closed-loop power and thermal management

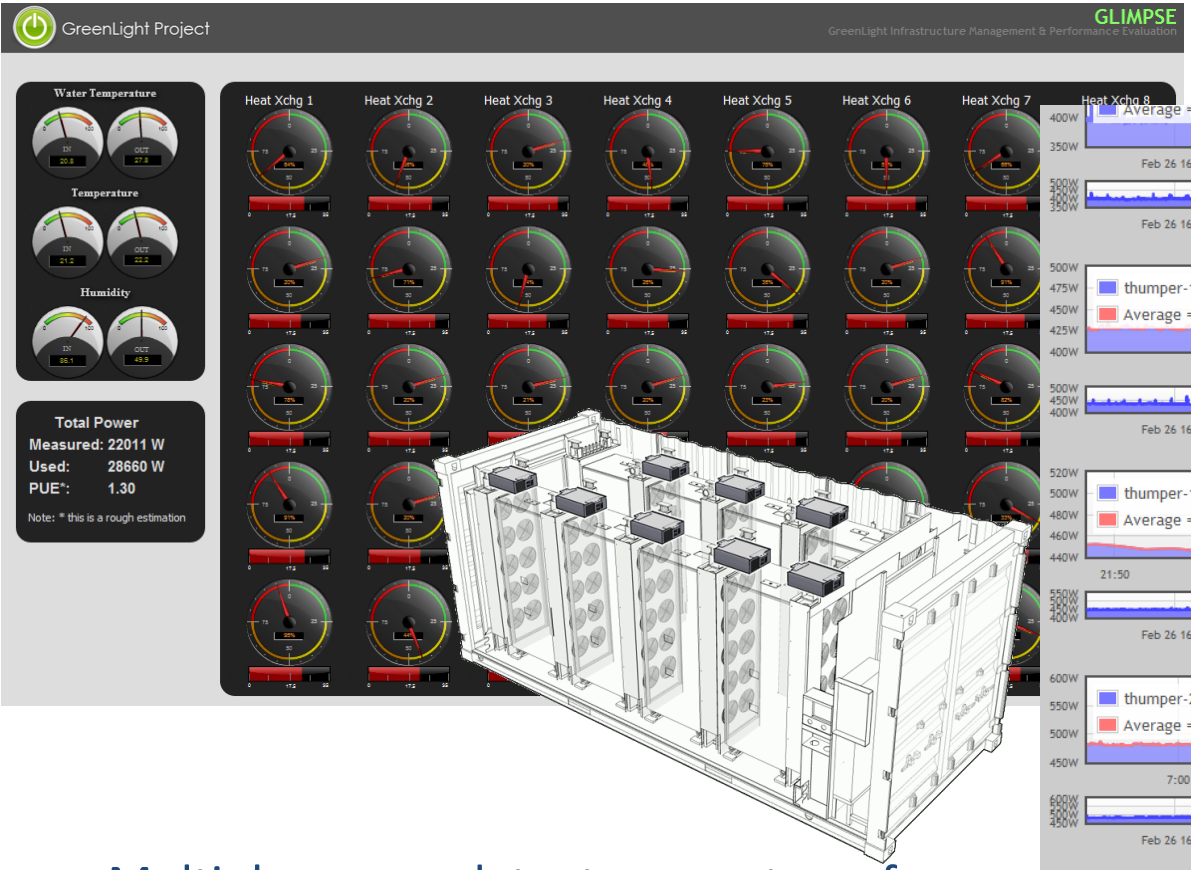
Dynamic power management (DPM)

- HW level: adaptive power gating gives 40% energy savings with no perf. impact
- SW level: 92% reduction in performance variability with DVFS
- Optimal DPM for a class of workloads
- Machine learning to adapt
  - Select among specialized policies
  - **Measured energy savings of 70%**

Dynamic thermal management (DTM)

- Workload scheduling:
  - Machine learning for dynamic adaptation
- Proactive thermal management
  - **Reduces thermal hot spots by average 80% with no performance overhead**
- Cooling aware management
  - **Savings of 70% in cooling subsystem**

NSF GreenLight: Dashboard & History plots

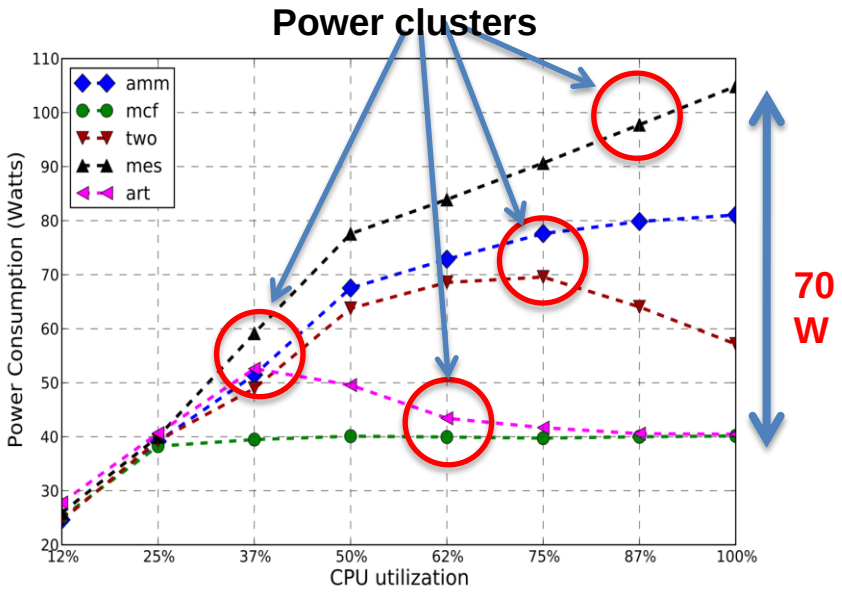


- Multiple sensor data: temperature, fan speed, liquid flow rate & temp, power
- Use measurements to develop models needed for energy management

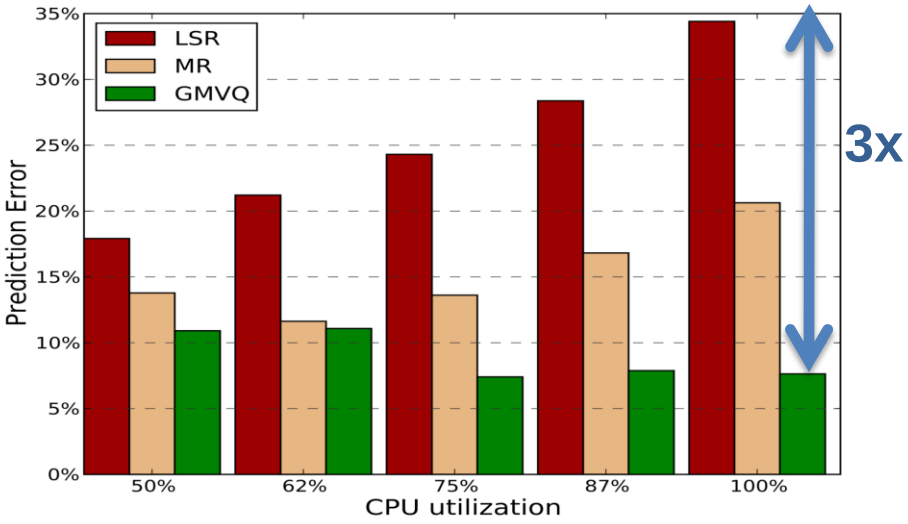
GMVQ VM Power Cost Prediction

- **Goal:** Estimate how much a VM consumes and predict what the cost would be if it migrates to another machine
- **Approach:** Gaussian Mixture Vector Quantization (GMVQ) to fit a GMM to the training data

CPU utilization is not enough!

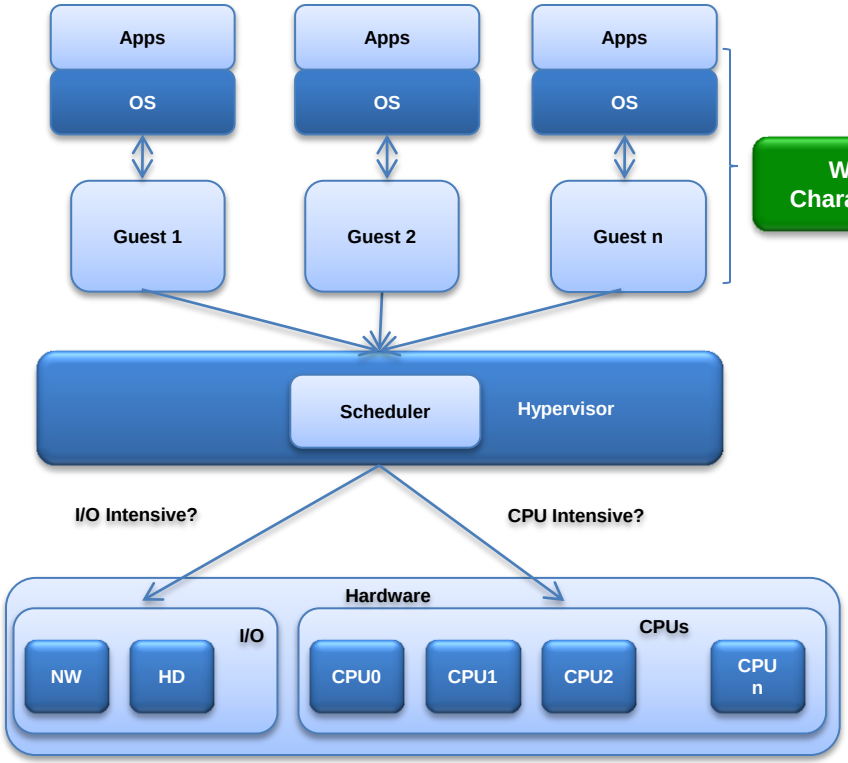


GMVQ is 3x better than regression

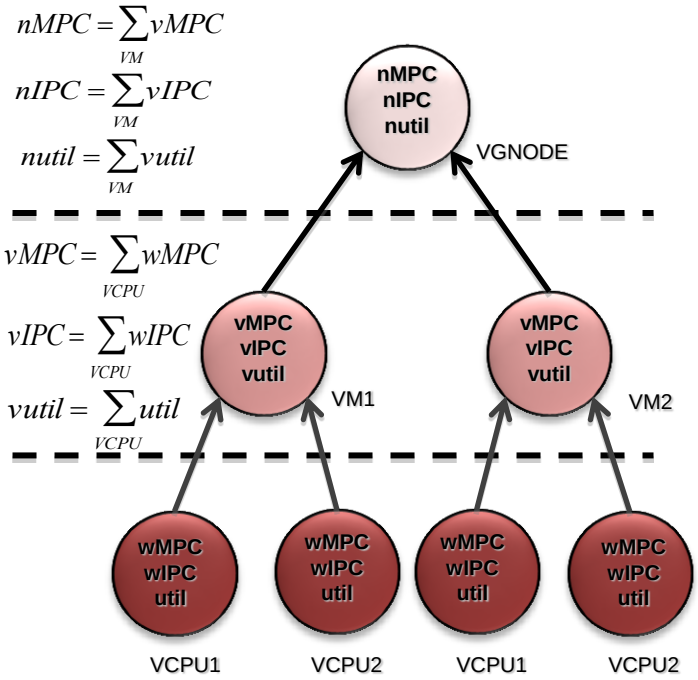


Energy management with virtualization

vGreen



Workload Characterization



- Scheduling
  - Co-locate guests with orthogonal characteristics
- Management policies
  - Based on the metrics maintained per guest

✓ **Avg 35% Energy Savings**

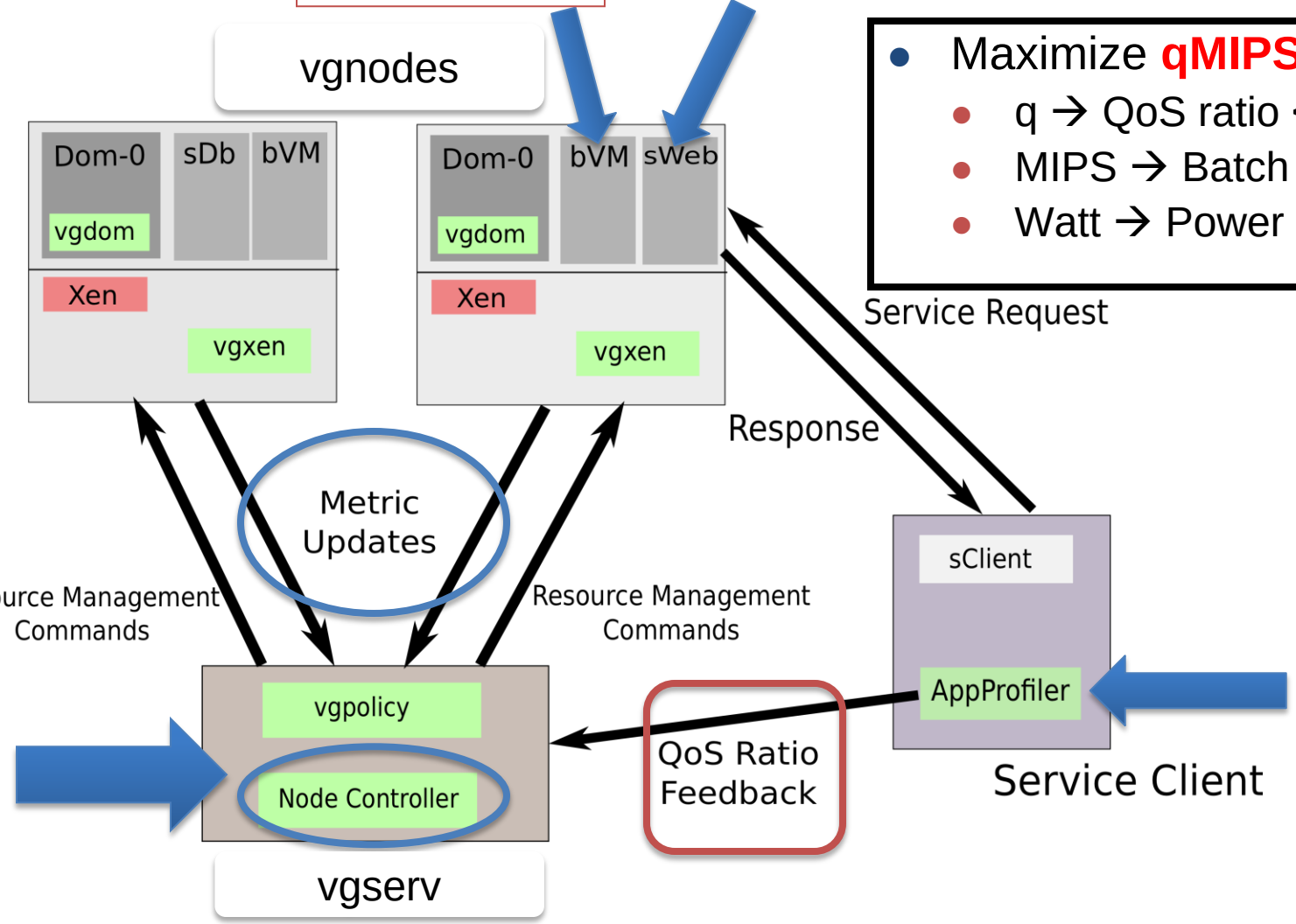
✓ **Avg 40% speedup**

VMs: Energy and QoS

vGreen+

Batch Jobs:
•MIPS driven

Services:
•Latency sensitive

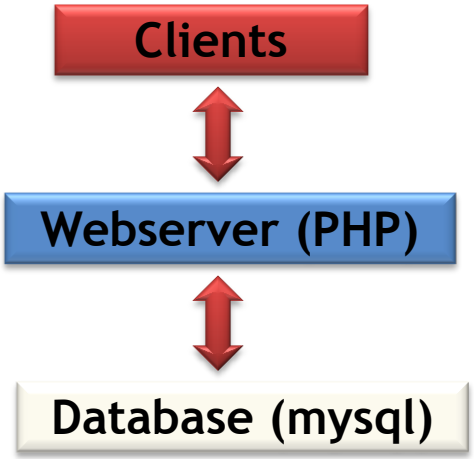


- Maximize **qMIPS/Watt**
  - q → QoS ratio < 1
  - MIPS → Batch job throughput
  - Watt → Power consumption

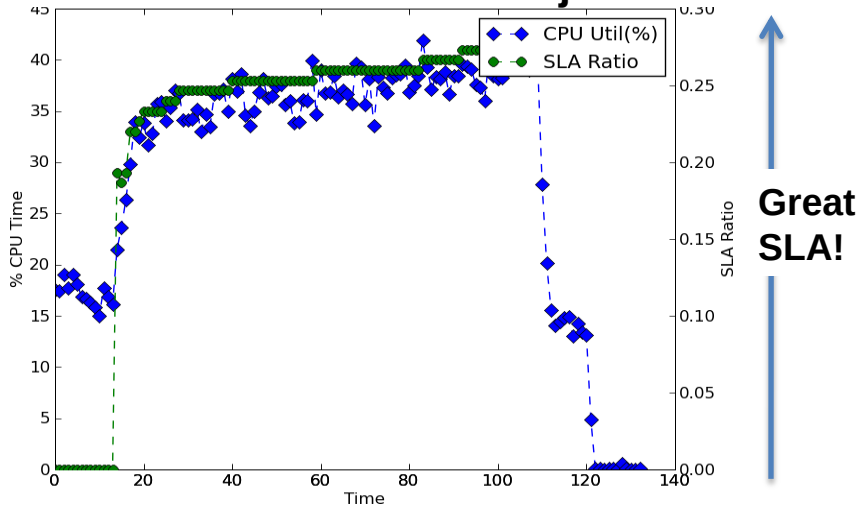
Managing multi-tier applications

- What happens when we combine:
  - Latency sensitive jobs (e.g. RUBIS)
  - Throughput sensitive (e.g. batch jobs)
- Preliminary results:
  - More than 10x improvement in SLA with background jobs relative to the default scheduler

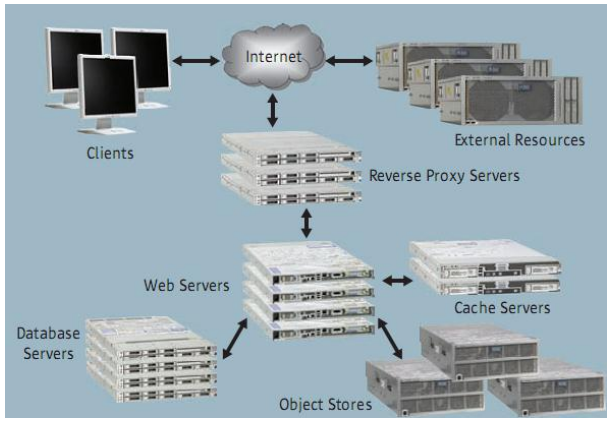
RUBiS: auction website



vGreen: Rubis & batch job

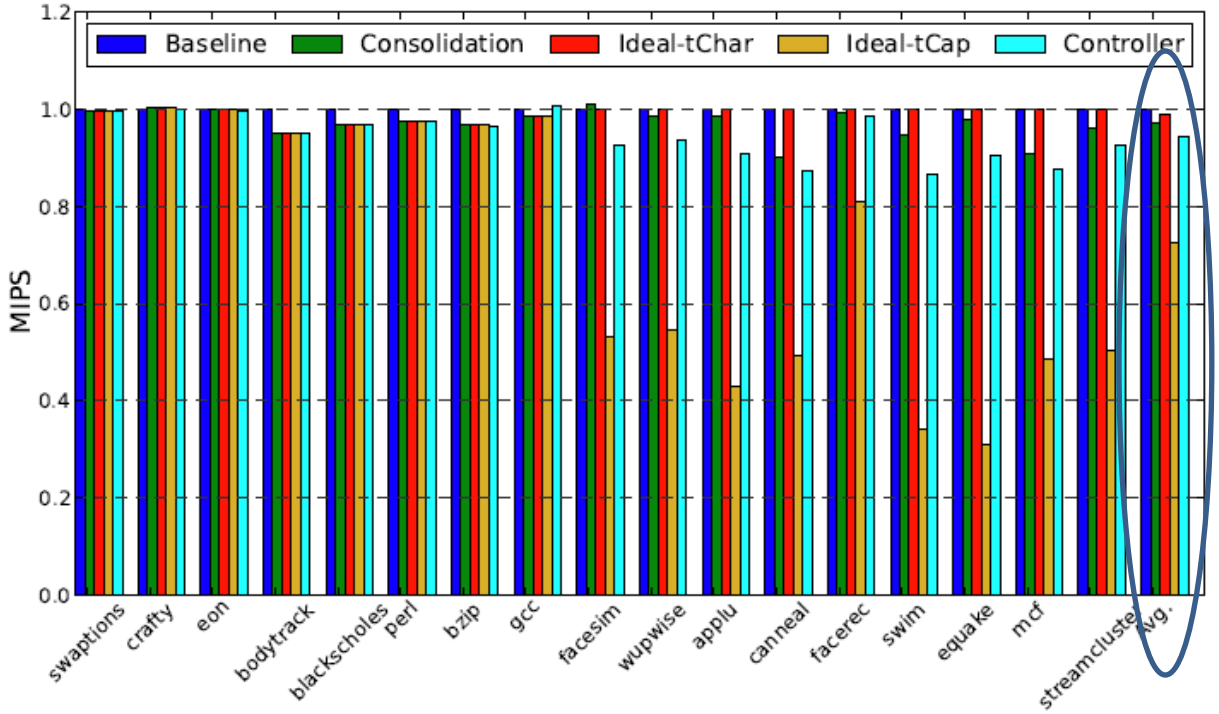


Great SLA!



VM Scheduling Policies Throughput

- State of the art:
  - Baseline: running services and batch jobs on separate servers
  - Selective Consolidation (tChar): vGreen
  - Capping (tCap): Cap the CPU allotment to bVM to mitigate interference effects (Padala@EuroSys'09, Nathuji@EuroSys'10)
- Controller:
  - Dynamically control the vCPU allocation of the service VM to maximize the batch job throughput while meeting service response times

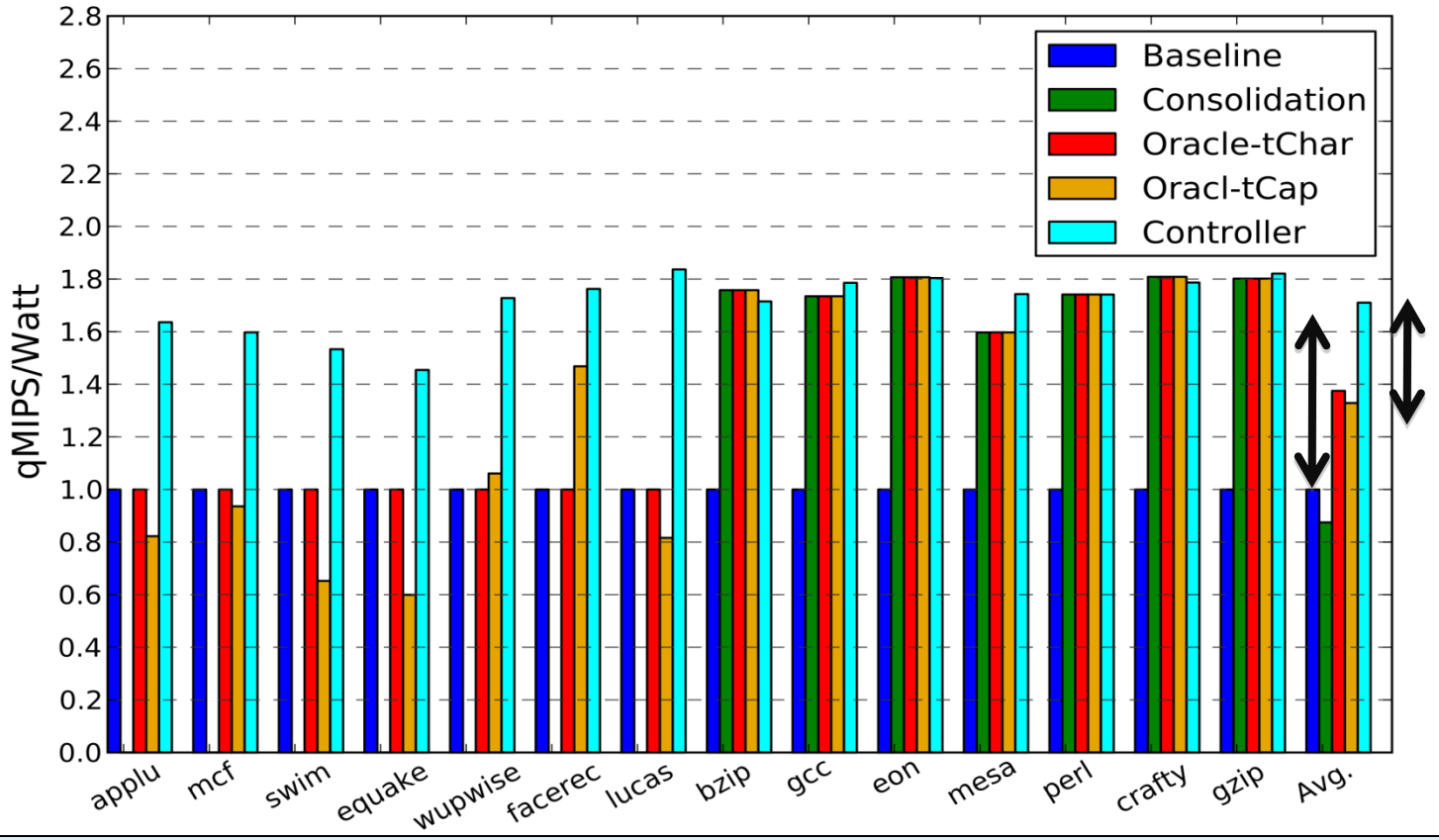


Batch job throughput

Our controller is within 7% of baseline; CPU capping has 25% lower average throughput

Energy Efficiency Improvements

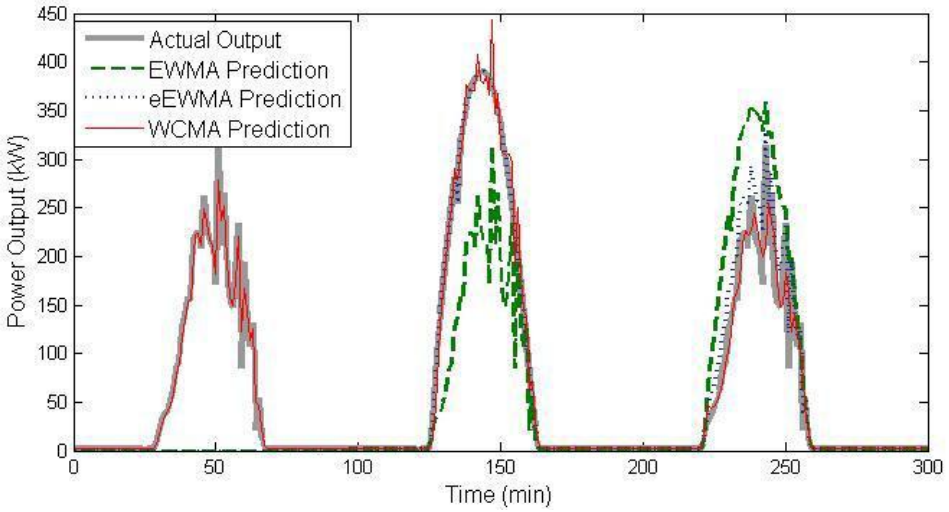
- Use a controller to manage virtual CPUs dynamically
  - Maximize CPUs of various batch jobs while meeting Rubis SLAs

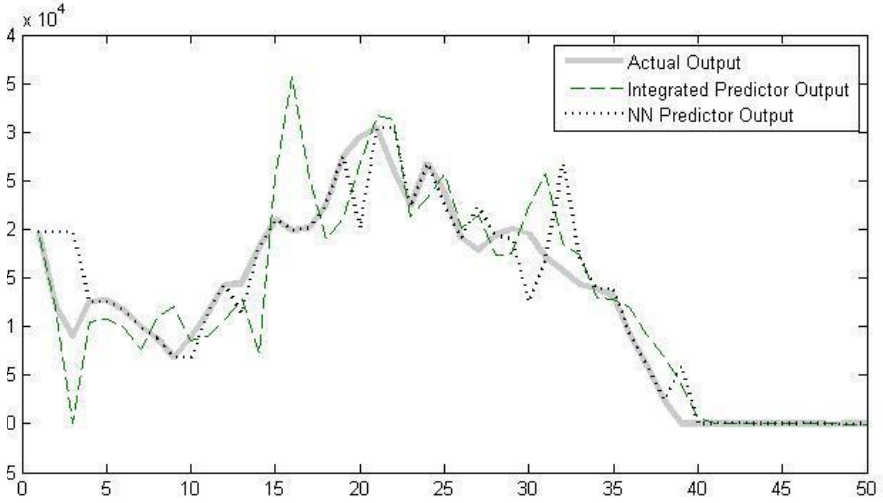


- **70% more efficient than running service & batch VMs separately while within 7% of maximum batch job throughput**
- **35% more efficient than the ideal version of state of the art**

Green Energy Prediction

- **Predict green energy availability for the next 30min window**
  - Schedule additional MapReduce jobs accordingly; they take max 30mins



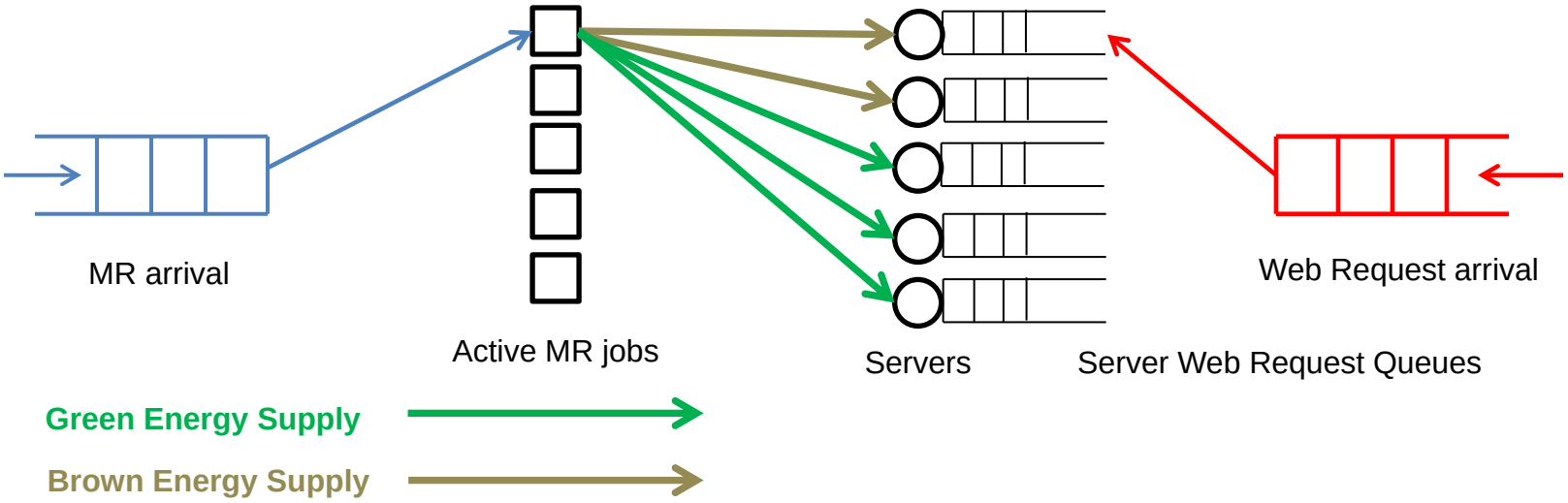


- Data from solar panels at UCSD
- State of the art: exponential weighted average
  - EWMA: 32.6 % error
  - Extended eEWMA: 23.4% error
- **Our algorithm:**
  - **WCMA: less than 9.6% error**

- Data gathered from a wind farm in Lake Benton, available by NREL
- State-of-the-art:
  - Integrated predictor: 48.2% error
- **Our algorithm: 21.2% error**
  - Combination of a weighted nearest-neighbor (NN) tables and wind power curve models

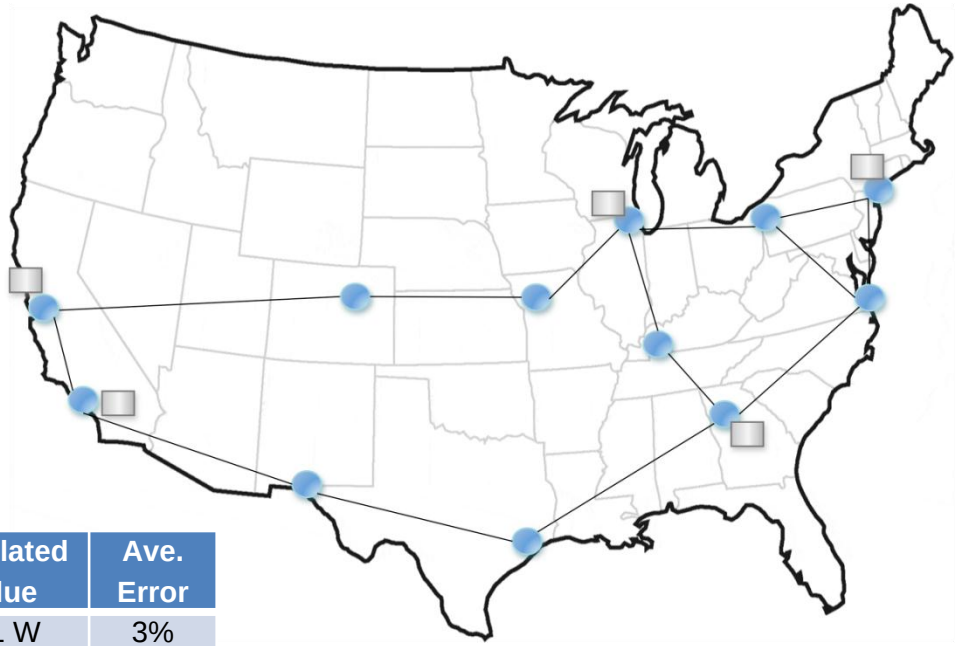
Methodology

- Use green energy to schedule “extra” batch jobs.
  - MapReduce (MR) type jobs for this purpose.
  - Initiate more subtasks with the available green energy.
    - Increased throughput
    - Reduced completion time - limit reduction to 10%
    - Kill a subtask if the green energy supply level drops



Experimental setup & validation

- Globally distributed datacenters connected
  - 5 datacenters, 12 routers, modeled after ESnet
  - Solar traces from UCSD
  - Wind traces from NREL

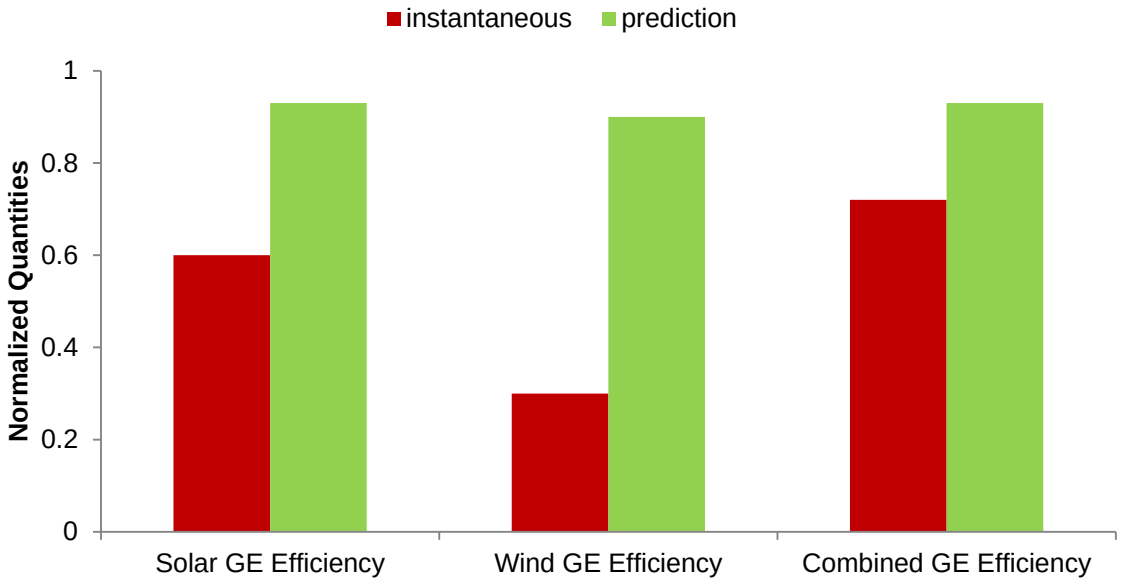


Simulator validation	Measured Value	Simulated Value	Ave. Error
Avg. Power Consumption	246 W	251 W	3%
Rubis QoS ratio	0.08	0.085	6%
Avg. MapReduce Completion Time	112 min	121 min	8%

- Jobs run within vGreen VMs on Nehalem server
  - Rubis used for services with 100ms 90th%ile response time constraint
  - MapReduce used for batch jobs with 10% max job completion time reduction (max 5 cores on Nehalem server)
  - Use VM migration (with quantified performance impact)

Benefits of Green Energy Prediction

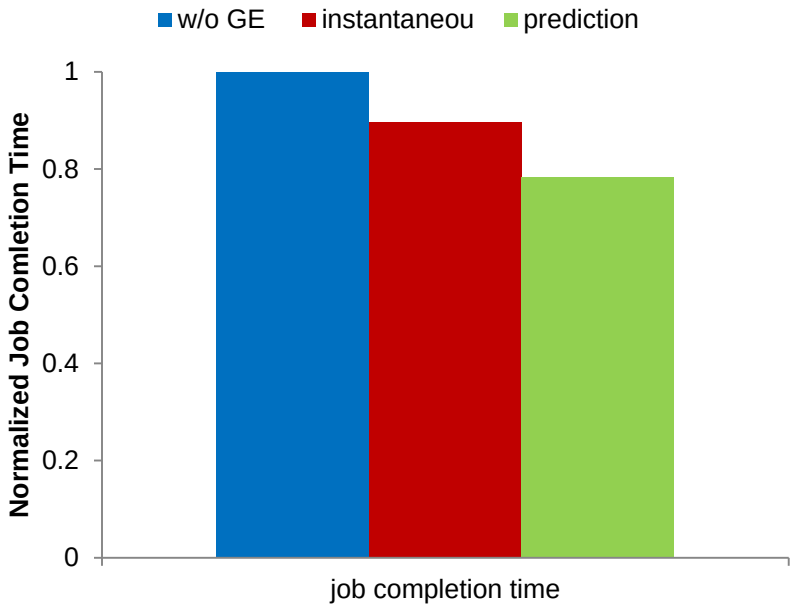
- Compare our green energy *predictive* jobs scheduler with *instantaneous* usage of green energy

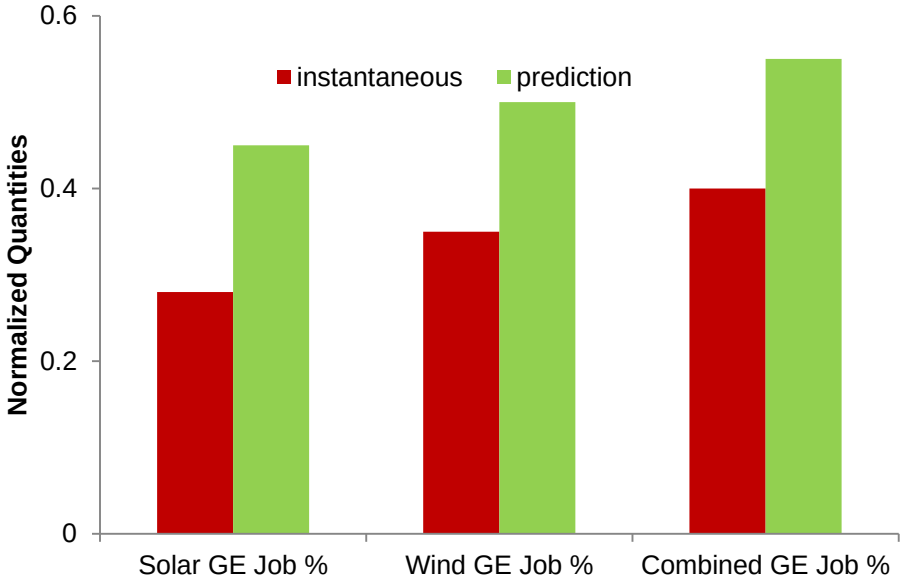


Prediction has 93% GE Efficiency

GE Efficiency: ratio of green energy consumed for useful work vs. the total green energy available

Benefits of Green Energy Prediction





Prediction has 22% faster batch job completion time vs. instantaneous

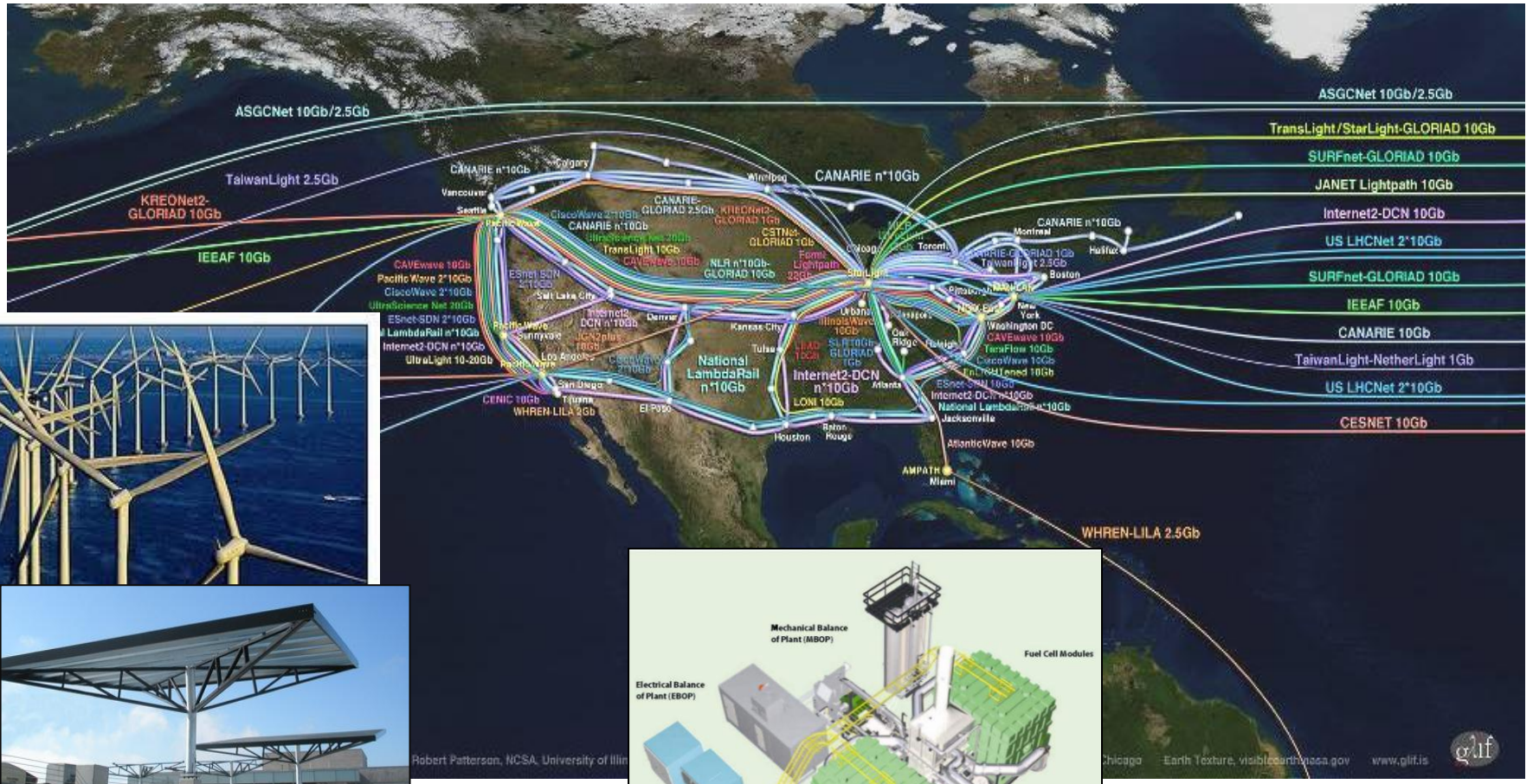
On average, **5x fewer batch tasks need to be terminated** when using GE prediction vs. instantaneous usage

38% more jobs complete with prediction vs. instantaneous

GE Job %: ratio of batch jobs completed with GE over all jobs.

Next steps: Green energy powered global routing

(jointly with Inder Monga, Esnet)



Wind from NREL, 2MW Solar @UCSD

2.8MW Fuel Cell Power Plant , UCSD

Focus Center Research Program

Sponsors

SIA Semiconductor Industry Association

Intel
Freescale
IBM
GLOBALFOUNDRIES
MICRON
Texas Instruments
Xilinx
AMD

semi

Applied Materials
Novellus
Cadence

DOD & DARPA

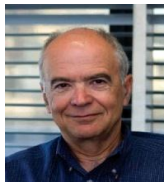
Raytheon
United Technologies

Multi-Scale Systems Research Center [10 Universities]
Director: Prof. Jan Rabaey (UC-Berkeley). High-level systems design addressing distributed sense and control systems, large-scale and small-scale information technologies systems.

Gigascale Systems Research Center [15 Universities]
Director: Prof. Sharid Malik (Princeton). Platform architectures; concurrent systems programming; platform viability; resilient systems and alternative computation models.

Center for Circuit & Systems Solutions [13 Universities]
Director: Prof. Larry Pileggi (CMU). Circuit/module infrastructure; enterprise systems; portable electronics; functional diversity and emerging circuits for post CMOS.

Interconnect Focus Center [13 Universities]
Director: Prof. Paul Kohl (Georgia Tech). Nanoscale electrical & optical interconnects; energy delivery and thermal management; wireless connectivity; modeling, analysis and assessment of new connectivity solutions.

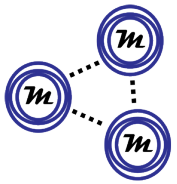


Materials, Structures and Devices [15 Universities]
Director: Prof. Dimitri Antoniadis (MIT). Integration of new materials enabling CMOS extension; carbon-based devices; novel embedded memory; functional diversification; theory, modeling and simulation of new devices.

Functional Engineered Nano-Architectonics [14 Univ.]
Director: Prof. Kang Wang (UCLA). Novel materials and processes which enable fabrication of nanoscale devices and interconnects.

MuSyC in a Nutshell

GRAND CHALLENGE : "Energy-smart" distributed systems, that
- Are deeply aware of balance between energy availability and demand
- Adjust behavior through dynamic and adaptive optimization through all scales of design hierarchy.

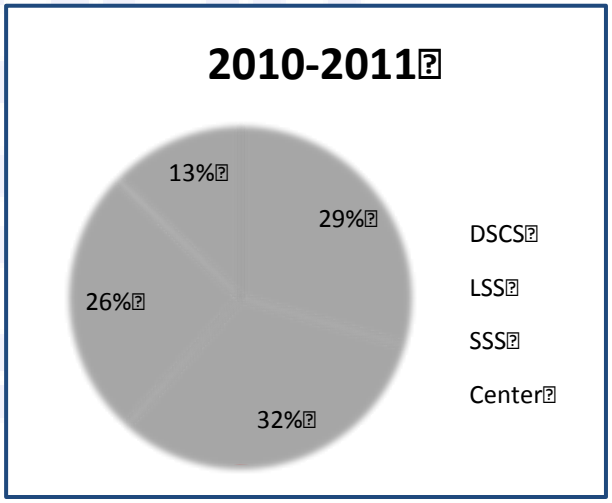


DSCS Theme
Distributed sense and control systems
Target: Airborne Platforms (Avionics)
5 faculty (reduced from 8), 3 locations

LSS Theme
Large-scale "energy-intensive" systems
Target: Data centers
9 faculty (stable), 5 locations

SSS Theme
Small-scale "energy-frugal" systems
Target: Human enhancement
5 faculty (reduced from 8), 4 locations

- 19 Faculty Distributed over 9 US Universities
- 60 Students (many only partly funded)
- 109 publications
- Monthy e-seminars and bi-annual e-workshops

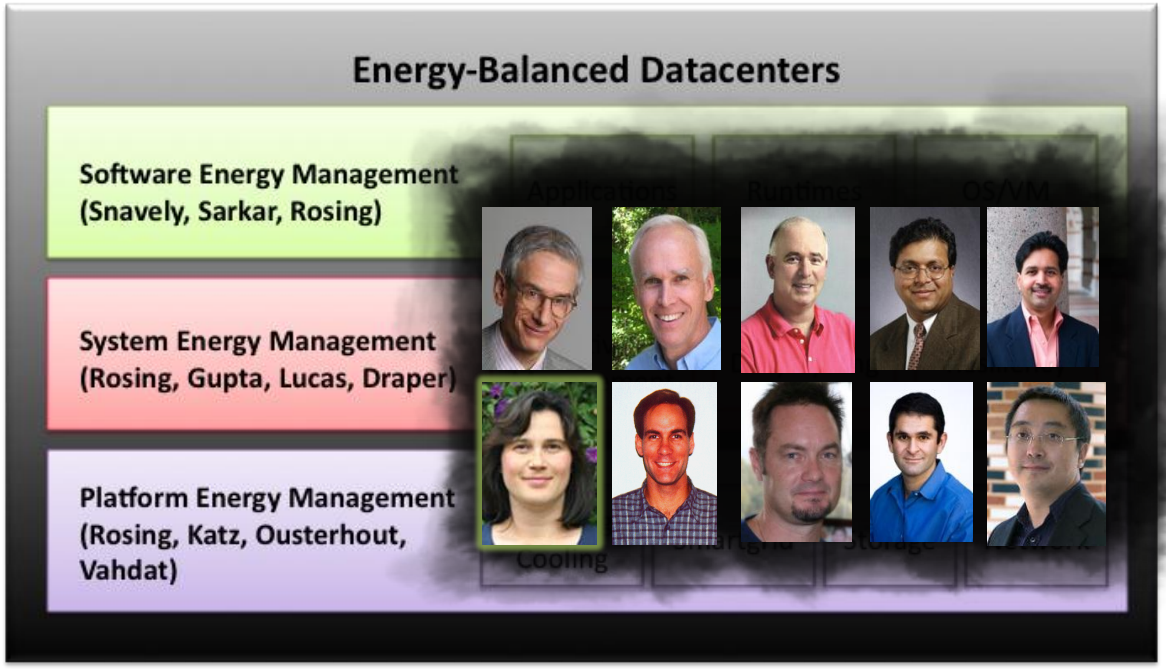


Multiscale Systems Center: Energy Balanced Datacenters

Theme leader: Tajana Simunic Rosing

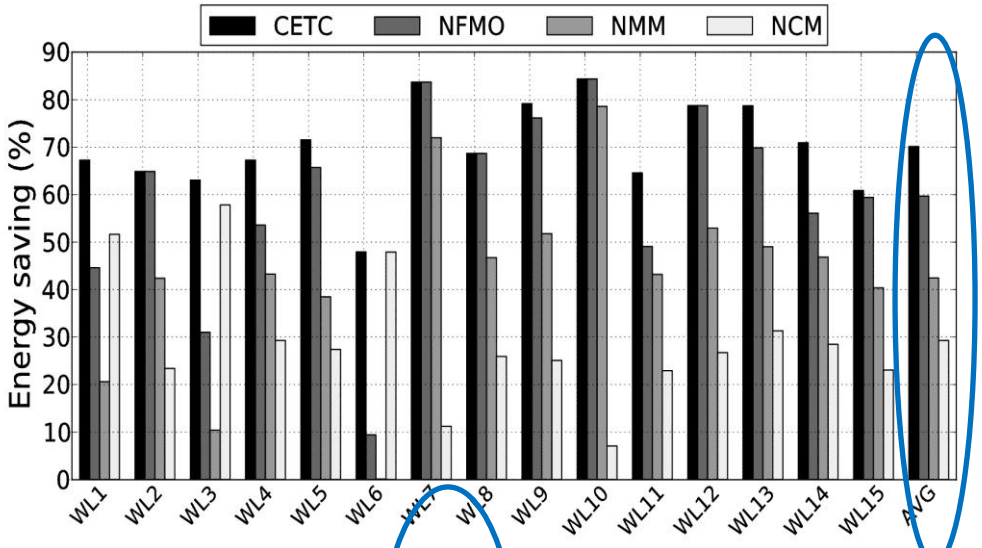
Realize **closed-loop energy management strategies** that enable **"energy-intensive" large-scale systems** to be **orders of magnitude more energy-efficient**, while ensuring that mission-critical goals are met.

"Doing nothing well"



MuSyC
The Multiscale Systems Center

Combined Energy Thermal & Cooling, CETC Management Results



Local ambient Temp.	# DIMM	CETC %	NFMO %	NMM %	NCM %	DLB %
45 °C	8	0.175	0.184	0.093	0.102	0
35 °C	8	0.115	0.120	0.069	0.100	0
45 °C	16	0.175	0.187	0.109	0.102	0

- **CETC performance overhead < 0.2%**
- **CETC page migration rate < 5 pages/sec -> negligible overhead & high stability**

- **Average cooling savings of 70% relative to state-of-the-art**

Intel Xeon Dual Socket Quad core Server; with state-of-the-art PI fan controller

- **CETC:** Our policy
- **DLB:** Dynamic load balancing (baseline)
- **NFMO:** Only page migrations allowed
- **NMM**: No memory clustering
- **NCM**: No CPU scheduling optimizations

Workload	Socket A	Socket B	Ambient °C
W1	*3eon*	*eon + mcf + gcc*	42
W2	*2eon + mcf*	*eon + bzip2 + mcf*	42
W3	*2bzip2 + 2mcf*	*2bzip2 + 2mcf*	42
W4	*2perl + 2eon*	*2gcc + 2mcf*	42
W5	*2perl + 2bzip2*	*2gcc + 2mcf*	39
W6	*2perl + bzip2*	*gcc+2mcf*	42
W7	*perl + 3gcc*	*perl + 2gcc*	40
W8	*perl + 3gcc*	*perl + 2gcc*	42
W9	*3eon*	*3eon*	42
W10	*2eon+mcf*	*2eon+mcf*	42
W11	*2bzip2*	*2bzip2*	42