Sustainability (and Energy Management) in the Cloud

Cullen Bash
Distinguished Technologist
Hewlett Packard Laboratories
June 14, 2010

OUTLINE

- Motivation
- Sustainability
- The Sustainable Cloud
- Concluding Remarks

Motivation

Impact of IT on Environmental Sustainability

using Carbon Emissions as a Proxy

Industry challenge:

1) The Cloud will continue to grow and consume resources
2) Build a sustainable cloud (Focus of this talk)
3) Use the cloud to impact the 98% (Panel topic?)

Total carbon emissions

IT industry

2%

The rest of the global economy

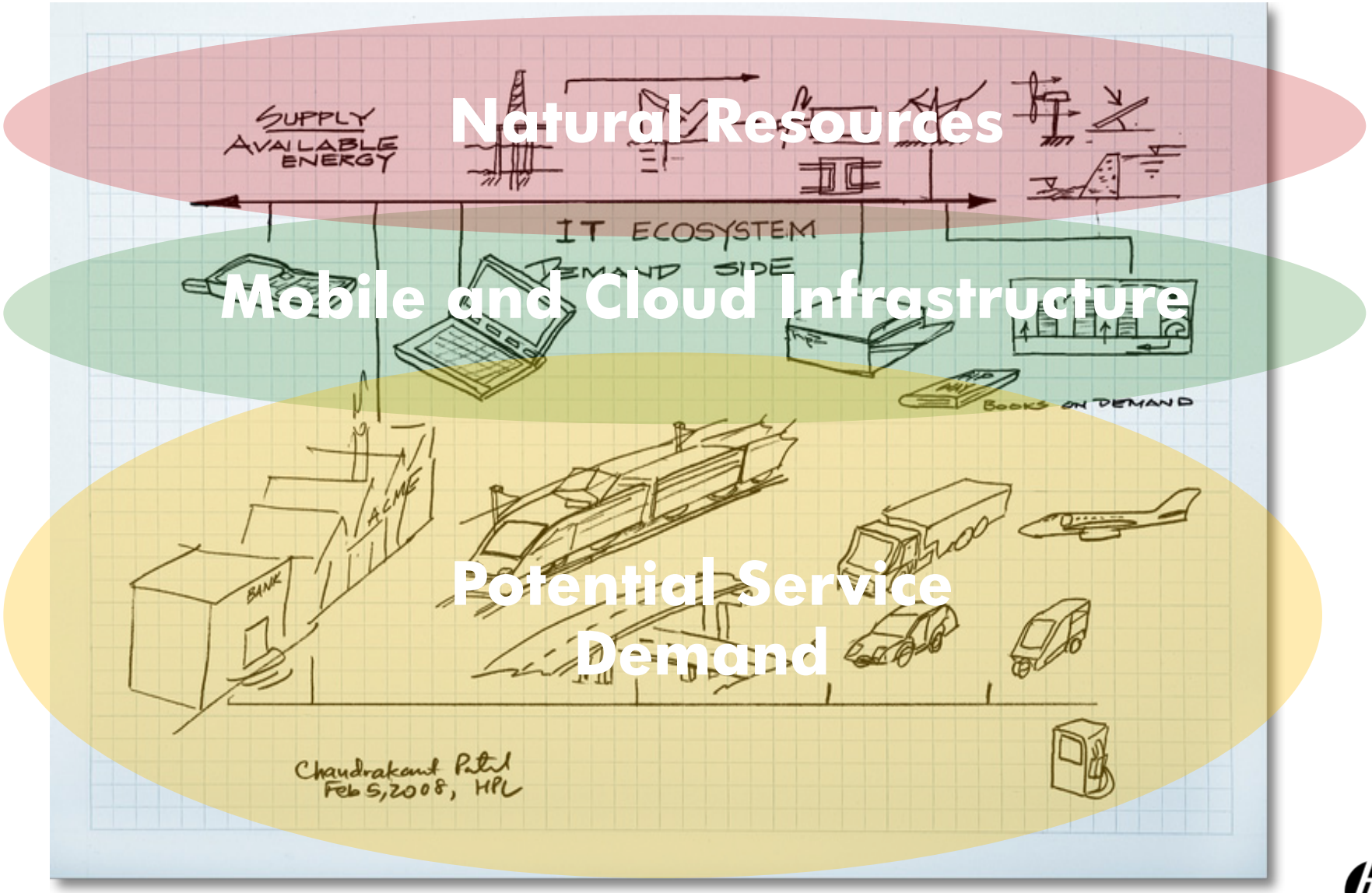
98%

IT *could* play a central role in addressing the global sustainability challenge.

All Rights Reserved. No copies may be made, in full or part, without written permission of the authors.

The Global Cloud Computing Ecosystem

billions of service oriented client devices and thousands of data centers….to deliver net positive impact



All Rights Reserved. No copies may be made, in full or part, without written permission of the authors.

Sustainability

Sustainability: "Triple Bottom Line"

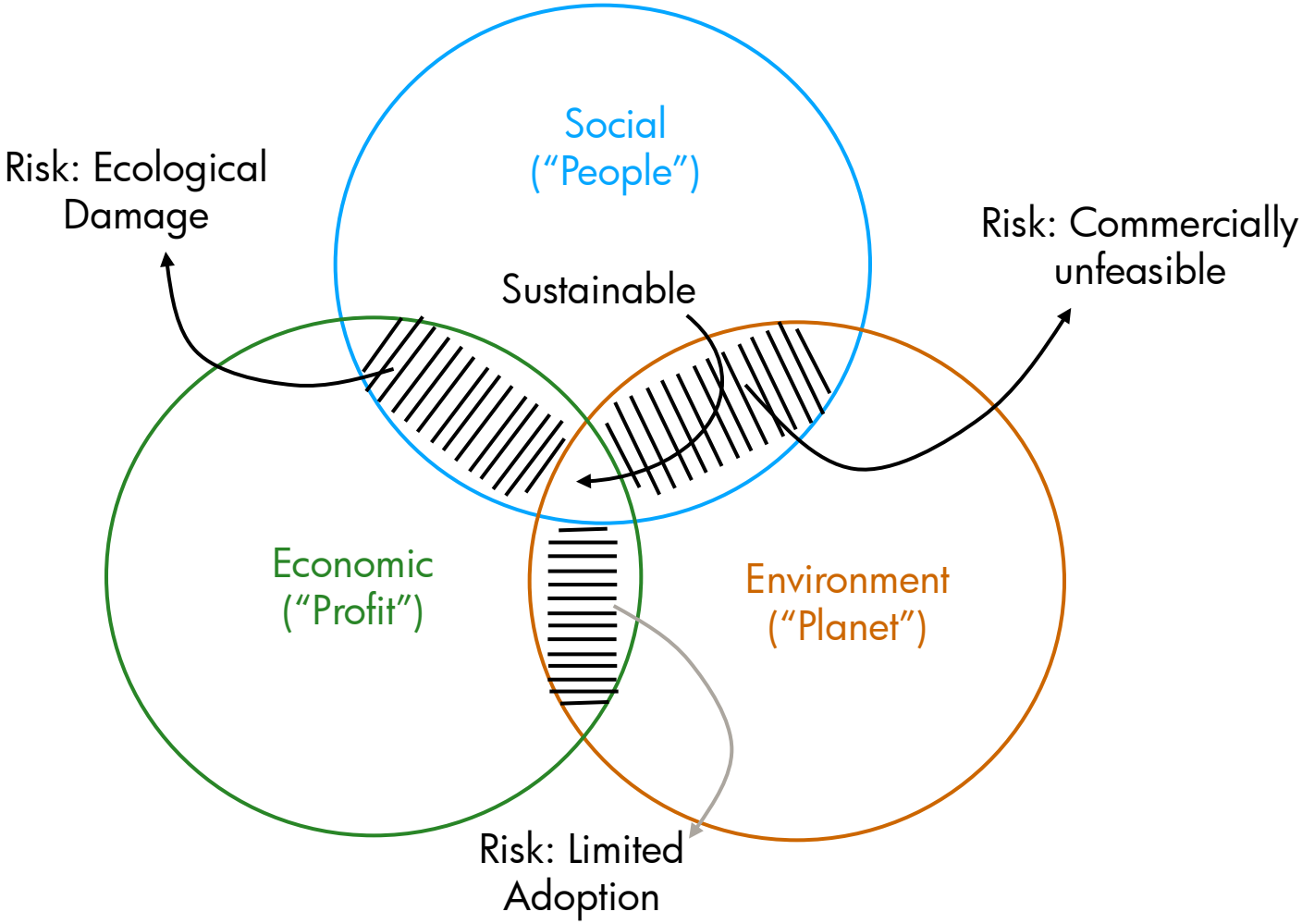


Figure Credit: A. Agogino, UC Berkeley

The Sustainable Cloud

Elements of a Sustainable Cloud

1. Lifecycle Design and Sizing
2. Sustainability monitoring and metric development
3. Sustainable service management
4. Matching resource demand with supply

None of the above are unique to Clouds. But, Clouds provide an opportunity to scale the benefits globally.

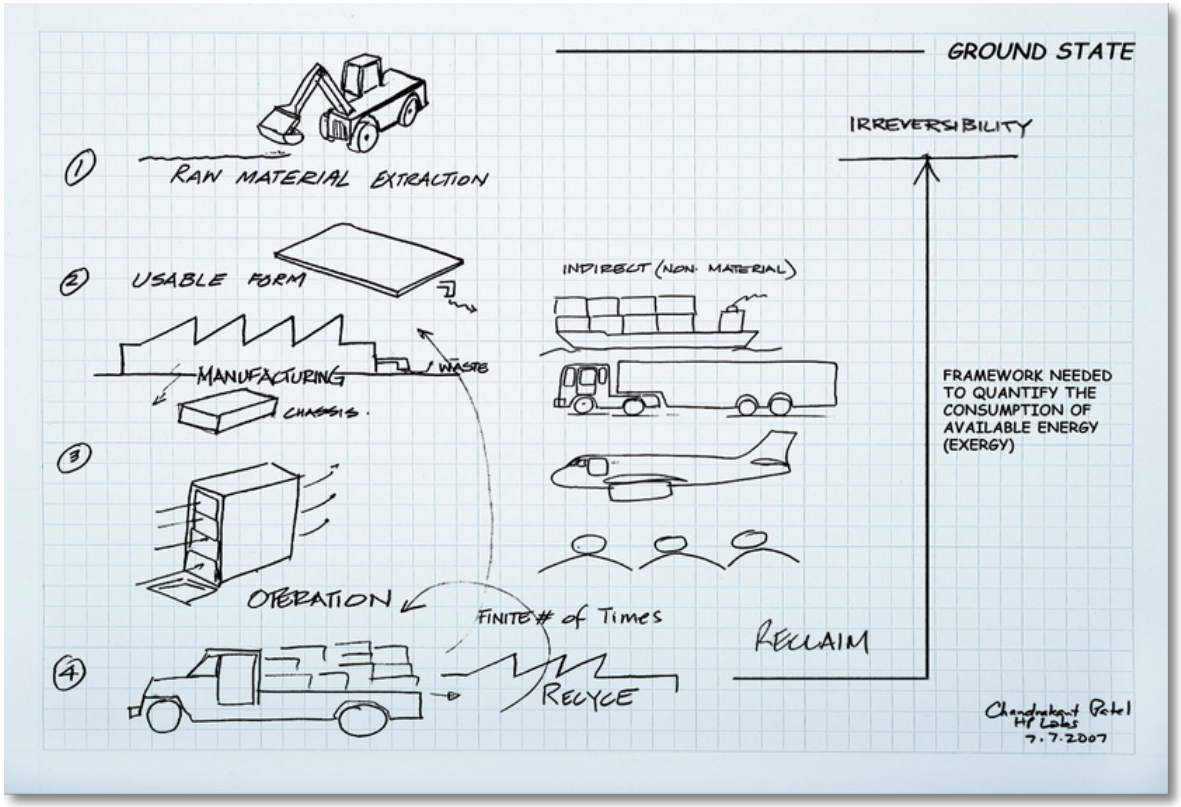
Topic #1: Lifecycle Design and Sizing

Goal: Take a lifecycle view of infrastructure.
(i.e., don't focus *only* on operation)

Lifecycle Design and Exergy *(not Energy)*

Exergy: Quantification of the *quality* of energy

- An extension of Life-Cycle Assessment (LCA)
  - Track flow of material and energy across the entire life-cycle
  - Perform an exergy analysis of each flow

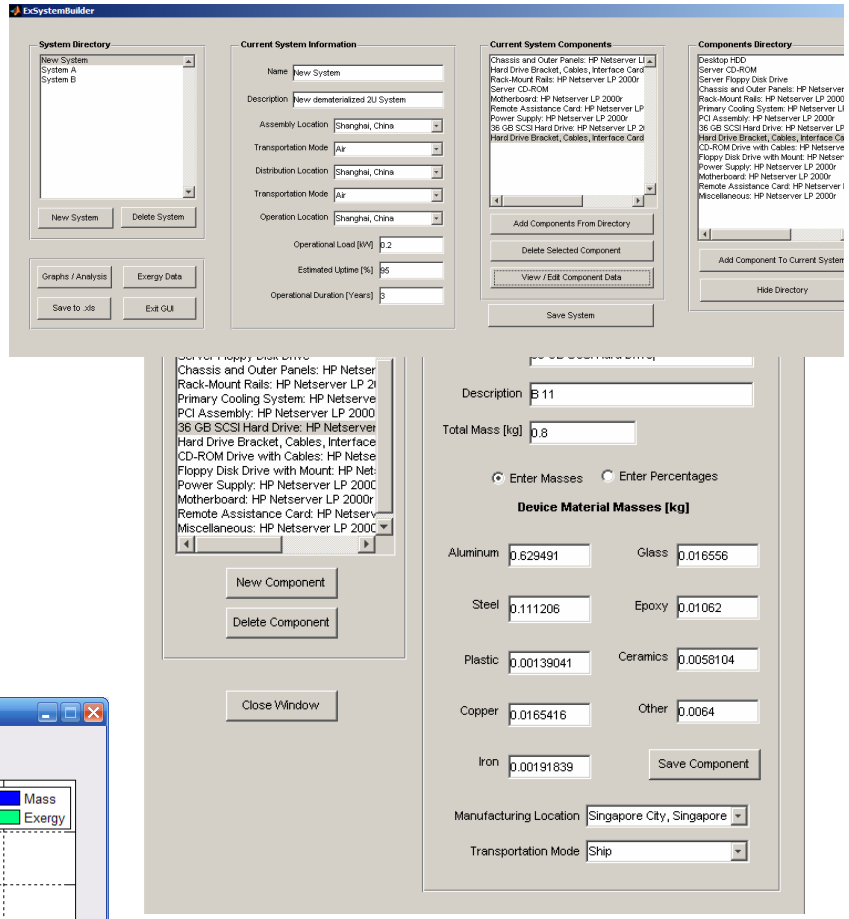


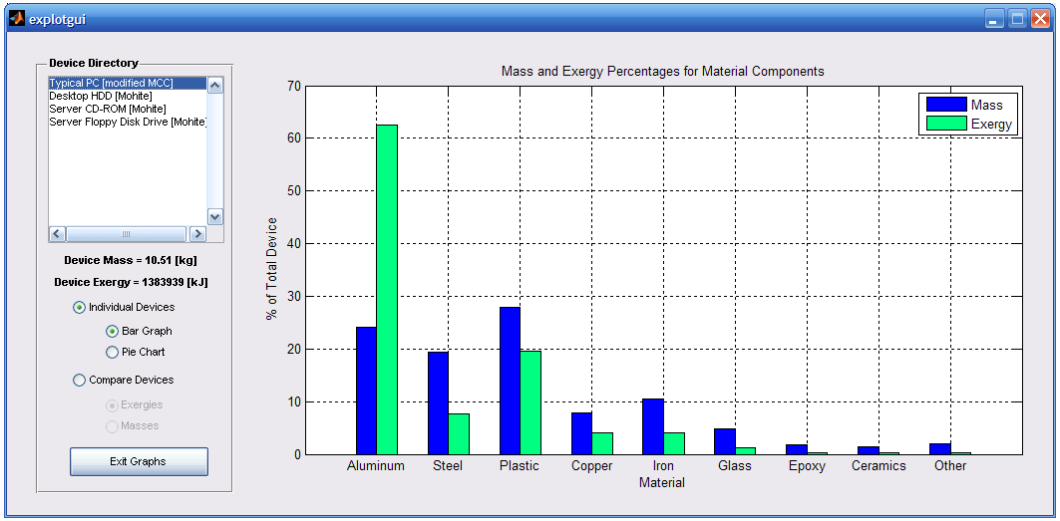
Lifetime exergy loss framework can predict consumption of lifetime resources as well as operational energy

Lifetime Exergy Advisor

Joint work with UC Berkeley

- Tool can be used to devise environmental design strategies for different product families
- Illustrative case study:
  - *Mobile clients* have a significant fraction of embedded vs. operational exergy destruction
  - But, nearly 95% of environmental impact for an *enterprise server* due to use-phase power and cooling





LEA will be a unique software package to help reduce product environmental footprints

Case Study: An Enterprise Server

Can we do better if we design servers specifically for Clouds?

Similar study applied to data centers shows a more balanced distribution of impact among embedded and operational segments.

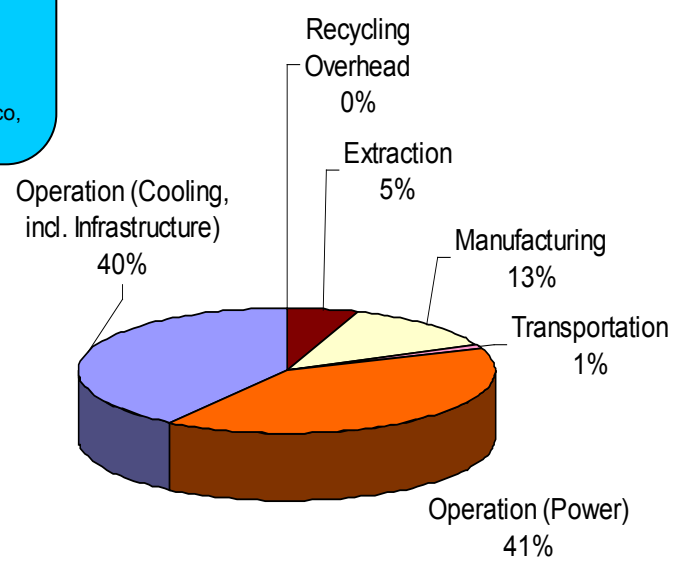
→Holistic optimization is important.

Shah, A., et. al. "The Environmental Footprint of Data Centers", InterPack, San Francisco, July 2009

Destroyed

Operation:
A_d ~ 29 GJ

Across the life-cycle:
A_d ~ 26 GJ to 34 GJ



Hannemann, C., et. al., "Lifetime Exergy Consumption as a Sustainability Metric for Enterprise Servers", Proceedings of ASME Energy Sustainability, August 2008

Ref: Shah et al., IEEE Itherm (2010)

Infrastructure Optimization

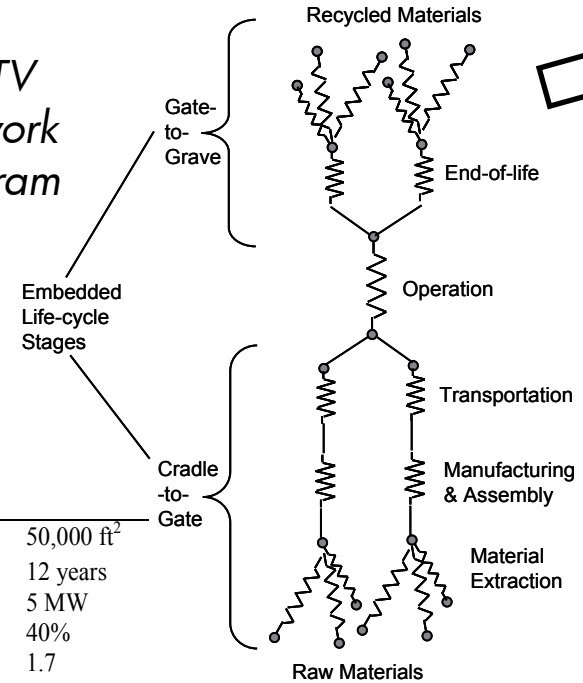
The exergo-thermo-volume (ETV) approach

– A method for quantifying trade-offs between thermal management requirements, operational power, and embedded exergy consumption

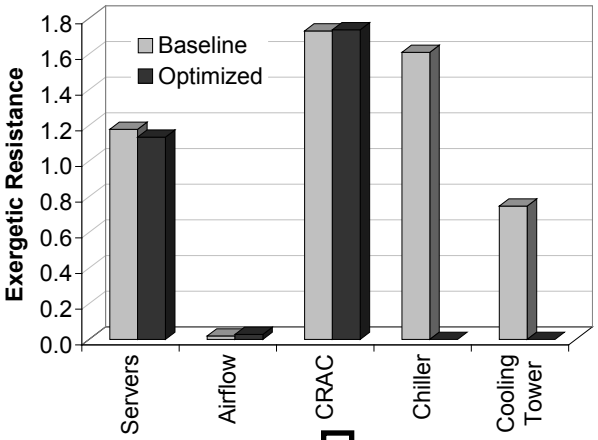
(1) Data Center Specification

Floor Area of Data Center:	50,000 ft^2
Lifetime of case study:	12 years
Design (peak) IT load:	5 MW
Average IT load:	40%
COP_G:	1.7
Power Delivery Losses:	10%
Miscellaneous Building Load:	2%
Number of servers:	20,000
Redundancy:	N+1

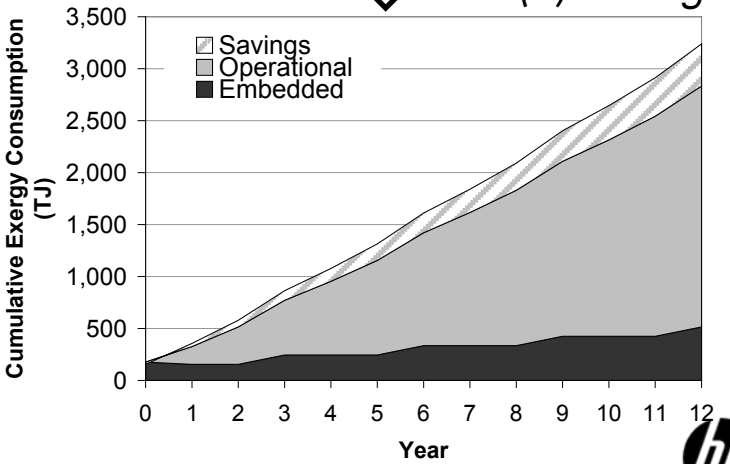
(2) ETV Network Diagram



(3) ETV Resistances



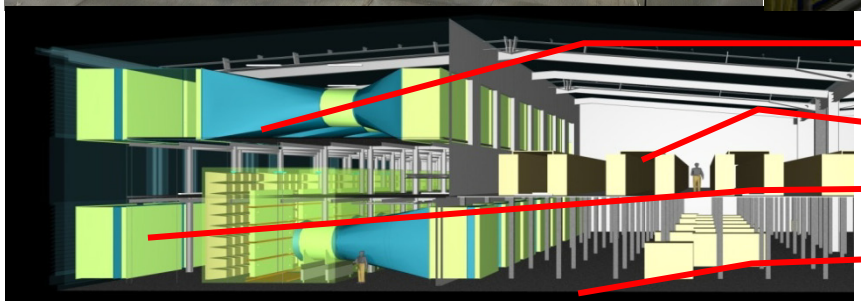
(4) Savings



Using ETV to identify bottlenecks within data center infrastructure

An Example: The Wynyard Cooling Solution

Resource Economization



Low-speed exhaust fans – 8 x 2.1m Diameter

Enclosed cold aisles

Mixing & Humidification Chamber

Large floor plenum (5m)

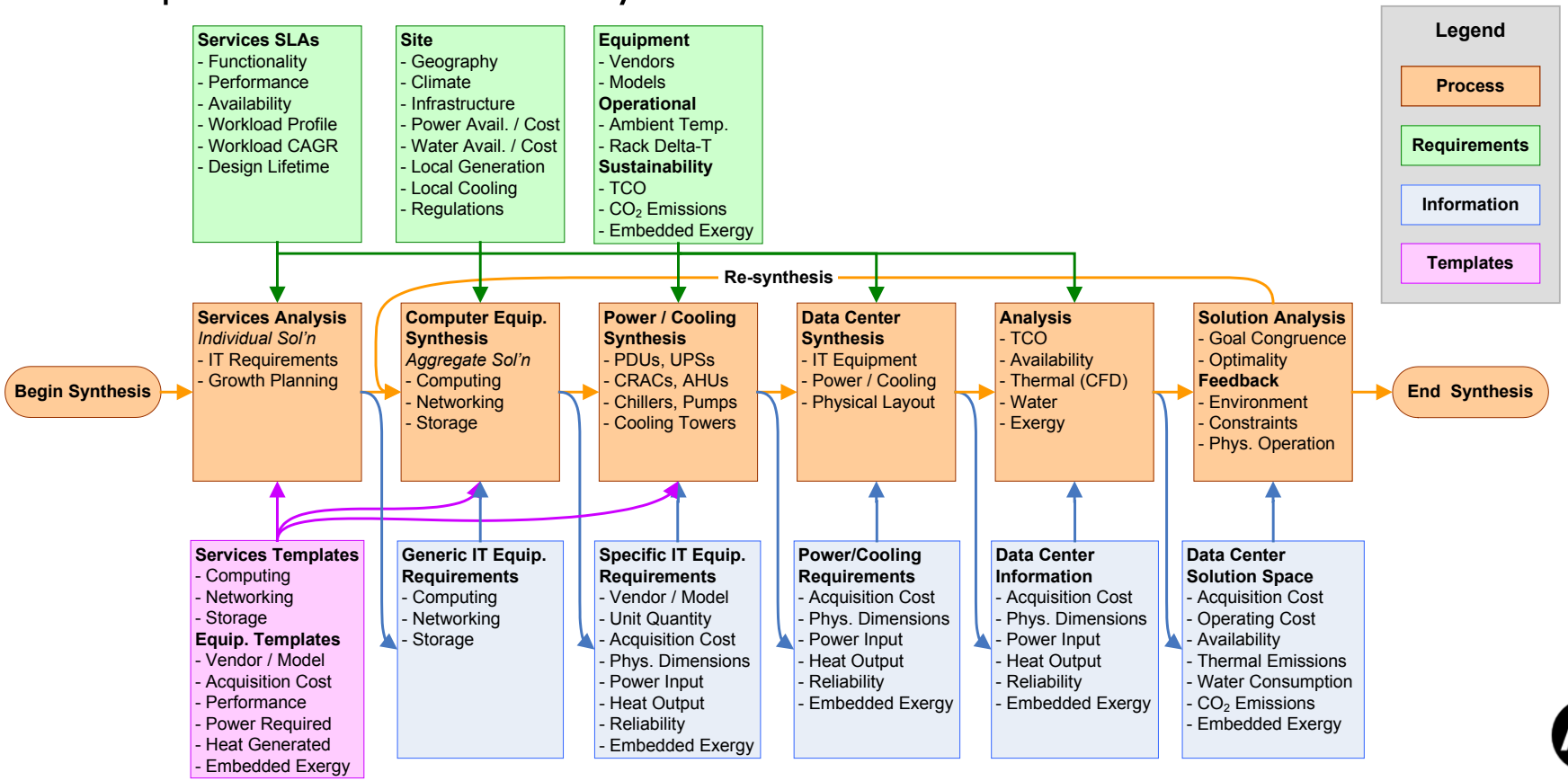
Measures to Reduce Environmental Impact:
Rainwater harvesting
Re-use of cladding and concrete
Replacement of existing paved parking areas with soft landscaping
Increased biodiversity through planting and habitat improvements
Intelligent lighting system
High reflectance racking

Wynyard Data Center Facts:
300,000 square feet
4 data halls built of 8
Tested PUE 1.16 per data hall, 1.2 overall
Converted distribution center
Tier III capabilities
BREEAM Excellent Rating

Lifecycle Design and Sizing: Workloads

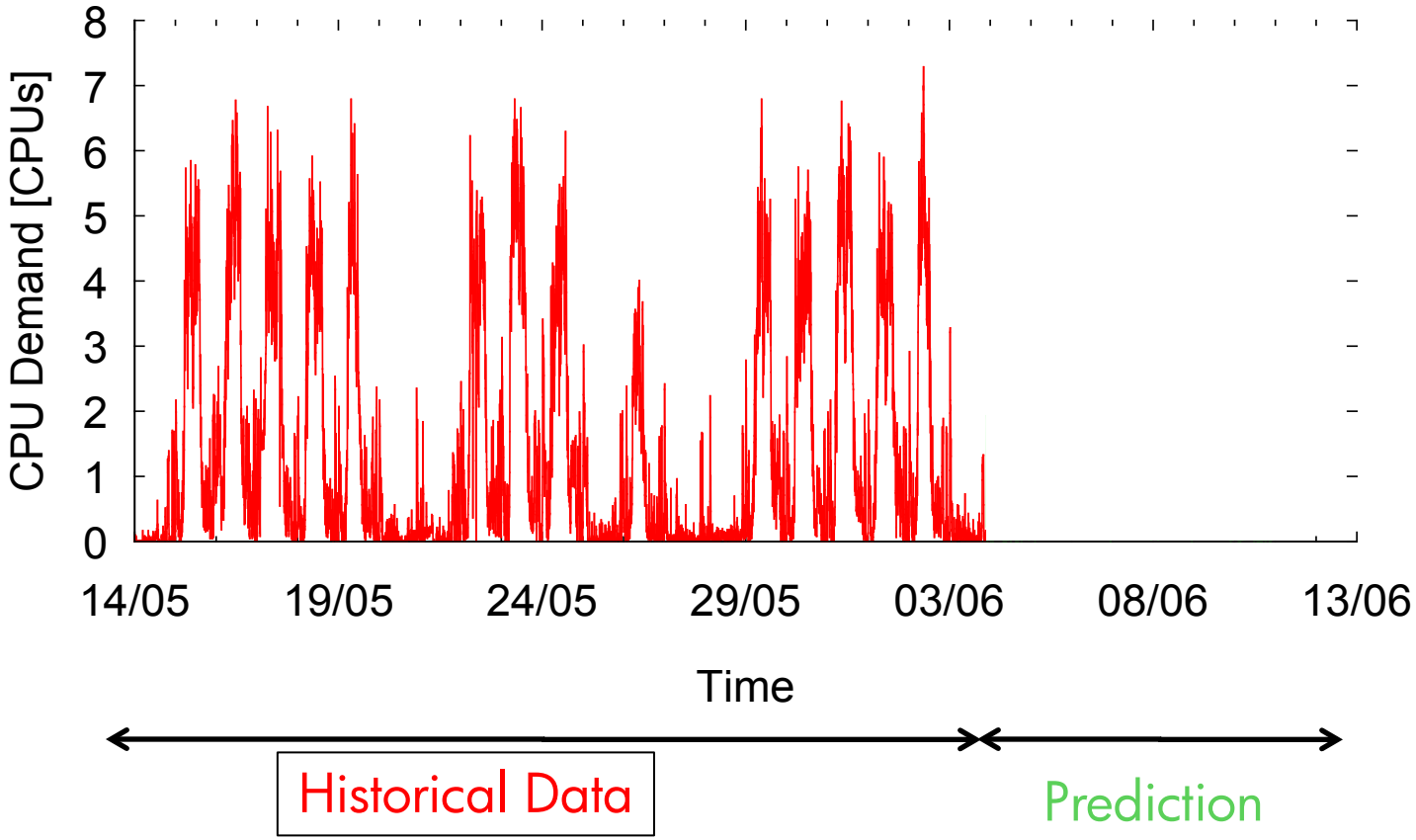
– Idea:

- Facilitate design of an optimal data center
- Consider future workload characteristics
- Emphasis on sustainability criteria and cost



Know Your Workloads

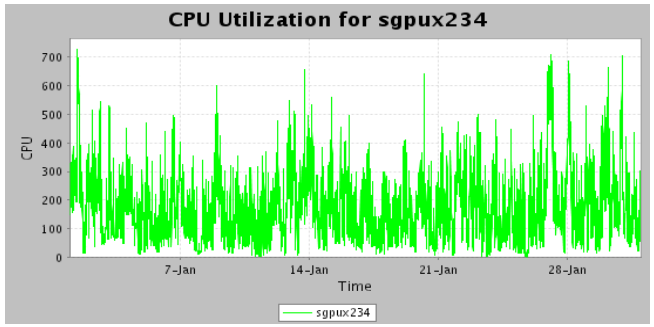
– Monitor, analyze, and predict resource demands of workloads



Capacity Planning—Size Your Data Center

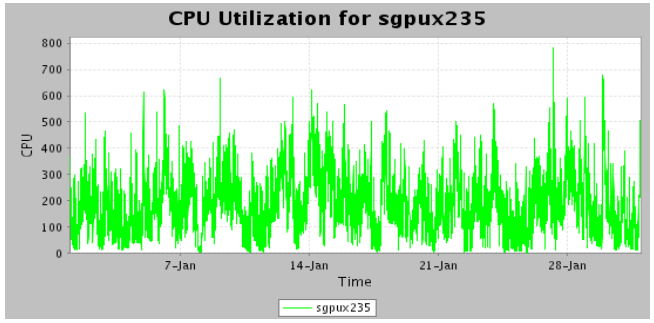
- Virtualize resources
- Consolidate workloads based on history
- Save resources through sharing: New math: 8 + 8 = 12

8 CPU Peak



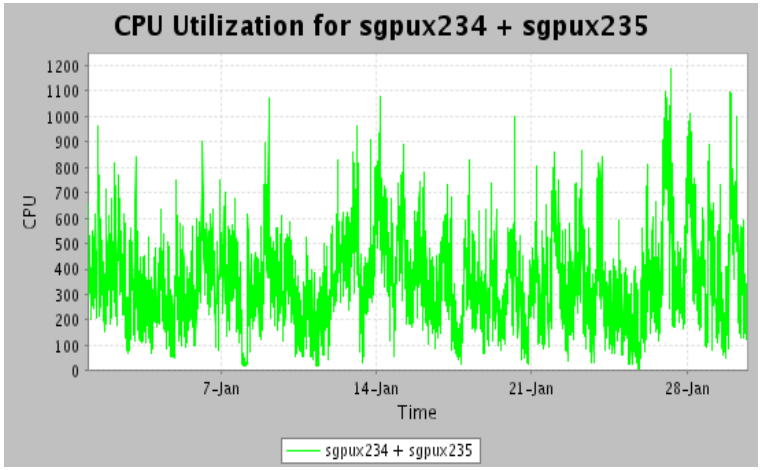
+

8 CPU Peak



=

12 CPU Peak



Topic #2: Sustainability Monitoring and Metric Development

Goal: Quantify service delivery impact on sustainability.

All Rights Reserved. No copies may be made, in full or part, without written permission of the authors.

Infrastructure Operational Cost

Sustainability through end to end design and management => Least Cost

$$Cost_{total} = \underbrace{\left(\frac{\$}{ft^2}\right)\left(A_{critical}, ft^2\right)}_{\text{Real Estate}} + \underbrace{\left(1 + K_1 + L_1 + K_2 L_1\right) U_{\$,grid} P_{\substack{consumed \\ hardware}}}_{\textbf{Burdened power consumption}} + \underbrace{R\left(M_{total} S_{avg} + IT_{dep} + \sigma_1\right)}_{\text{Personnel, equipment, SW per rack}}$$

J_1 : capacity utilization factor,

- ratio of maximum design (rated) power consumption to the actual data center power consumption

$K_1 = F(J_1)$: burdened power delivery factor,

- ratio of amortization and maintenance costs of the power delivery systems to the cost of grid power

$K_2 = F(J_1)$: burdened cooling cost factor,

- ratio of amortization and maintenance costs of the cooling equipment to the cost of grid power

Depreciation factors

L_1: cooling load factor,

- ratio of power consumed by cooling equipment to the power consumed by compute, storage and networking hardware

Patel and Shah, Cost Model for Planning, Development and Operation of a Data Center
http://www.hpl.hp.com/techreports/2005/HPL-2005-107R1.html

Infrastructure Operational Cost

Sustainability through end to end design and management => Least Cost

$$Cost_{total} = \underbrace{\left(\frac{\$}{ft^2}\right)\left(A_{critical}, ft^2\right)}_{\text{Real Estate}} + \underbrace{\left(1 + K_1 + L_1 + K_2 L_1\right) U_{\$,grid} P_{\substack{consumed \\ hardware}}}_{\textbf{Burdened power consumption}} + \underbrace{R\left(M_{total} S_{avg} + IT_{dep} + \sigma_1\right)}_{\text{Personnel, equipment, SW per rack}}$$

How much energy does a service consume?

How many resources are needed to meet service requirements?

How sustainable is a service?

Service Energy Estimation

Power Model

- *Server:* based on utilization

$$P_{server} = P_{idle} + (P_{busy} - P_{idle}) * (2u - u^r)$$

- *Networking:* use (chassis + line-card) power and port speed and utilization

$$P_{switch} = P_{chassis} + \Sigma_{cards} P_{card} + \Sigma_{rates} (num_ports_rate * power_port_rate)$$

- *Storage:* based on number of drives
- *Service:* derive from application workload

  Workload -->component resource usage --> component power --> service power

- *Facility (power delivery + cooling):* use Power Usage Effectiveness (PUE) or Coefficient of Performance (COP)

$$PUE = (IT\ Power + Facility\ Power) / (IT\ Power)$$

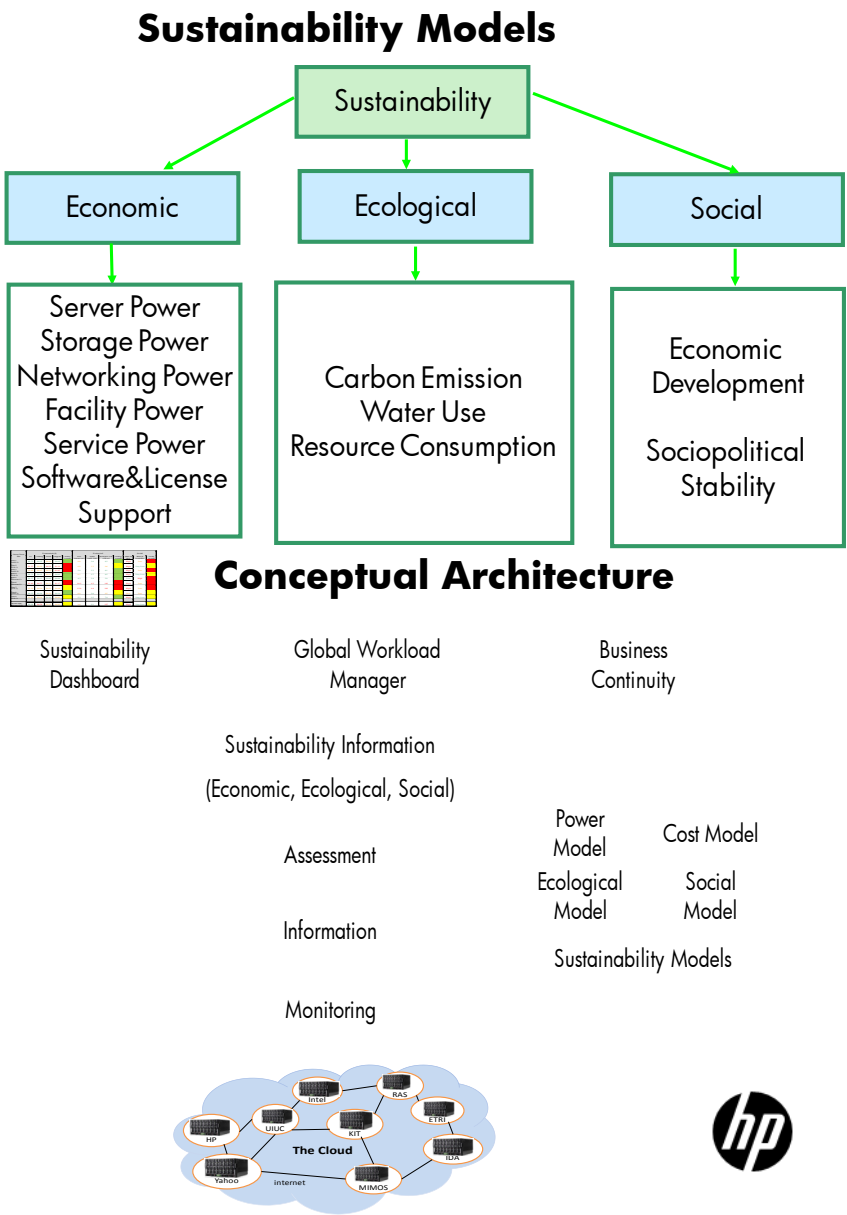
$$PUE = 1 + 1/COP_G + \%\ Power\ Delivery\ Losses$$

Service Sustainability

Cloud Sustainability Dashboard

<u>Quantify, assess and monitor Cloud sustainability and enable sustainability-aware management</u>

- Represent and model the overall sustainability of Clouds and Cloud Services
- Automated, real-time Cloud Sustainability Dashboard
- Enable sustainability-aware management tools



Cloud Sustainability Dashboard: A Case Study

Empirically evaluated the sustainability of Open Cirrus

- 10 sites distributed across North America, Europe, and Asia
- More than 2,300 computer servers (>15,000 CPU cores)
- Approximately 2.7TB main memory and over 5PB storage

An Illustrative Sustainability Dashboard (Infrastructure View)

Open Cirrus Sites	Economical ($)					Ecological				Social		
	IT	cooling	ntwk	support	econo. overall	CO_2 (tonnes-eq)	water (mill. Gal)	resource (GJ-eq)	ecolg overall	state of devt.	risk of instability	social overall
Site 1	$0.72	$0.35	$0.16	$0.43		6.0	2.6	83		High	Low	
Site 2	$1.27	$0.59	$0.21	$1.11		6.8	3.3	96		High	Very Low	
Site 3	$1.05	$0.47	$0.12	$1.07		5.9	2.3	81		High	Low	
Site 4	$0.75	$0.35	$0.12	$0.61		6.1	2.7	85		High	Very Low	
Site 5	$0.27	$0.13	$0.05	$0.09		4.3	2.4	59		Low	High	
Site 6	$1.82	$0.77	$0.11	$1.17		10.2	4.3	142		High	Low	
Site 7	$1.23	$0.54	$0.11	$0.98		15.0	4.4	192		High	Low	
Site 8	$0.55	$0.26	$0.10	$0.16		6.9	2.6	95		Med.	Low	
Site 9	$1.01	$0.44	$0.10	$0.83		5.3	2.5	74		High	Very Low	

Topic #3: Sustainable Service Management

Goal: Improve scheduling decisions.

All Rights Reserved. No copies may be made, in full or part, without written permission of the authors.

Integrated Data Center Management

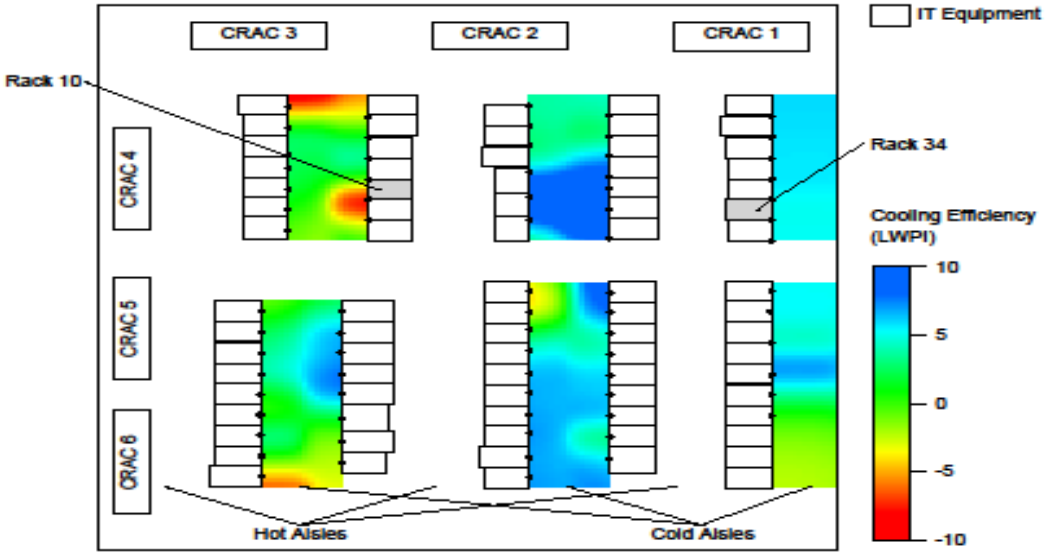
- Idea: Optimize Utilization of
  - Computing resources
  - Cooling resources
- Our Approach: Integrated Data Center Management
  - Dynamic computing resource allocation
  - Dynamic workload placement
    - Live migration of workloads
  - Cooling integration
    - Place workload on most cooling efficient servers
    - Adjust cooling resources to computing demand

Integrated IT-Facility Management
Policy-based Control

Experiment setup at HPL Palo Alto Demonstrator:

- 20 physical servers
  - 9 in Rack 10; 11 in Rack 34
- 35 Virtual Machines
  - 2 interactive 3-tier apps
  - 29 computational workloads
- 10 hour experiment
- Integrated controllers
  - Application Controller
  - Node Controller
  - Pod Controller
  - DSC Controller

Example: Real-time thermal-aware placement of virtualized IT load



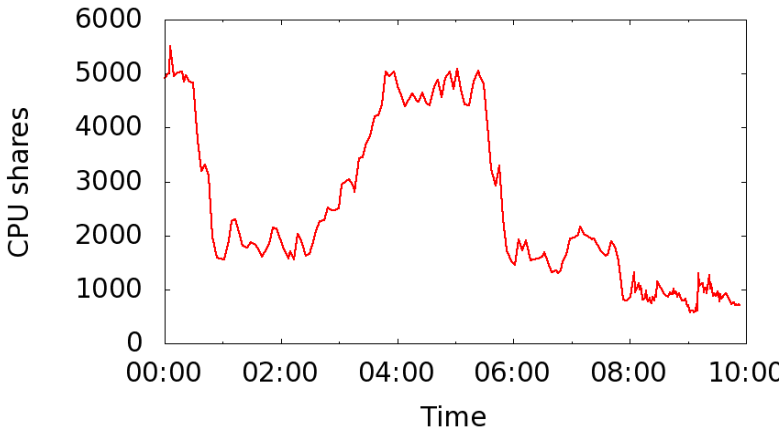
Savings: 26.5% IT power, 16.5% cooling power

EXTRACTION | MANUFACTURING | **OPERATION** | END OF LIFE

Y. Chen *et al.*, Integrated Management of Application Performance, Power and Cooling in Data Centers Proceedings of the 12th IEEE/IFIP Network Operations and Management Symposium (NOMS 2010), April 2010.

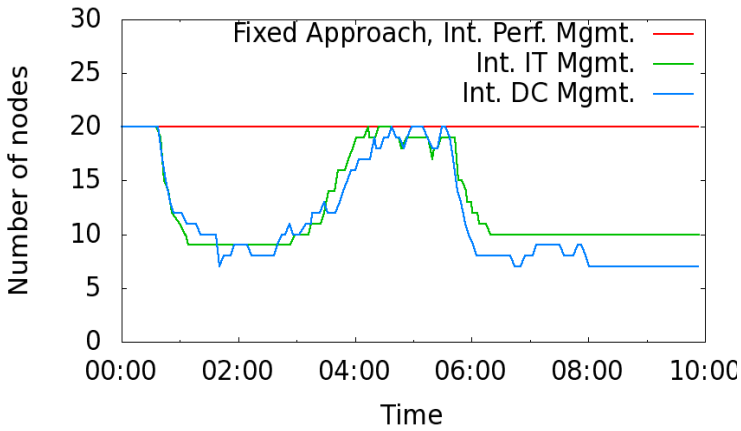
Integrated IT-Facility Management
Results

- Total CPU demand



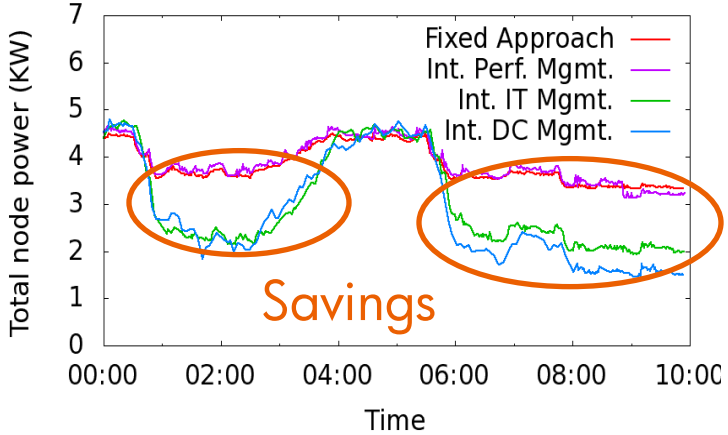
Workload demand

- Reducing computing resources



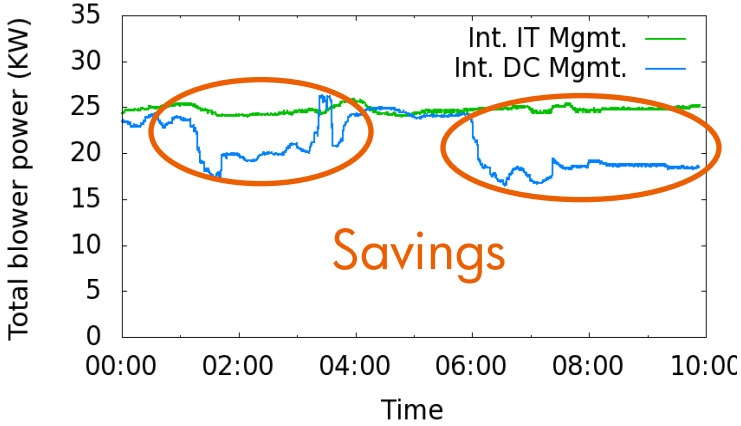
Number of running servers

- Reducing IT power



Server power

- Reducing cooling power

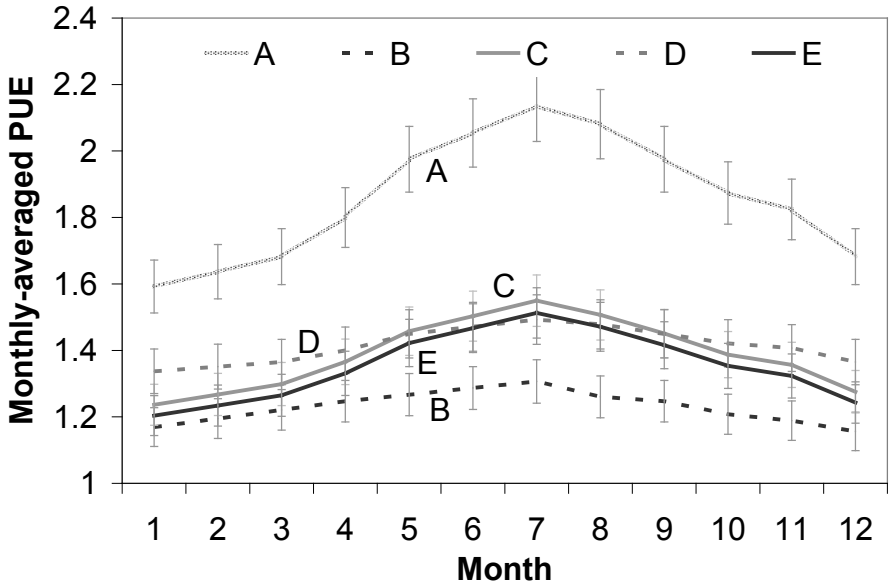


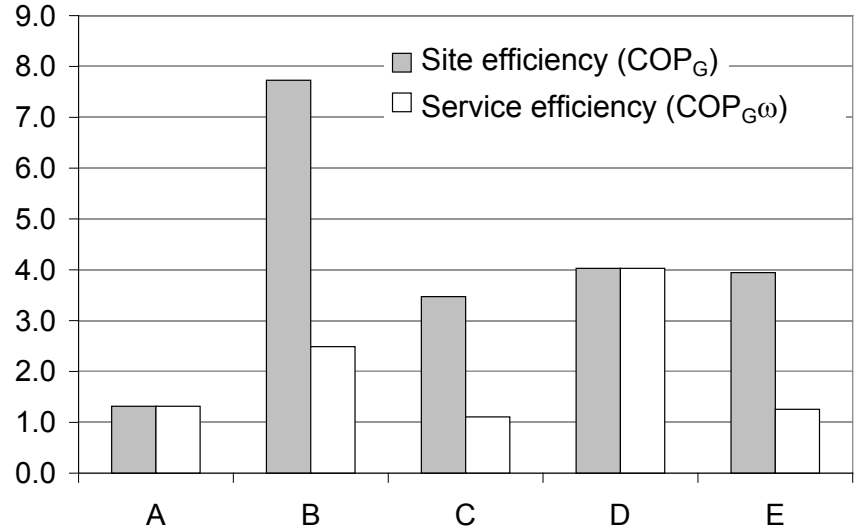
CRAC blower power

Geographic Workload Management

Thermal Management Considerations

A: typical enterprise data center (3 MW), in Houston, TX
B: best in class data center (3 MW), in UK
C: distributed data center (3 x 1 MW), in Houston, UK, Bangalore, India
D: containerized data center (30 x 100 kW), in Houston
E: distributed containerized data center (30 x 100 kW), in Houston, UK, Bangalore





Service efficiency is based on penalty function: $\omega = \sum_{i=1}^{n} d_i \dot{Q}_i / d_{avg} \sum_{i=1}^{n} \dot{Q}_i$

A. Shah et al., "Thermal Management Considerations For Geographically Distributed Computing Infrastructures", *IHTC-14*, Washington, DC, August 8-13, 2010.

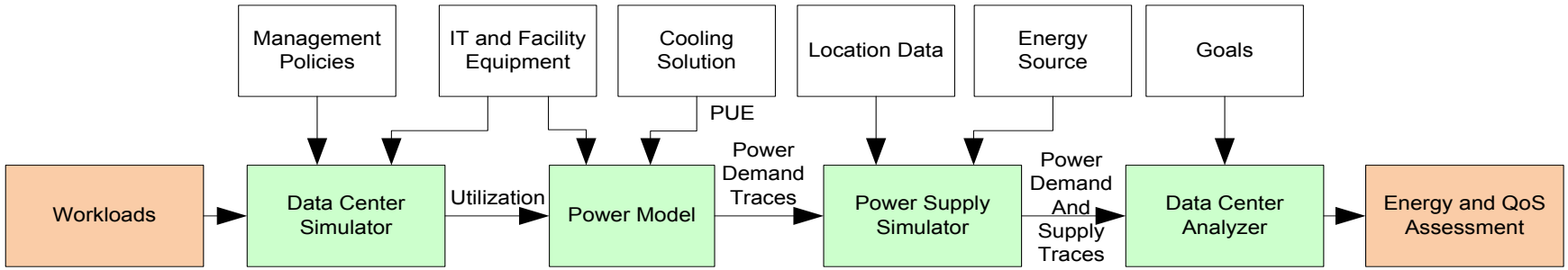
Topic #4: Matching Supply to Demand

Goal: Fulfill energy and resource demands from sustainable sources.

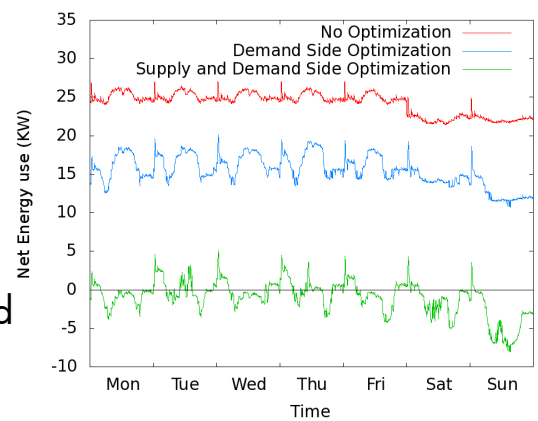
All Rights Reserved. No copies may be made, in full or part, without written permission of the authors.

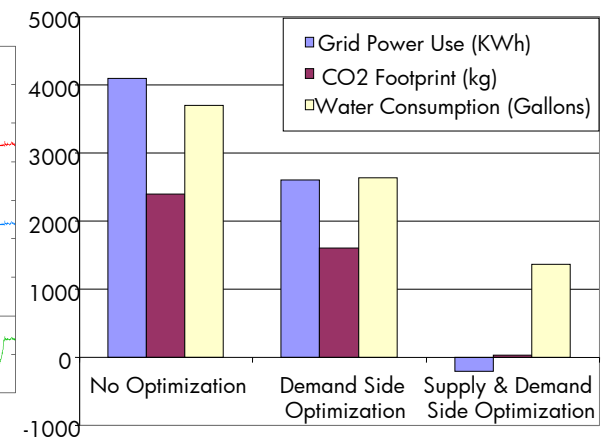
Integration of Power Supply with Demand

- Evaluate alternative IT and facilities management policies
  - Design: quantify the impact of alternative power supplies
  - Operation: consider different workload and power management policies via simulation
  - Take into account time-varying power demands and location-specific power supplies



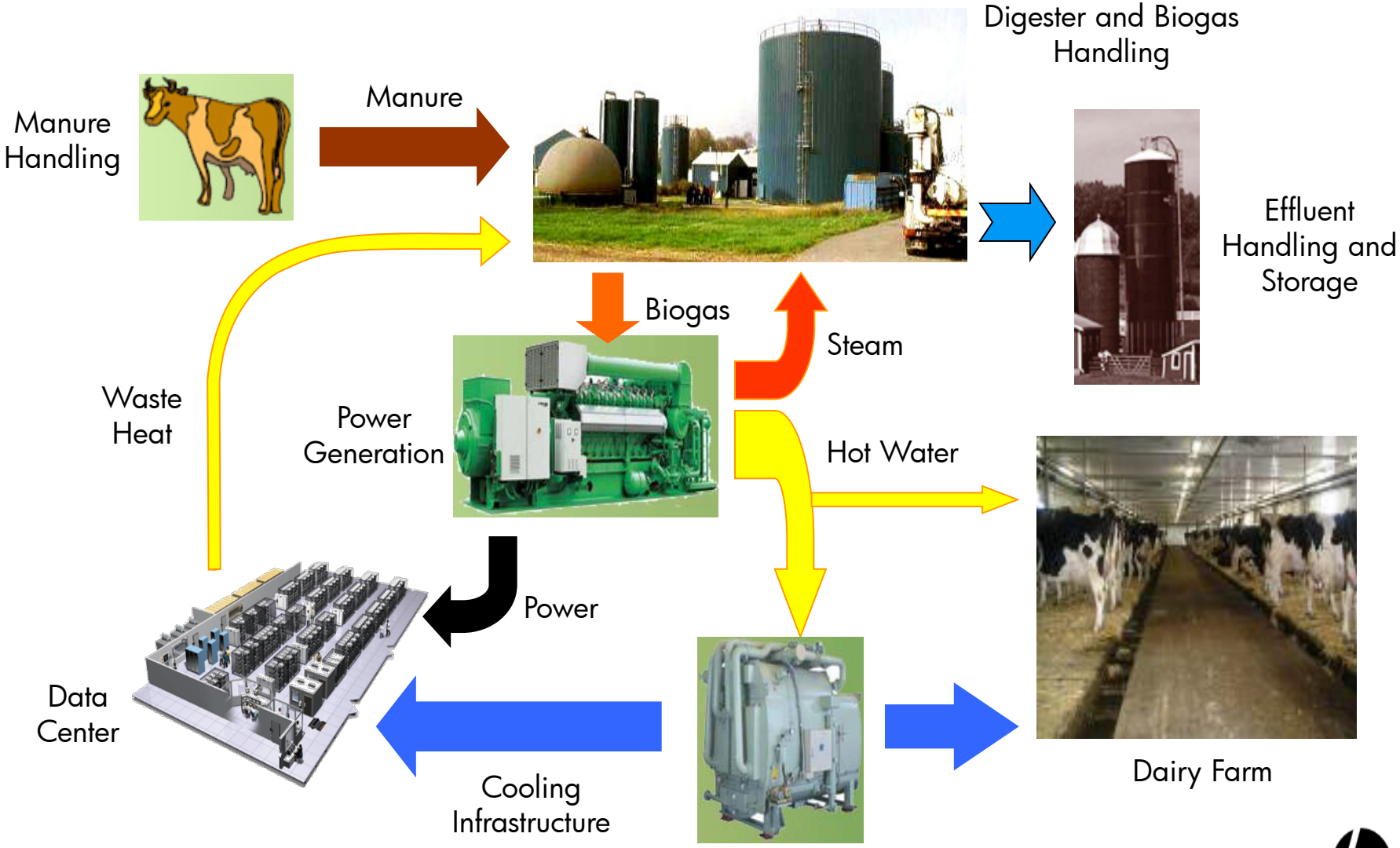
- Make recommendations
  - how much power is required
  - what mix of power sources is desirable
  - how power sources should be allocated





Power Supplies: Think Locally

Farm Waste and IT



Concluding Remarks

The Sustainable Cloud

1. Cloud computing will grow to meet the fundamental needs of society (more on this in the panel);
2. Despite the benefits of the cloud, there is a risk that this growth may result in increased resource consumption;
3. A Sustainable Cloud is key to meeting this need while reducing the impact of IT on the environment;

Q&A