ON THE THREE LAYERS OF SUSTAINABLE COMPUTING RESEARCH

Chao Li, Ph.D.
Professor, Dept. of CSE
Shanghai Jiao Tong University

Oct. 2022

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Intro.

李超，上海交通大学教授，博导
Ph.D. (University of Florida'14)

Research Area: Computer Architecture and System Design

Vice Chair, CCF Technical Committee on Computer Architecture
Lead Topical Editor, IEEE Transactions on Computers (TC)
Program Co-Chair, IEEE Int. Symp. on Reliable Distributed Systems

Sustainable Architectures and Infrastructure Laboratory

Department of Computer Science and Engineering
Shanghai Jiao Tong University (SJTU)

http://www.cs.sjtu.edu.cn/~lichao

The Technology Trend of Computing Infrastructure

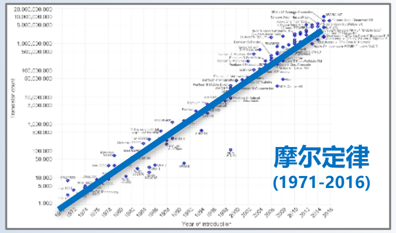
PC Era

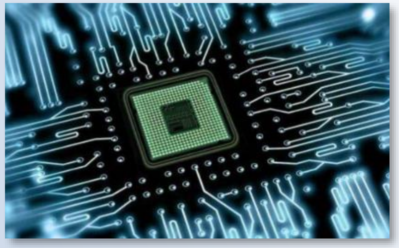
Internet Era

Cloud Era

AI+ Era

Support

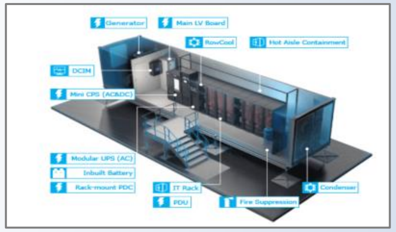




Core / Server

Server Cluster

Big Data Center



Cloud-Edge


Tsunami of Data： Could Consume 20% of
Global Electricity by 2025 …
climatechangenews.com

Contents

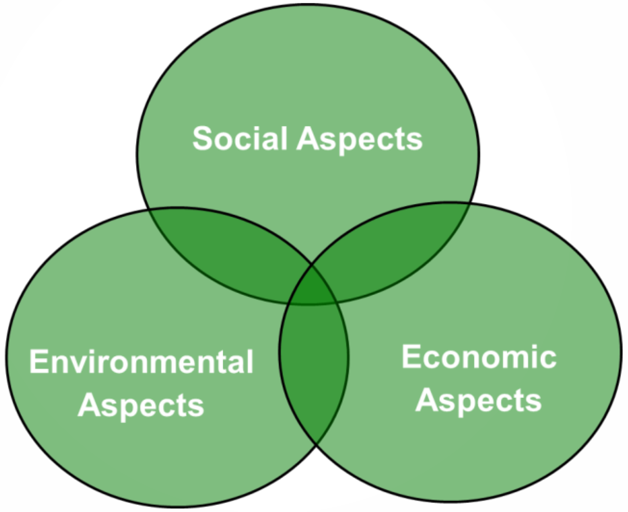
Different Interpretations of Sustainability

There are lots of discussions on the meanings of sustainability

P (population) × A (affluence) × T (technology)

The IPAT Equation (1970s)

A rough estimation of environmental impact



The Three Pillar (1980s)

We must balance three key elements in our design

1. People
2. Prosperity
3. Planet

Triple Bottom Line (1990s)

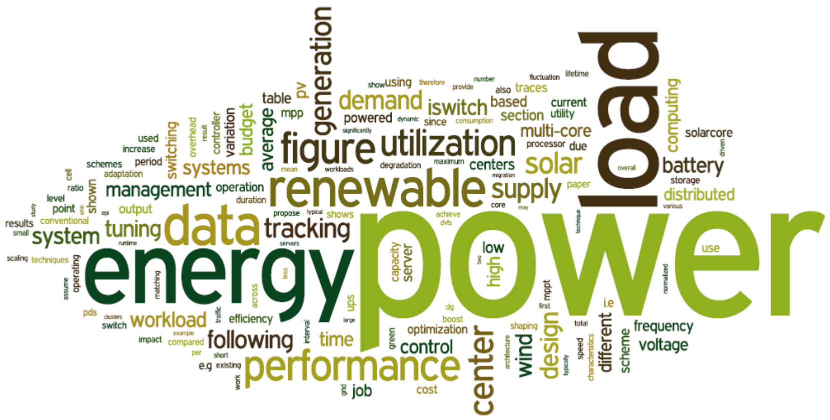
It is important to keep three key bottom lines in mind

When Computing Meets Sustainability

Sustainable Computing is not Computational Sustainability

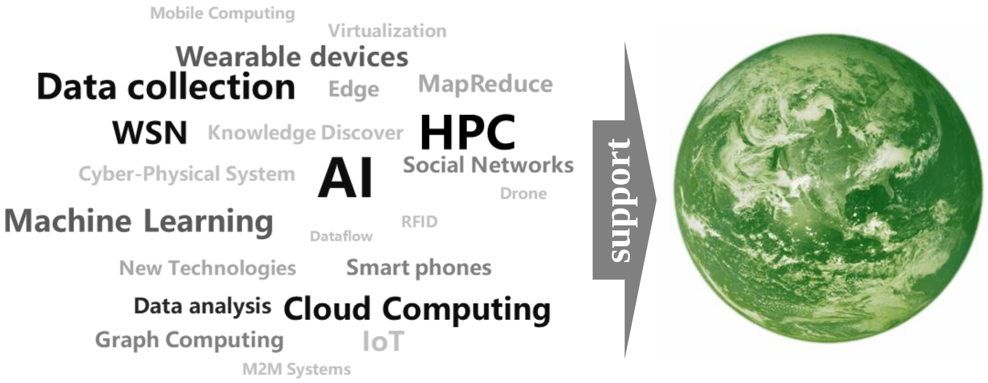
Sustainable Computing

-- developing energy-efficient, low-carbon computing systems and apps



Computational Sustainability

-- Leveraging computing techniques to improve the sustainability



Sustainable Computing is a Very Hot Topic

Sustainability has attracted great attention from researchers


POLICY FORUM

ENERGY

Recalibrating global data center energy-use estimates

Growth in energy use has slowed owing to efficiency gains that smart policies can help maintain in the near term

As demand for data centers rises, energy efficiency improvements to the IT devices and cooling systems they house can keep energy use in check.

Bottom-up analyses tend to best reflect this broad range of factors, generating the most credible historical and near-term energy-use estimates (7). Despite several recent national studies (8), the latest fully replicable bottom-up estimates of global data center energy use appeared nearly a decade ago. These estimates suggested that the worldwide energy use of data centers had grown from 153 terawatt-hours (TWh) in 2005 to between 203 and 273 TWh by 2010, totaling 1.1 to 1.5% of global electricity use (9).

Since 2010, however, the data center landscape has changed dramatically (see the first figure). By 2018, global data center workloads and compute instances had increased more than sixfold, whereas data center internet protocol (IP) traffic had increased by more than 10-fold (1). Data center storage capacity has also grown rapidly, increasing by an estimated factor of 25 over the same time period (1, 8). There has been a tendency among analysts to use such service demand trends to simply extrapolate earlier bottom-up energy values, leading to unreliable predictions of current and future global data center energy use (3–5). They might, for example, scale up previous bot-

"…new technologies is needed to manage future energy demand growth … once current efficiency trends reach their feasible limits" - ***Science, 2020***

"…the computing community should embrace a grand challenge to reduce the carbon-emissions and environmental impact of computing in absolute terms" - ***CACM, 2022***


COMMUNICATIONS OF THE ACM

HOME | CURRENT ISSUE | NEWS | BLOGS | OPINION | RES

Home / Magazine Archive / October 2022 (Vol. 65, No. 10) / Computing's Grand Challenge for Sustainal

EDITOR'S LETTER

Computing's Grand Challenge for Sustainability

By Andrew A. Chien

Communications of the ACM, October 2022, Vol. 65 No. 10, Page 5

10.1145/3559163

Comments

VIEW AS: SHARE:

There are Many Research Opportunities

SJTU

Both NSFC and NSF today put an emphasis on sustainability

NSFC of China

-- Sustainable Development International Cooperation Program 2022

国家自然科学基金委员会
National Natural Science Foundation of China

Sustainable Development International Cooperation Program 2022

Call for Proposals

日期 2022-04-25 来源: 作者: 【大 中 小】 【打印】 【关闭】

I Background

A systematic understanding of the synergies and conflicts among Sustainable Development Goals (SDGs) is of great significance for balancing the economic, social, and environmental dimensions of sustainable development and for the ultimate realization of the SDGs. With the theme of "human-environment system dynamics and UN' s SDGs", NSFC and international partners launch the program on Sustainable Development International Cooperation (SDIC) to contribute to the global effort for a sustainable future.

NSF of USA

-- Dear Colleague Letter: Design for Sustainability in Computing

NATIONAL SCIENCE FOUNDATION
2415 EISENHOWER AVENUE
ALEXANDRIA, VIRGINIA 22314

NSF 22-060

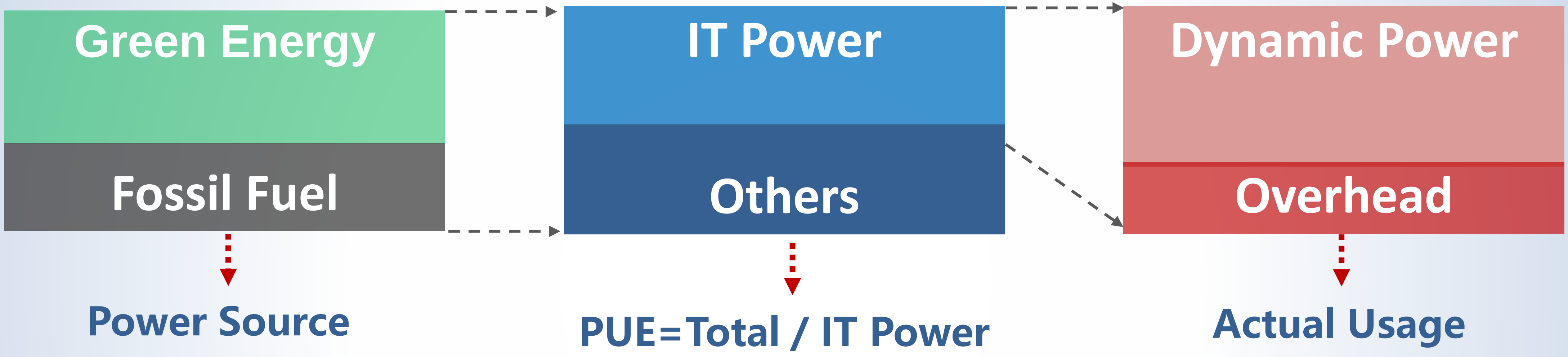
Dear Colleague Letter: Design for Sustainability in Computing

March 15, 2022

Dear Colleagues:

Environmental impacts of computing technologies extend well beyond their energy consumption and require a holistic focus on broader sustainability. Negative impacts of greenhouse gas emissions, depletion of rare earth elements, and e-waste are exacerbated by the proliferation of computing throughout society and treatment of computing systems as disposable commodities with planned obsolescence. Furthermore, environmental concerns

Related Research (1): Defining New Metrics



Low-PUE Does Not Guarantee Sustainability, we need to re-design the metric to take into account other factors. For example:

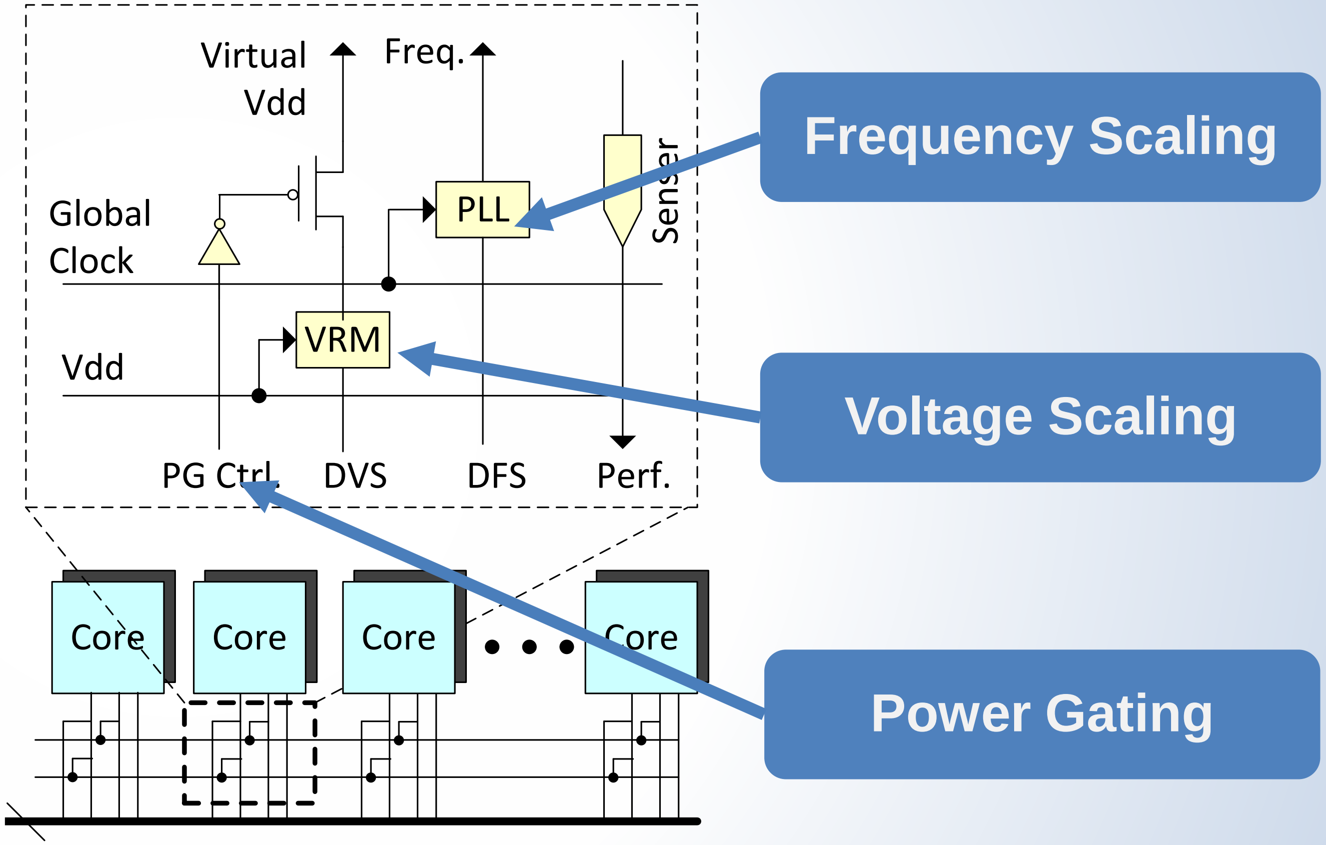
$$\frac{E_{green}}{E_{total}} \times \frac{E_{IT\_d}}{E_{total}}$$

By defining appropriate metrics, we can better design system

Related Research (2): Tuning the Hardware

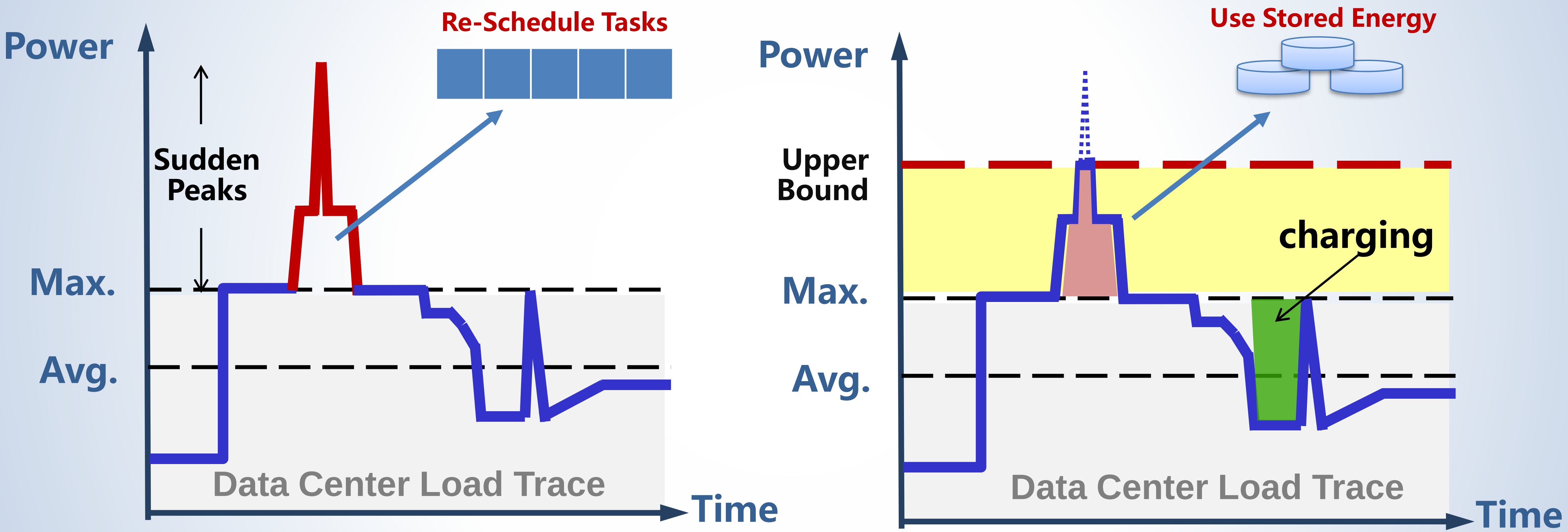
ns → ms

- **C0**: Keep Running
- **C1**: Rest Allowed
- **C2**: A Long Rest
- **C3**: Deep Sleep



By enabling hardware management, we can save CPU energy

Related Research (3): Scheduling the Workload



By enabling hardware management, we can save CPU energy

Related Research (4): Tapping into Renewables

By innovating the infrastructure, we can greatly reduce carbon

How to Better Understand Sustainable Computing?

How can we classify different works on sustainable computing?

- **Topic 1. Defining New Metrics**
- **Topic 2. Tuning the Hardware**
- **Topic 3. Scheduling the Workload**
- **Topic 4. Tapping into Renewables**

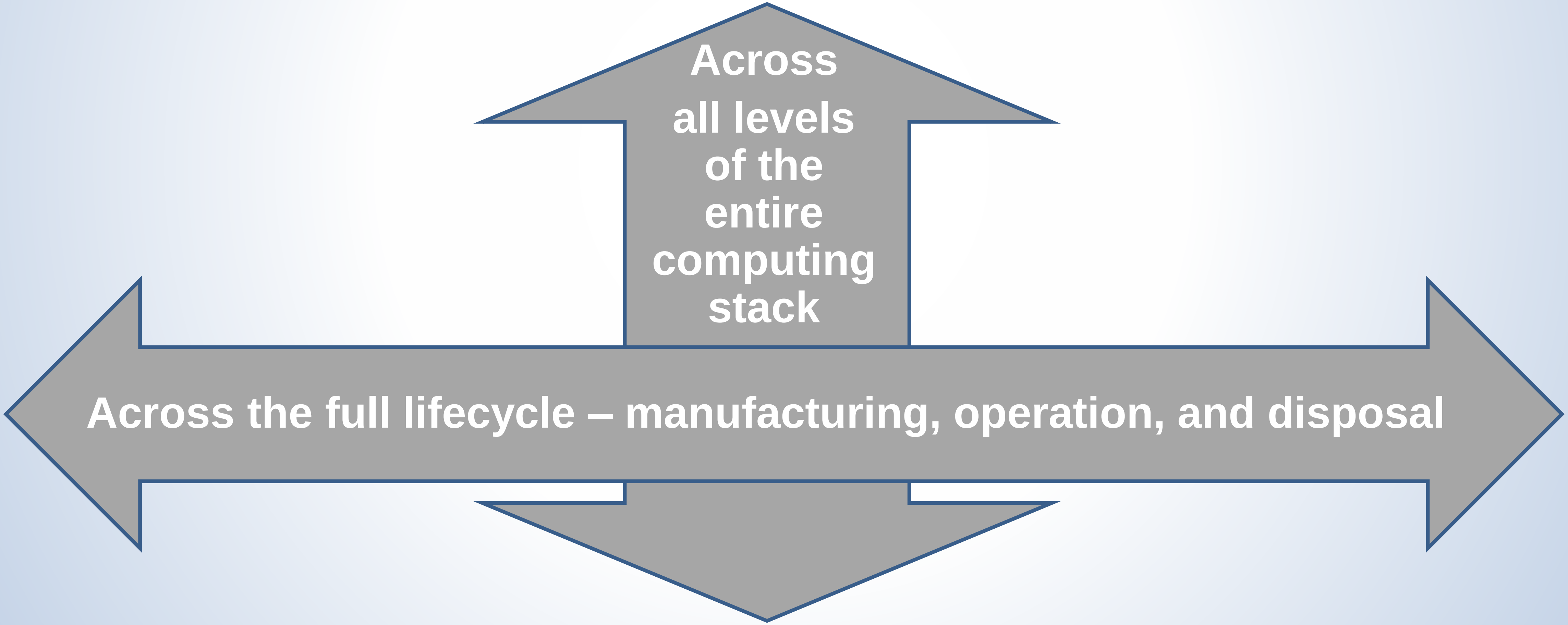
Key Questions:

Are they similar?
Are they different?
What have been done?
What haven't been done?
Are there any opportunities?

Contents

Sustainable Computing Research: Simple Classification

Sustainable computing research has two major dimensions



A Deeper Look at Sustainable Computing Research

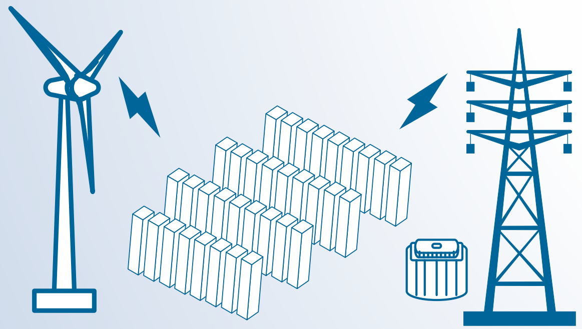
There are three almost-orthogonal tiers

- **Design for Sustainability**
  - **3. Enhancing Carbon Awareness**
    - e.g., Agile Service Provisioning
    - e.g., Indirect Energy Effects of IT
  - **2. Improving Carbon Efficiency**
    - e.g., Optimizing Resource Usage
    - e.g., Minimizing System Overhead
  - **1. Reducing Carbon Footprint**
    - e.g., Reduce Power/Energy Demand
    - e.g., Tapping into Renewables
    - e.g., Reducing Embodied Carbon

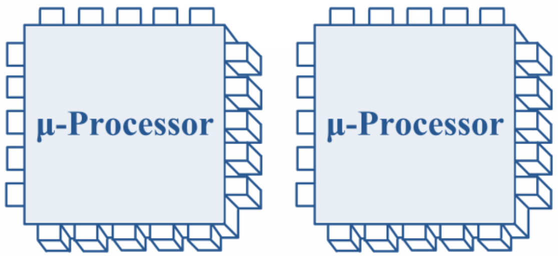
Tier-1. Reducing Carbon Footprint

- **Design levels: Hardware/Equipment Level**
- **Key objective: Minimizing the Power/Energy Usage and Carbon Footprint**
  - 1) Reducing the reliance on fossil fuels and “brown” energy
  - 2) Reducing the energy/power consumption of the hardware device
  - 3) Taking embodied carbon footprint into account and recycle hardware

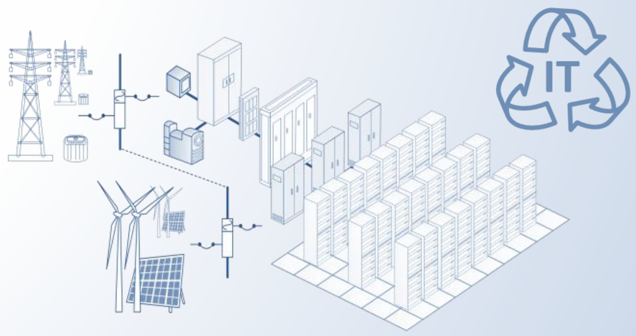
RES Integration



Low-Power Design



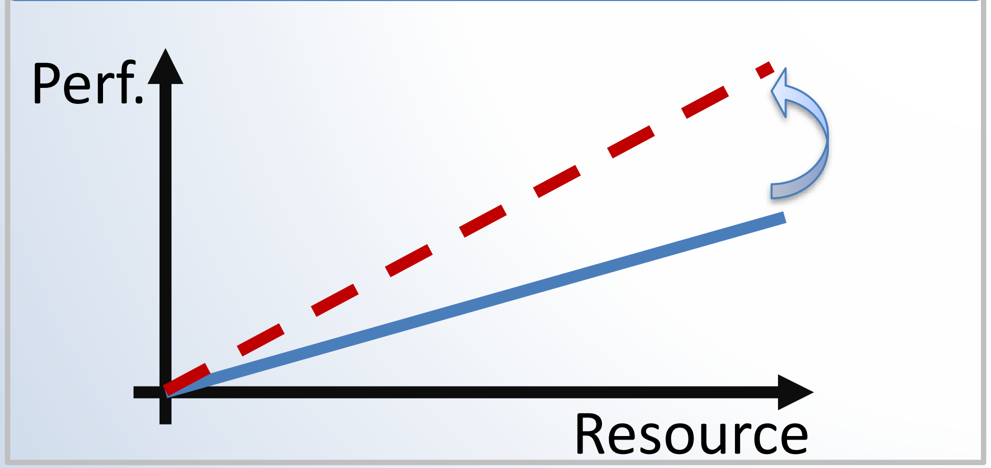
Embodied Carbon



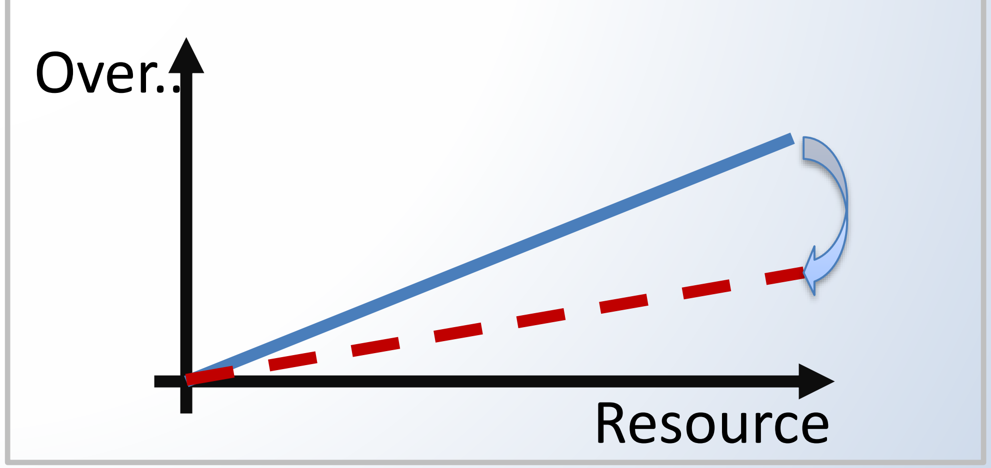
Tier-2. Improve Carbon Efficiency

- **Design levels: Architecture/System Level**
- **Key objective: Optimizing Perf. under the Given Energy/Carbon Budget**
  - 1) Make the best use of any computing resources (CPU、RAM、Network, etc.)
    - Re-designing algorithm and system architecture to improve resource utilization/sharing
  - 2) Eliminate any unnecessary system operations or resource overhead
    - Keep data preprocessing, network re-configuration, load migration simple and efficient

Boost Performance



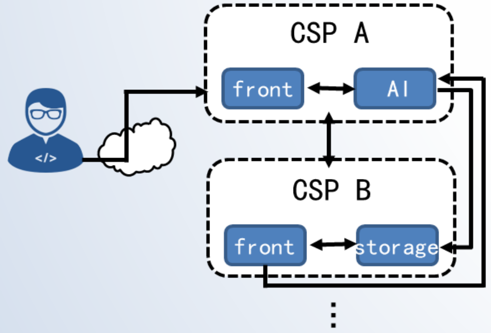
Reduce Overhead



Tier-3. Enhancing Carbon Awareness

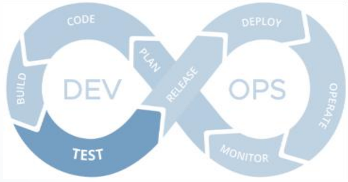
- **Design levels: Application/Service Level**
- **Key objective: Minimizing environmental impact of IT services**
  - Focusing on agile service provisioning to improve efficiency
    - If possible, upgrade rather than replace, unify rather than divide
  - Dealing with indirect carbon caused by the “rebound effect”
    - The carbon reduction achieved can be compromised by many other things

Agile Service Provisioning



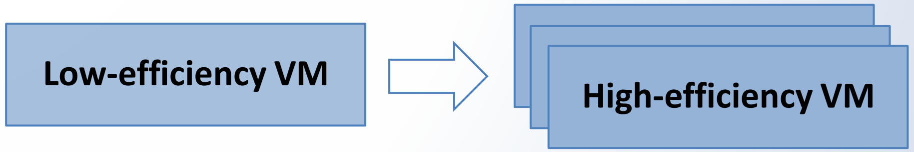
cloud federation and inter-cloud

Software-defined

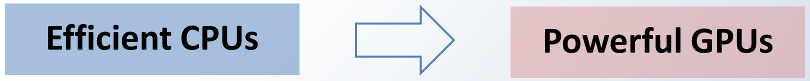


DevOps+AIOps

Rebound Effect Aware



Direct rebound: more consumption of the services



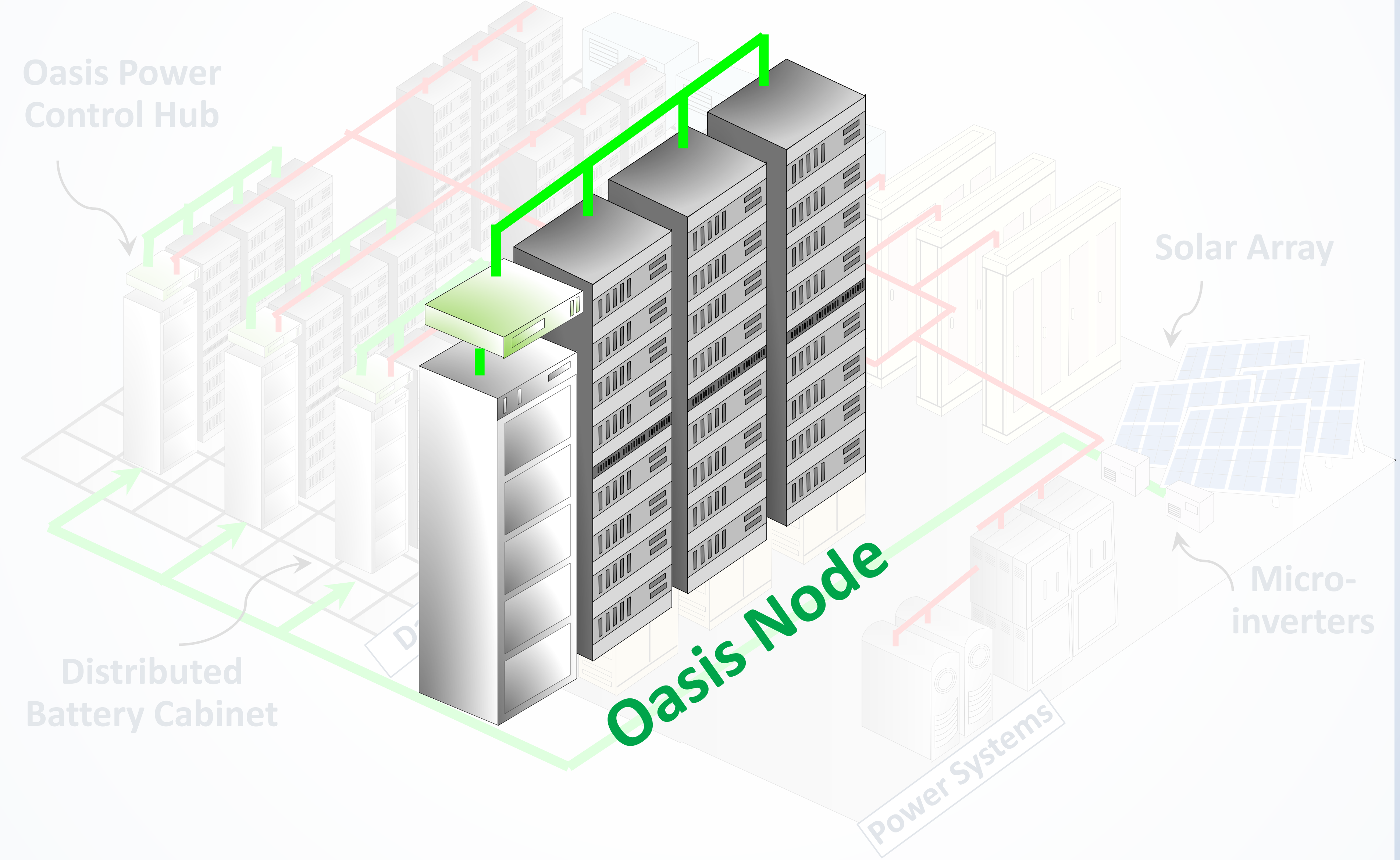
Indirect rebound: encourage the usage of other goods

Many Opportunities of Sustainable Computing

3. Carbon Awareness	**Application/Service Level**	Data-driven, AI-assisted design for continuous service optimization
2. Carbon Efficiency	**Architecture/System Level**	Fine-grained resource management in the “killer microseconds” era
1. Carbon Footprint	**Hardware/Equipment Level**	Unconventional power provisioning and energy storage strategies

We need a cross-layer optimization approach

OASIS: A Modular Green Data Center



OASIS: A Modular Green Data Center

Oasis Node

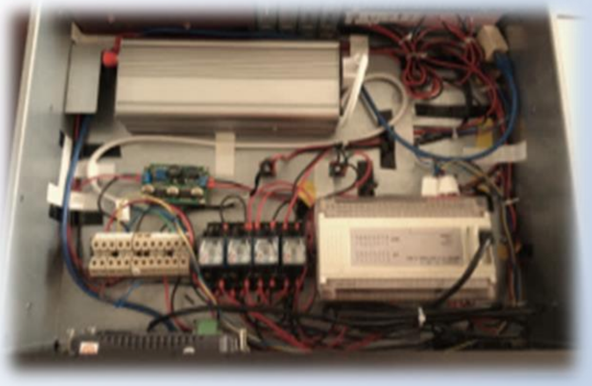
Solar Modules
- Grape Solar 1.6 KW PV panels

Server Nodes
- AMD low-power node cluster

Power Mgmt. Agent
- AMD low power node

Power Ctrl. Hub
- PLC + HMI + Inverter …

Battery Chassis
- Nine 2Ah lead acid batteries



Chao Li, et al. "Enabling Datacenter Servers to Scale Out Economically and Sustainably". MICRO 2013

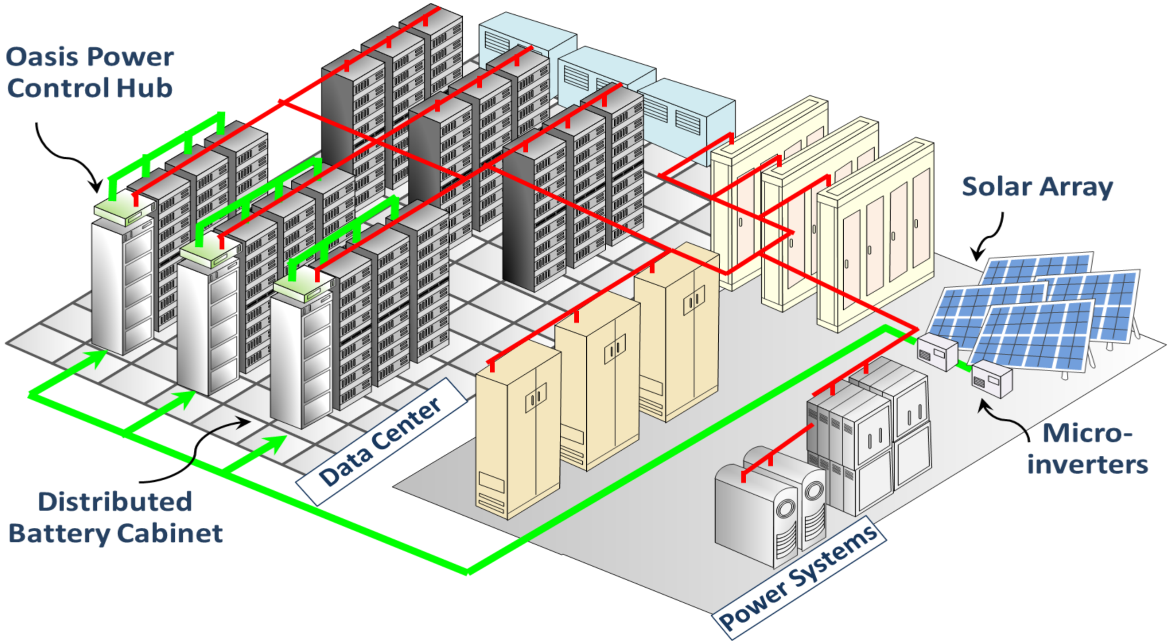
Go Green-1: Our Work at the Hardware/Equipment Level

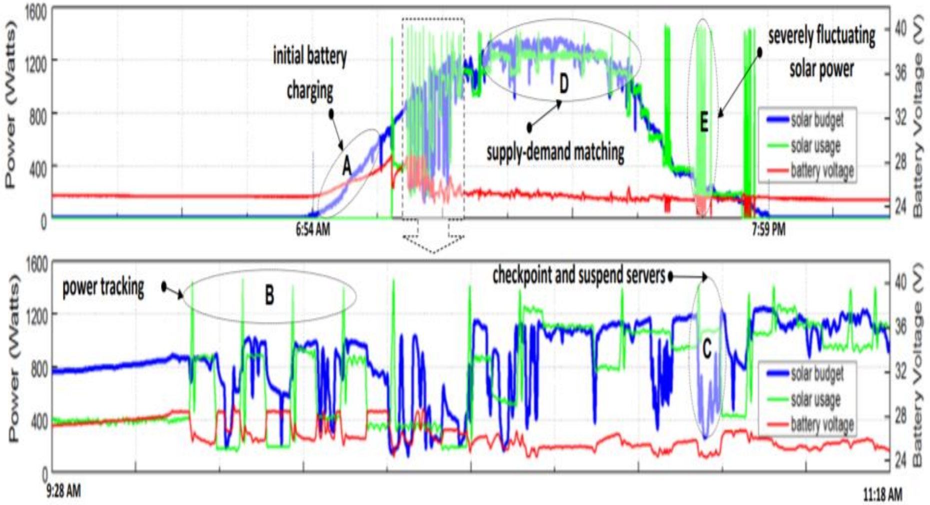
Background & Problem

Edge computing data center cannot handle time-varying renewable power supply

idea Allowing servers to adapt to renewable power

novelty One of the first green edge data center prototype





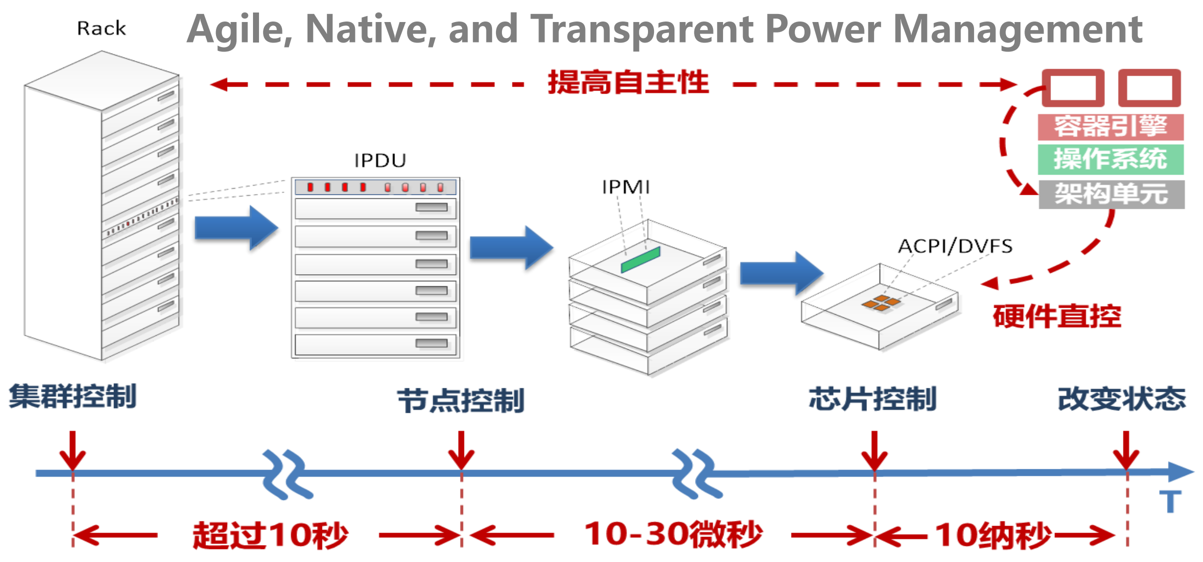
Towards Sustainable in-Situ Server Systems in the Big Data Era, ISCA 2015

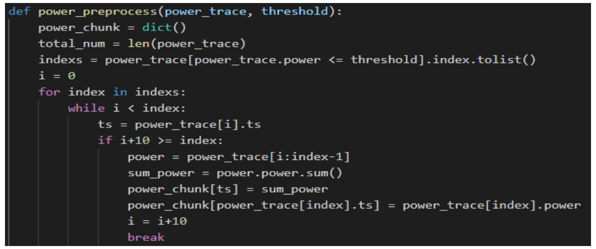
Go Green-2: Our Work at the Architecture/System Level

Background & Problem

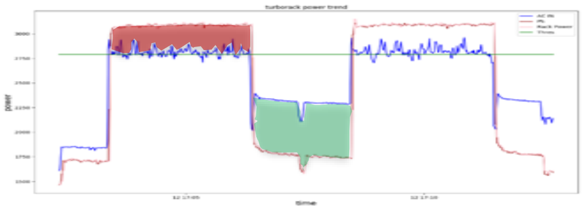
Data center incurs high load power peaks which is costly and less sustainable

idea	Re-configure data center power control hierarchy
novelty	A super-fast data center power capping strategy

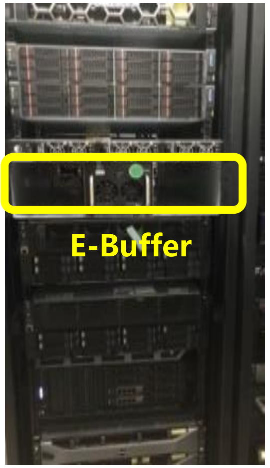




```
def power_preprocess(power_trace, threshold):
    power_chunk = dict()
    total_num = len(power_trace)
    indexs = power_trace[power_trace.power <= threshold].index.tolist()
    i = 0
    for index in indexs:
        while i < index:
            ts = power_trace[i].ts
            if i+10 >= index:
                power = power_trace[i:index-1]
                sum_power = power.power.sum()
                power_chunk[ts] = sum_power
                power_chunk[power_trace[index].ts] = power_trace[index].power
                i = i+10
                break
```



Peak Power Tracking



Prototype

ANT-Man: Towards Agile Power Management in the Microservice Era. SC 2020

Go Green-3: Our Work at the Application/Service Level

Background & Problem

Prior graph computing framework cannot efficiently utilize unified memory

idea	Re-design data structure to reduce overhead
novelty	The first UM-aware graph computing framework

Graph Data → **Prepare** → **Update** → **Generate** → Processing

```
void Kernel(){
  for vtx in frontier{  // Assign each active vertex
      to one warp
    for edge in vtx.edgeSet{ //Threads in the same
        warp process edges in parallel
      Update();  // Update labels
      Generate(); // Enqueue vertices based on
          criteria
}}}
bool CheckConverge(){
    if (frontier.size()==0) return true;
    return false;}
void Main(){
  Init();
  SetMemPolicy();    //Set adaptive UM policy for UM
      prefetching and hints
  while(!CheckConverge()){
    Prepare();
    Kernel();
    frontier.getNew();    //BDF gets the new frontier
        and reset the bitmap.
  }}
```

Grus: Toward Unified-Memory-Efficient, High-Performance Graph Processing on GPU. TACO. 2021

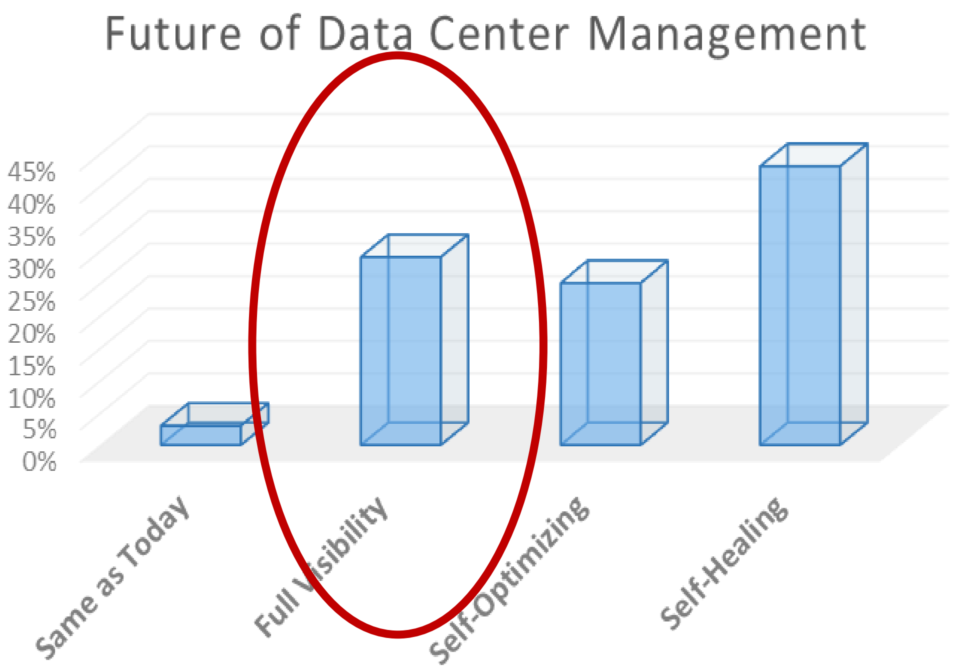
Contents

We Needs to Better Understand the Cloud Stack

Green cloud computing demands collaborative design

We need to know our infrastructure

- 《Data Center 2025：Exploring the Possibilities》



Key Challenges of Managing a Complex Data Center

How to Handle the Complexity?

Design Challenges

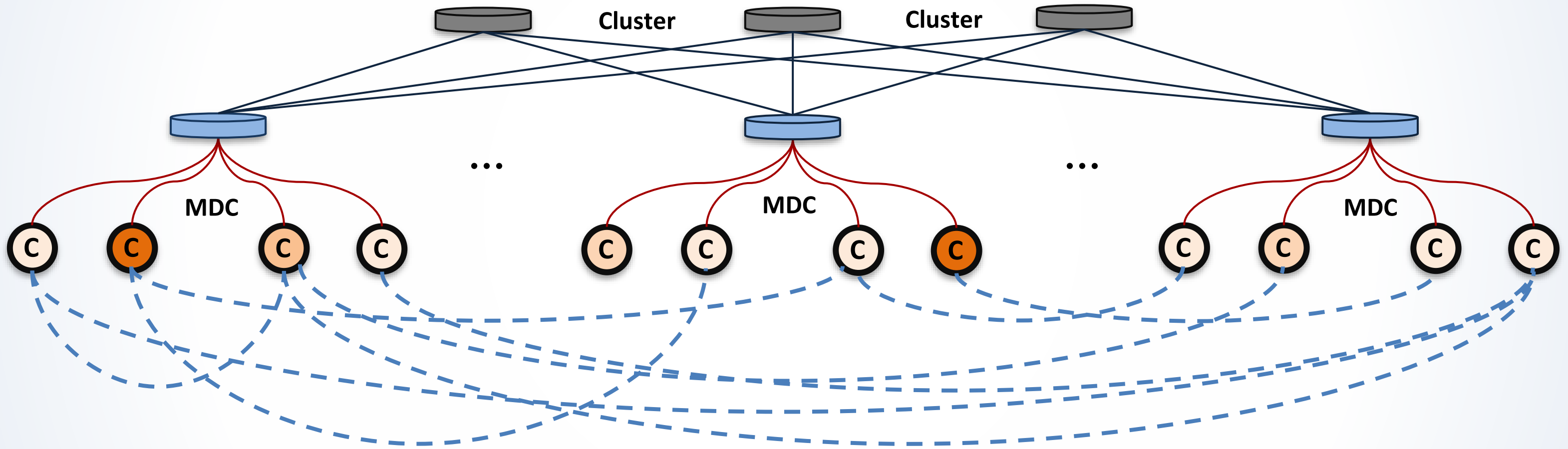
1. **Module-Level Heterogeneity**
   not all modules are equal
   — **"WHAT"**

2. **Cross-Layer Coordination**
   there are semantic gaps
   — **"WHICH"**

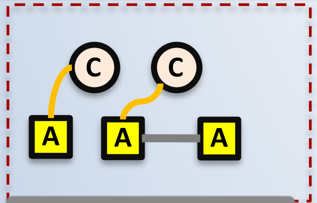
3. **Spatial/Temporal Variability**
   resources keep changing
   — **"HOW"**

Future Direction-1: Graph-Based Hardware Abstraction

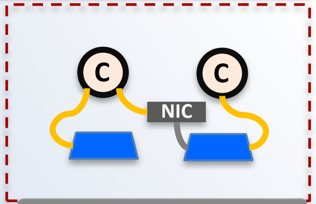
We need an abstraction/representation of the underlying resources



Hidden Relationship

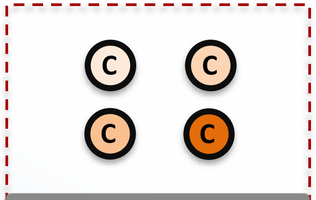


Interconnect

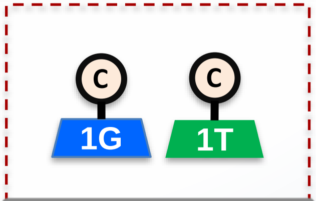


Mem Access

Runtime Properties

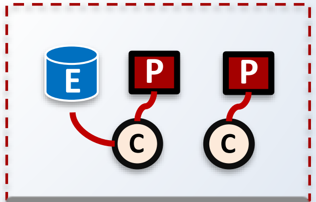


Core Variation

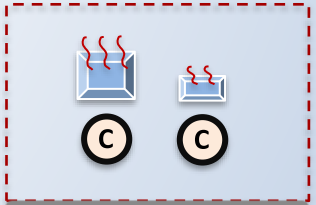


Capacity Issue

Indirect Interaction



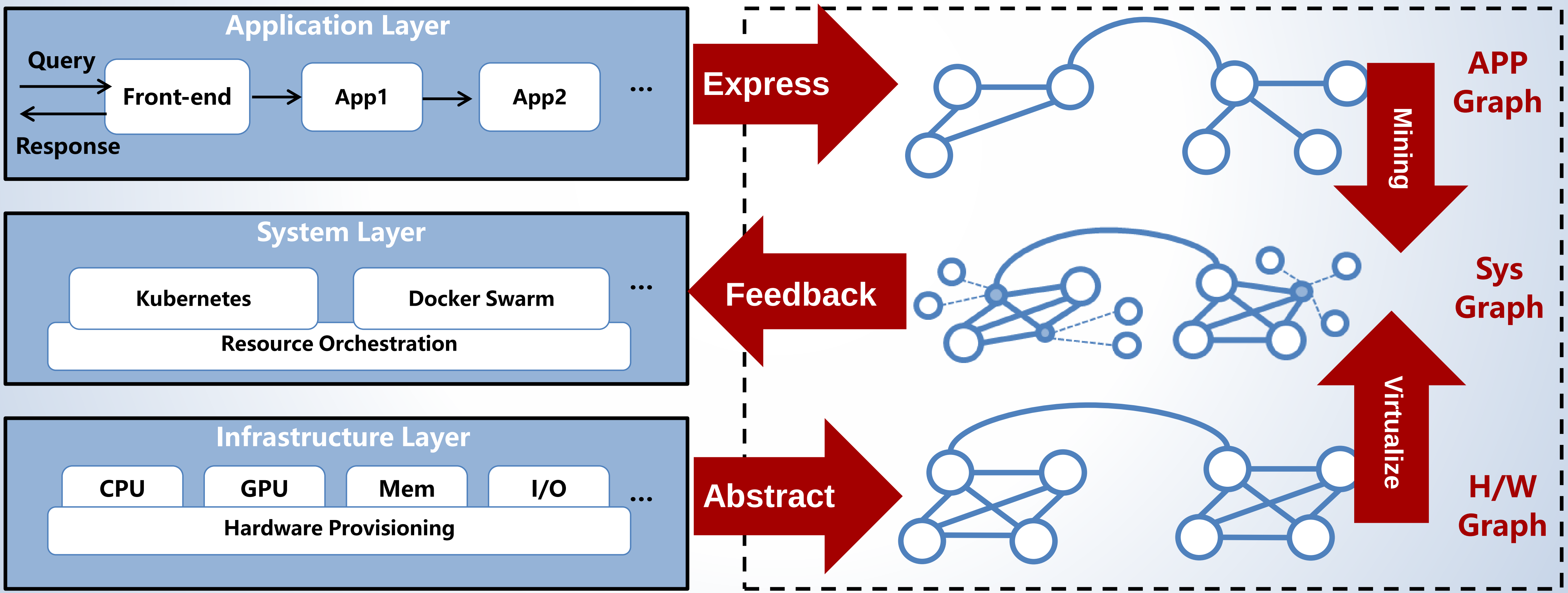
Power budget



Thermal Issue

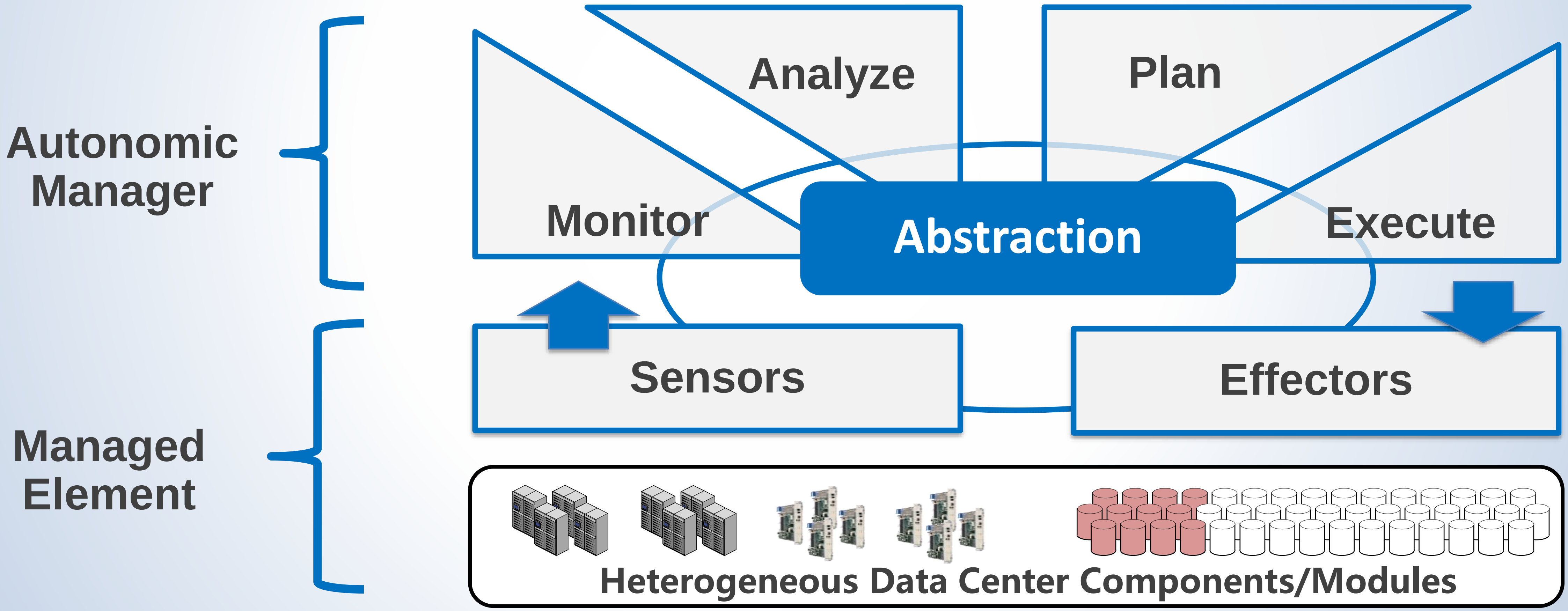
Future Direction-2: Data-Driven Resource Management

We need a data-driven resource management across the full stack



Future Direction-3. Self-Managing Control Framework

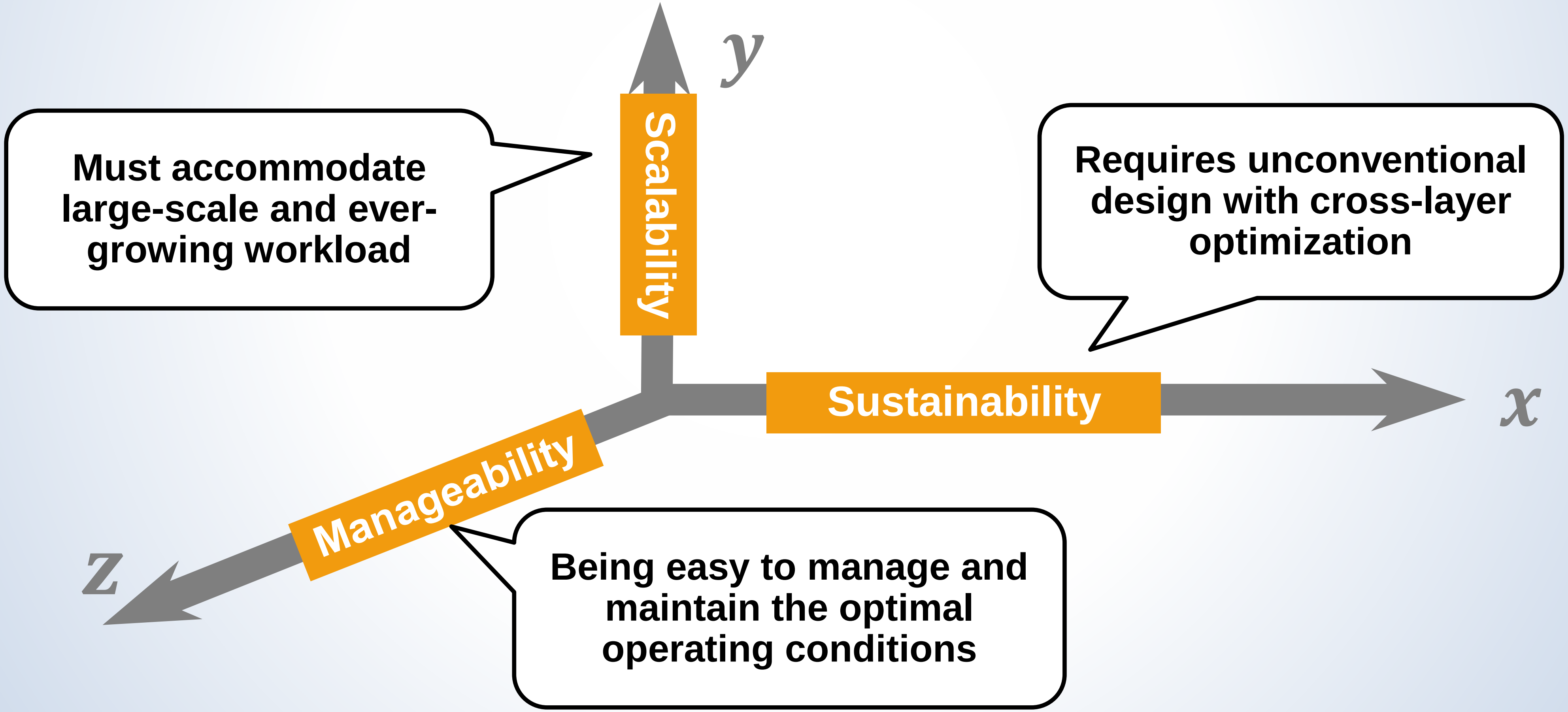
We need to enable the system to manage itself in the long run



Contents

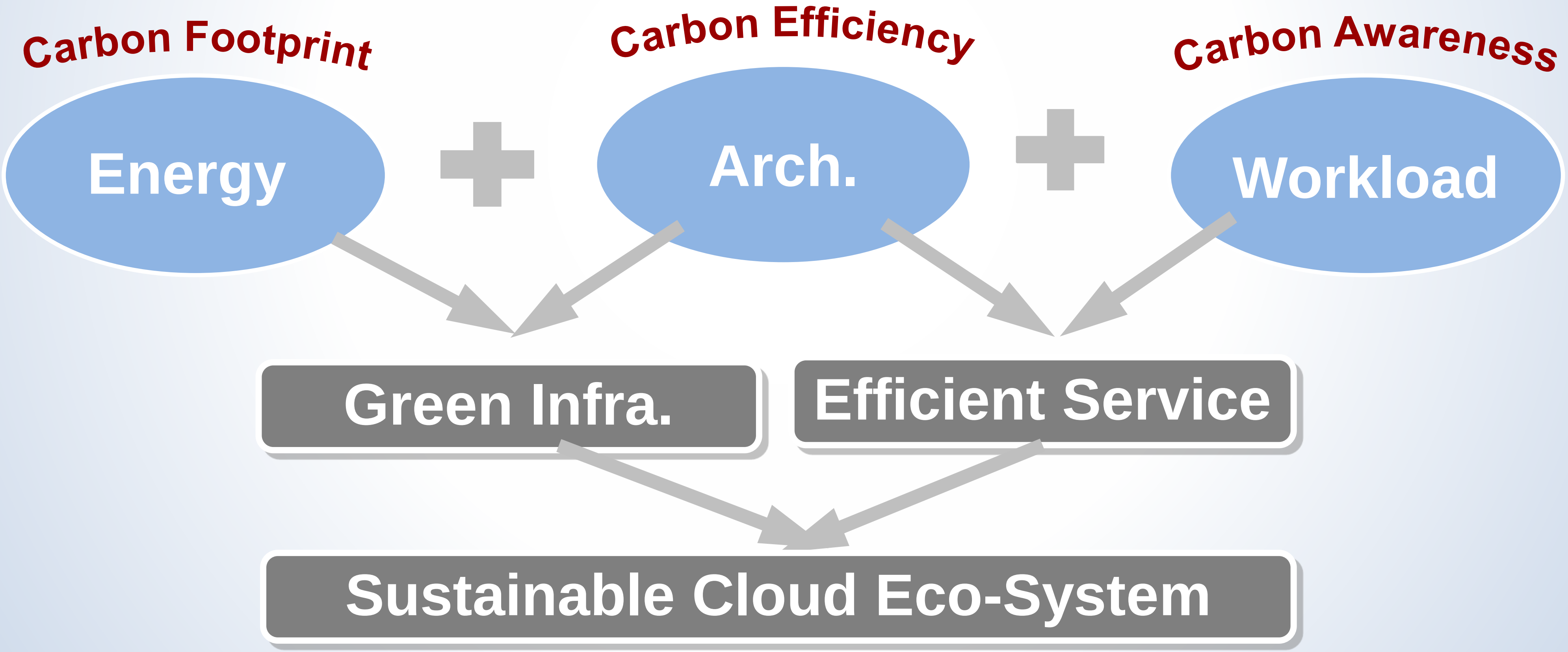
The "Sustainability-Scalability-Manageability" Challenge

It can be very challenging to build a highly scalable and sustainable computing system that is also easy to manage



Requires “Energy-Architecture-Workload” Co-Exploration

The Cloud should be designed for sustainability at three layers, combining the optimization of energy, architecture, and workload



Thanks!

Contact: CHAOL@SJTU.EDU.CN