Power Provisioning for Diverse Datacenter Workloads

Christopher Stewart
The Ohio State University

Jing Li
The Ohio State University

Datacenters: A New Research Area

- Datacenters are the *de facto* architecture for Internet, data-intensive, and increasingly HPC applications
  - Must revisit classic systems issues at datacenter scale

    Operating systems and middleware for massive FT

    Performance-to-power efficiency for processors

    TCP and network topology for high bandwidth, diverse workloads

- Introduce new power-architecture problems

  Deliver electricity to thousands of devices at low cost

  Cooling and long-term equipment maintenance

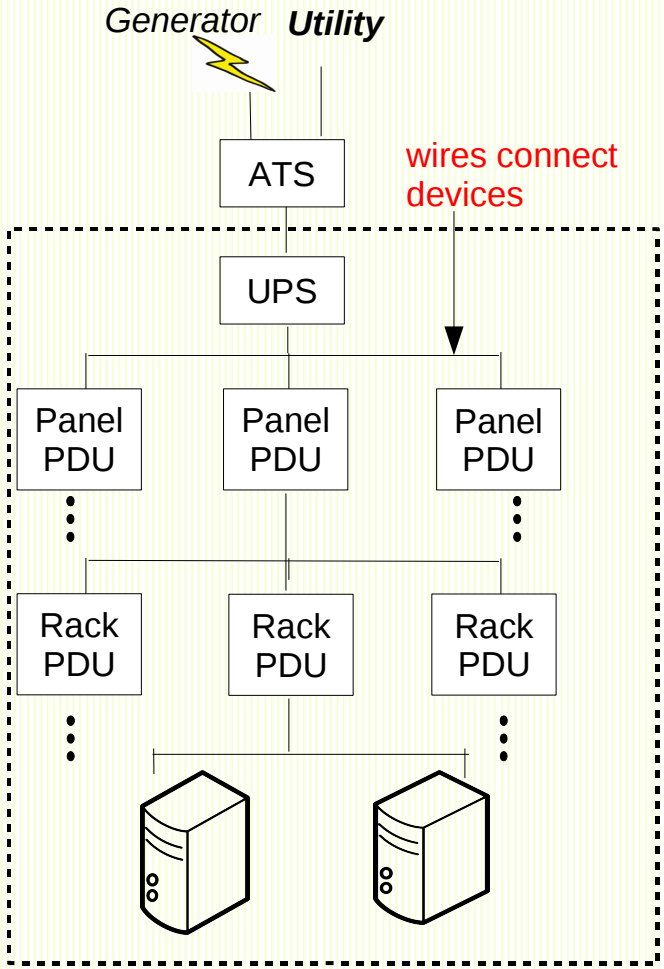
  Exploit low-cost and/or sustainable electricity

Power Provisioning in Datacenters

- Applications hosted in a datacenter share electrical circuits
  - Circuits have hard power capacity limits, i.e., fire code
  - Also, soft power limits/goals for load balancing
- Power provisioning maps application workloads to circuits

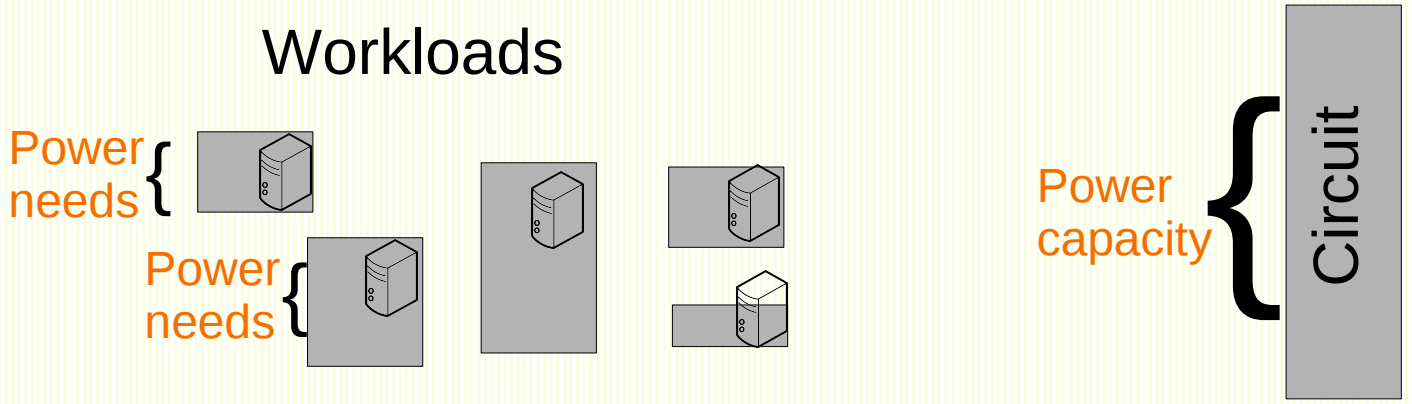
  Goal #1: Avoid exceeding limits

  Goal #2: Use all available capacity
  - Capacity is costly: $20--$100K for medium-sized upgrade



Power delivery subsystem in our local datacenter

Provisioning Non-Datacenter Workloads

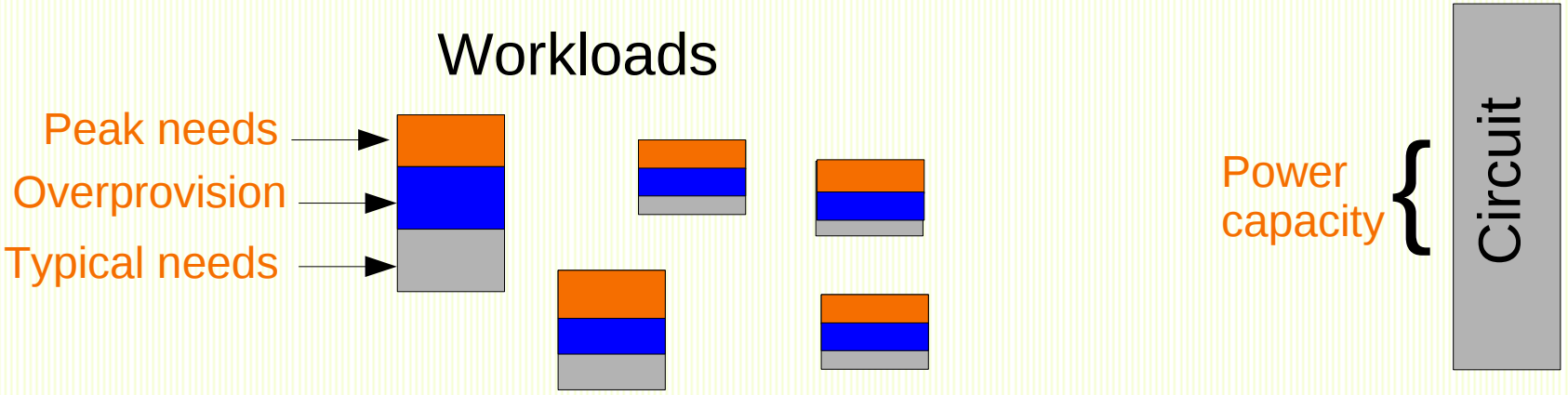


- Integer programming optimization
  1. Measure power needs of each workload
  2. Use circuit capacity as a constraint
  3. Find the assignment that places that uses as much of the circuit capacity as possible (knapsack)

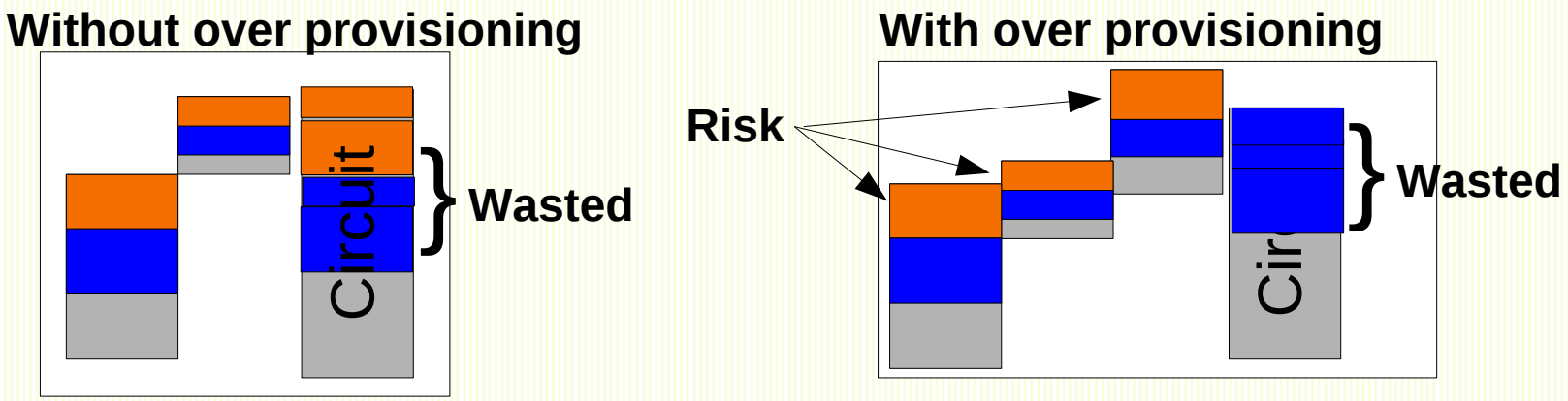
  Practical solution for static power needs, e.g., light bulbs

Datacenter Challenge

- Power needs can increase over time
  - Provisioning based on typical needs risks circuit breaks
- Practical solution: Provision based on peak needs
  - Use nameplate ratings [fan-isca-07] **or** measured peak power [govindan-eurosys-09]
  - Consider overprovisioning [fan-isca-07,govindan-eurosys-09,economou-mobs-06,pelley-asplos-10]



Datacenter Challenge cont.



- Peak-power provisioning → internal fragmentation

- Over provisioning offers an undesirable tradeoff:

  Reduce internal fragmentation but risk circuit breaks

Our Approach

- Working with real datacenters:
  - Study the whole distribution of power workloads
    - Typical, nameplate, and measured peak
  - Empirically observe common patterns
  - Exploit patterns for improved power provisioning
- Many datacenters face similar issues; Empirical results can have high impact
  - E.g, Nameplates exceed measured peaks by 60% on average [fan-isca-2007, economu-mobs-2006]
- Studying the whole-distribution allows us to find important patterns hidden in the common case

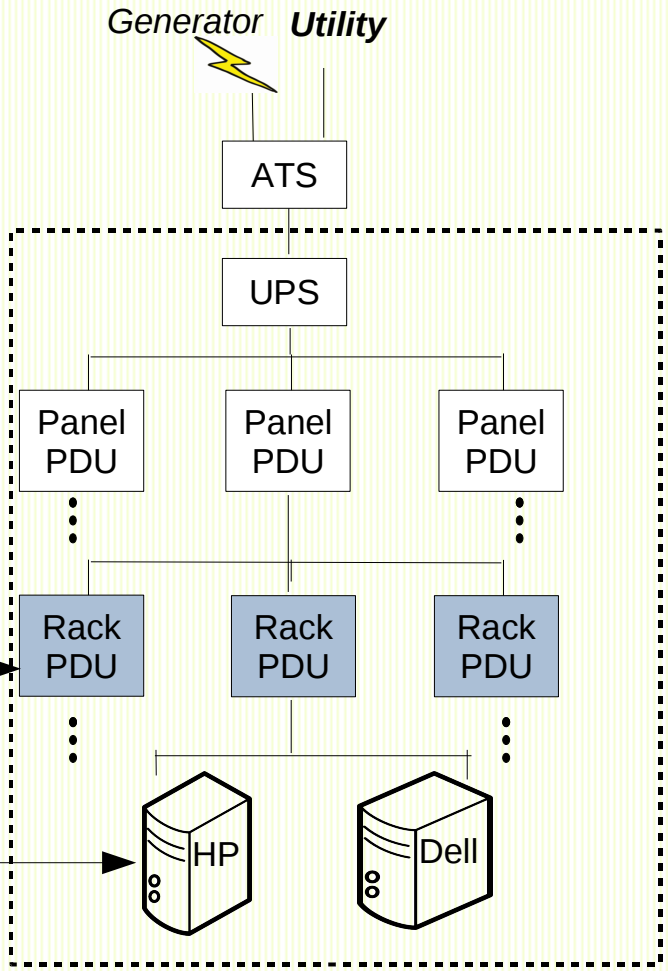
Outline

1. **Study of Power Workloads in 3 Real Datacenters**
2. Empirical Patterns: Diverse Power Utilization and Nonmonotonic Peak Power
3. Power Provisioning for Diverse Datacenter Workloads
4. Conclusion

Study of Power Workloads

1. Methodology
2. Datacenters
3. Unexpected Obstacles

- Power workloads studied:
  - Typical needs = statistical mode
  - Measured peak = largest observed over 3 mo. study
  - Nameplate = manufacturer provided data on peak needs
- Typical and measured peak were collected at rack-level PDU
- Nameplate data pulled from manufacturer websites
  - Per device nameplates are summed across a PDU



Power delivery subsystem in our local datacenter

Study of Power Workloads

- Codename OSU:
  - ~6500 compute cores 300 PDU
  - University-wide datacenter hosting a wide range of applications from enterprise to medical
- Codename CSE:
  - ~815 compute cores 35 PDU
  - Our departments server room
- Codename PROD:
  - ~980 compute cores 62 PDU
  - Academic services for the OSU College of Engineering

The Datacenter Floor is Sacred!

- Single point of failure, security breach, etc.
- Hard to gain access as a research group
  - "We have asked them to arrange a time to escort you all through the datacenter floor.... We have asked them not to provide any information about the server names or functions.... if your group does... then you will be asked to leave." --- System Administrator
- Call for research: Non-invasive techniques to profile datacenter power workloads
  - Allow more academics to participate
  - Encourage enterprises to become active

  [Open Platform, Facebook]

Outline

1. Study of Power Workloads in 3 Real Datacenters

2. **Empirical Patterns: Diverse Power Utilization and Nonmonotonic Peak Power**

3. Power Provisioning for Diverse Datacenter Workloads

4. Conclusion

Diverse Power Utilization

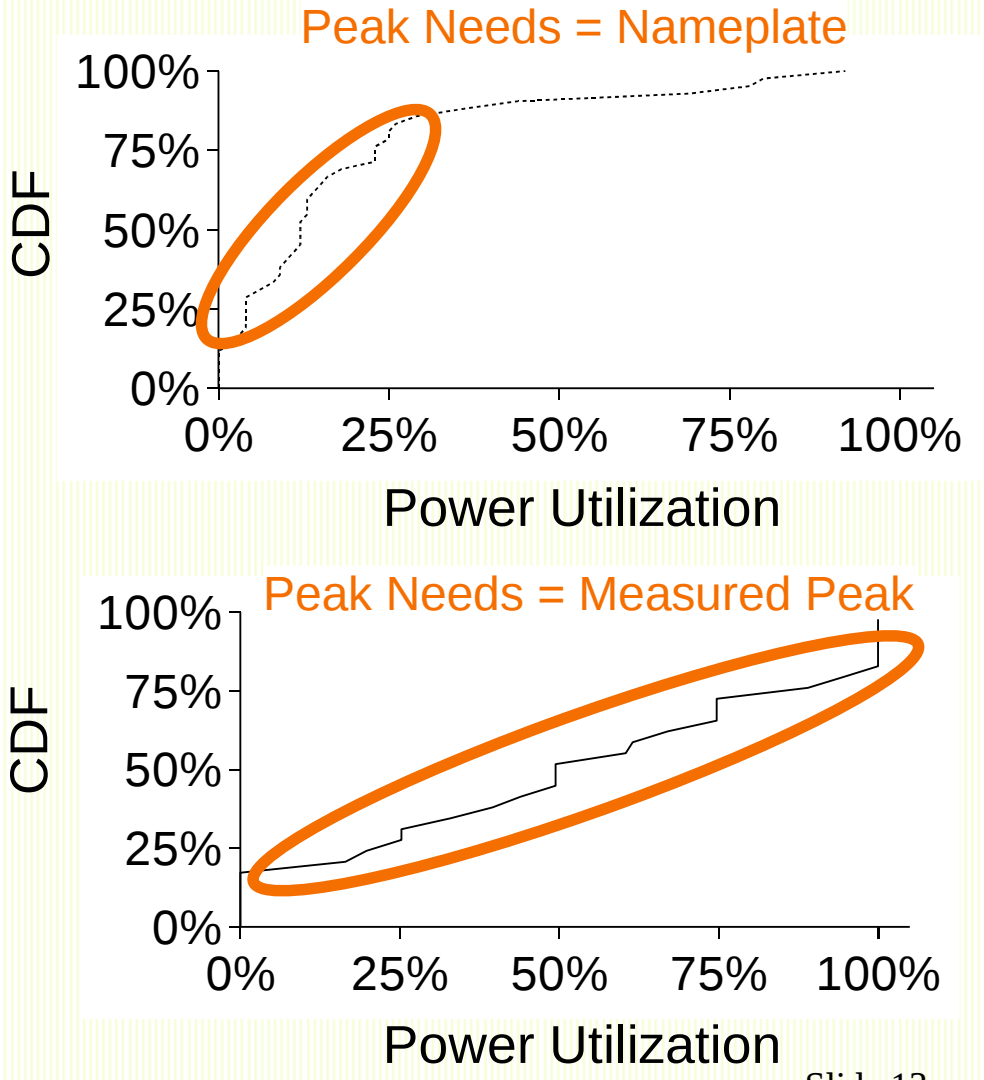
1. Nonmonotonic Peaks
2. Empirical Results
3. Impact of Factors

- Typical power utilization:

  Typical Needs / Peak Needs

- The majority of supported workloads span:
  - 0—30% under nameplate
  - 15—100% measured peak

- Confirmed: Nameplate rating overestimates typical needs by 65% [economou-mobs-06] 95% of the time

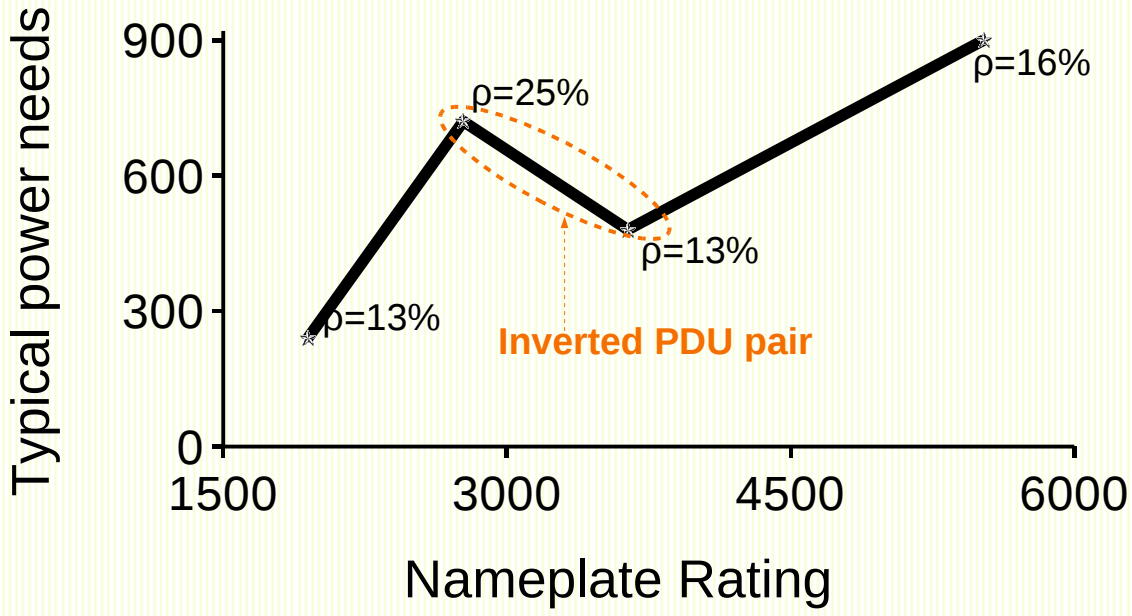


Diverse Power Utilization

1. Nonmonotonic Peaks
2. Empirical Results
3. Impact of Factors

- Diverse power utilization led to a surprising result:

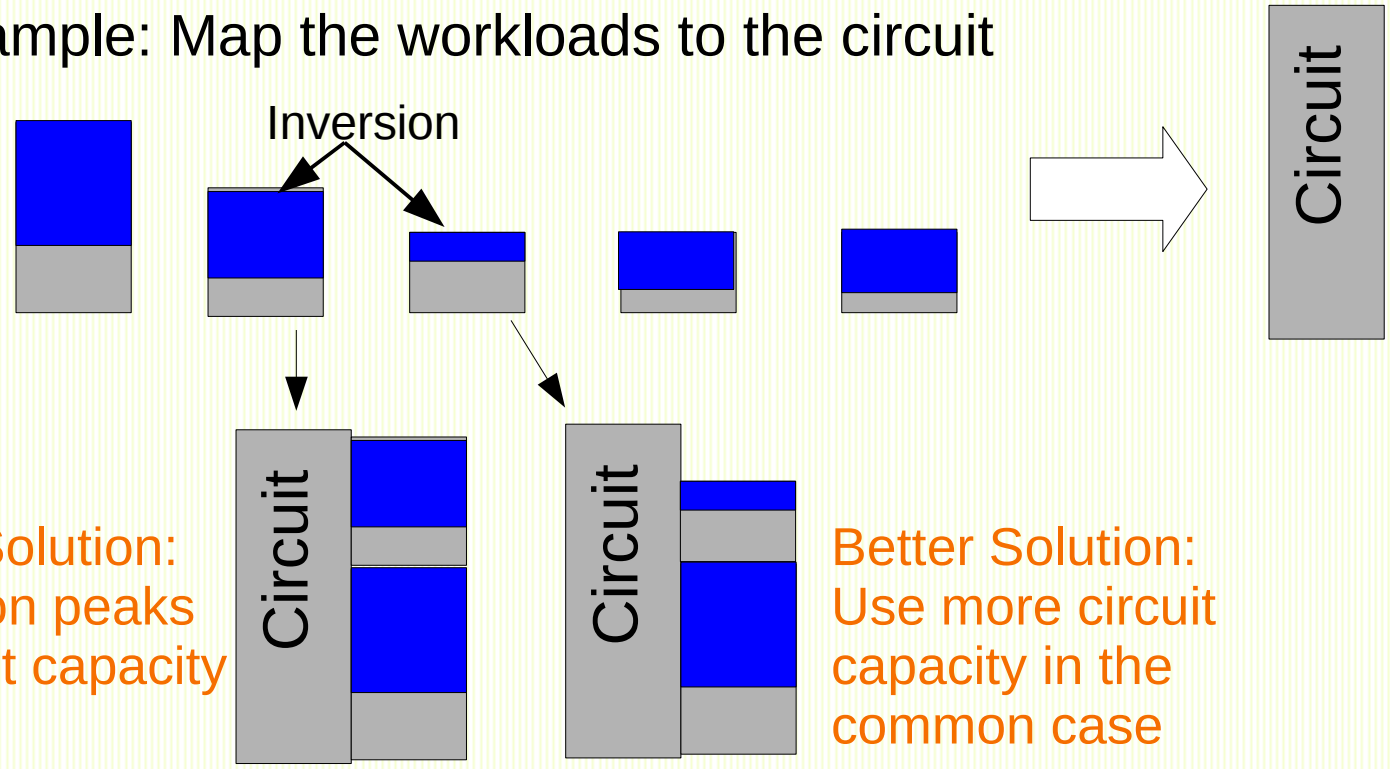
  The relationship between peak and typical power needs was nonmonotonic.



Diverse Power Utilization

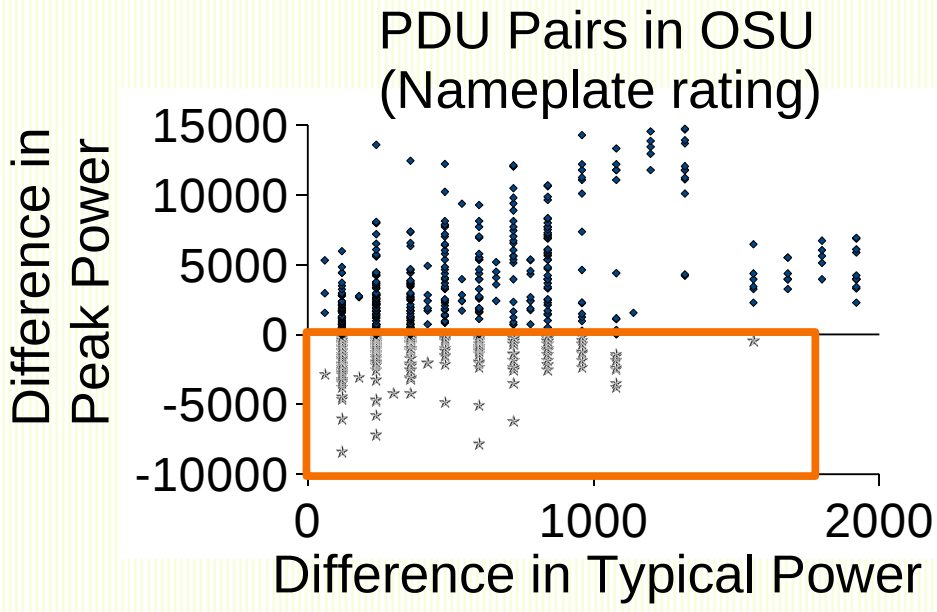
Nonmonotonic peak power can lead to poor power provisioning decisions since peak needs don't reflect typical needs.

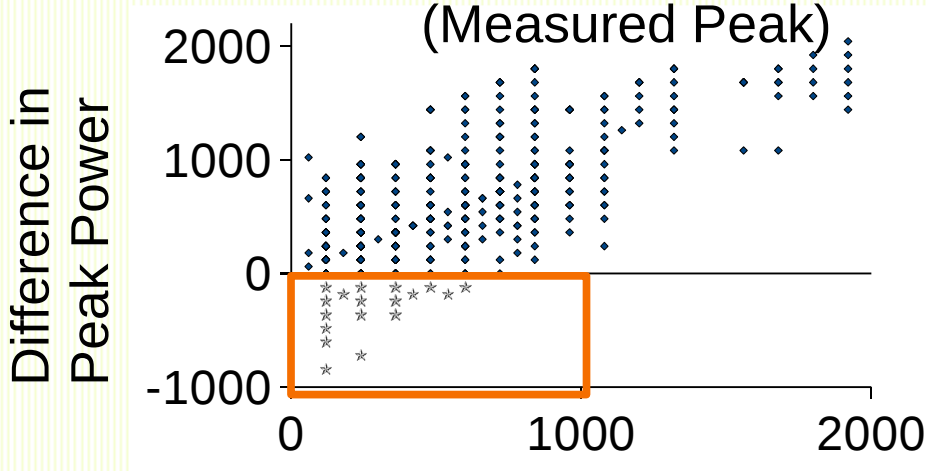
Example: Map the workloads to the circuit



Diverse Power Utilization

1. Nonmonotonic Peaks
2. Empirical Results
3. Impact of Factors





- *How frequently do order inversions occur?*
- Peak = Nameplate ratings
  - OSU – 27%
  - CSE – 38%
  - PROD – 23%
- Peak = Measured peak
  - OSU - 12%
- Note, whole distribution analysis. Order inversion is not the common case.
- Impact can be significant

Diverse Power Utilization

- In the paper, we studied the impact of several factors on nonmonotonic peaks, including:
  - Utilization distribution stability
  - Time of day
  - Persistence throughout the day
  - Inversion impact
  - Power utilization clustering

Outline

1. Study of Power Workloads in 3 Real Datacenters

2. Empirical Patterns: Diverse Power Utilization and Nonmonotonic Peak Power

3. Power Provisioning for Diverse Datacenter Workloads

4. Conclusion

Power Provisioning

1. In Practice
2. Our Approach
3. Prelim Results

8 Recently added PDU in OSU

- Challenge: Given new incoming workloads (right table), map workloads to circuits such that:
  - The sum of peak power does not exceed circuit capacity
  - Maximize the amount of power used in the common cas
  - *Mapping must be based on peak power only*
- Overprovisioning → Study many circuit capacities

Power Utilization	Measured Peak
4%	3119
4%	3119
13%	3823
9%	3823
100%	480
33%	1080
5%	2500
4%	960

Power Provisioning

- Integer programming with peak power (IP)
  - Assumes peak and typical needs are monotonic
  - Picks wrong workloads under diverse utilization
  - Can perform worse as circuit capacity increases
- Smallest peak power first (SPPF)
  - In the single circuit case, SPPF avoids all order inversions
  - Picks wrong workloads when large-peak workloads have high utilization
  - Stable across capacity increases

Power Provisioning

- Best of both worlds
  - Stability of SPPF, common case capacity usage of IP w/ monotonic peaks
- Approach:
  - Create SPPF assignment
  - Many IP assignments
  - Use the observed power-utilization distribution to estimate the potential loss due to inversions
- *Honest moment: Introduce a new parameter (cert) that reflects an administrators fear of inversions*

Power Provisioning

1. In Practice
2. **Our Approach**
3. Prelim Results

Create multiple solutions

Compare solutions

New parameter to control *certainty* of avoiding inversions

Use empirical observations of power utilization

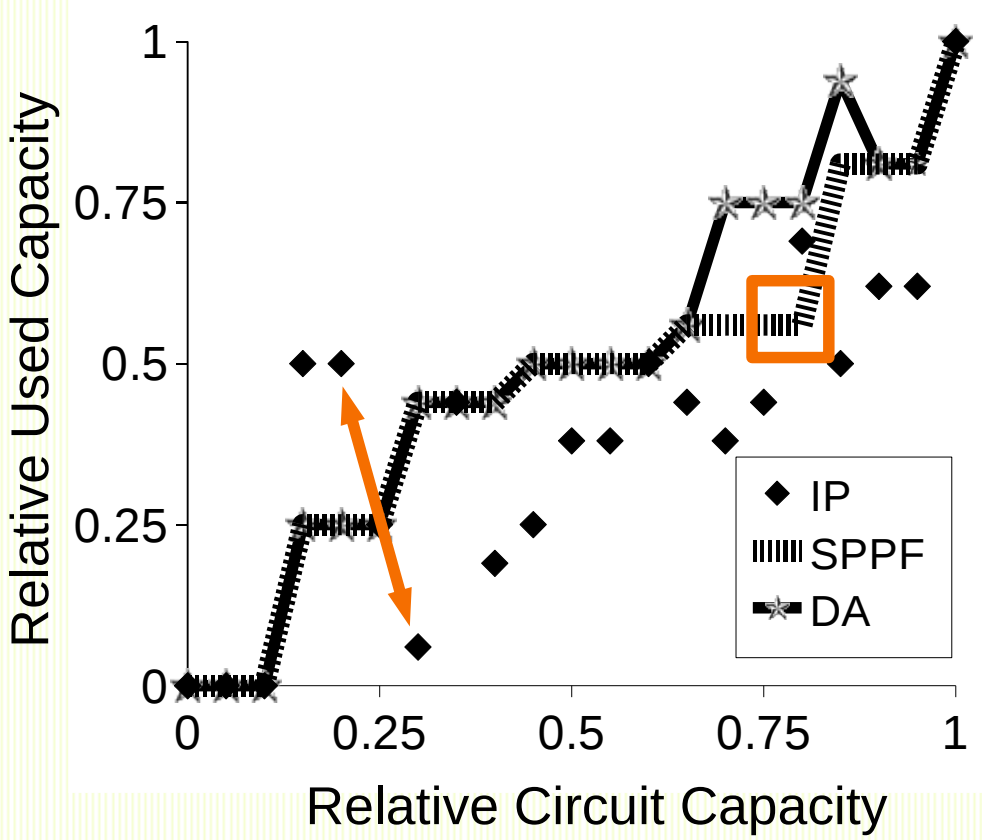
```
DA_Provsioning (candidates, capacity, utilCDF) {
# candidates -> {Pmp(0),...,Pmp(i)}
# capacity -> int C
# utilCDF -> Hashtable(keys=K^th percentile, val=power util.)

    assignment base_solution = {};
    assignment alt_solution = {};

    base_solution = SPPF(candidates, capacity);
    int alt_count=sumPeakNeeds(base_solution);
    while (alt_count < capacity)
        alt_count++;
        alt_solution = knapsack(candidates, capacity);
        if (DA(alt_solution) > DA(base_solution)
            base_solution = alt_solution;
    return base_solution;
}

float DA(assignment A) {
    float tot_cost = 0;
    float cert = 0.1;
    forall a in A
        float max_cost = 0;
        forall c in candidates and not in A
            if (Pmp(c)*utilCDF{1 - cert} >
                                Pmp(a)*utilCDF{cert})
                if (Pmp(a) > Pmp(c)) # Inverted PDU
                float this_cost = (Pmp(a) - Pmp(c))
                if (this_cost > max_cost)
                    max_cost = this_cost;
        tot_cost += max_cost;
    return sumPeakNeeds(a) - tot_cost;
}
```

TABLE III: Pseudo-code of our provisioning approach.

Power Provisioning

1. In Practice
2. Our Approach
3. **Prelim Results**

- X- and Y-axis are relative to the needs of studied PDU
- IP is unstable as circuit capacity
- SPPF can perform poorly
- *Diversity-aware is stable and uses the most capacity*



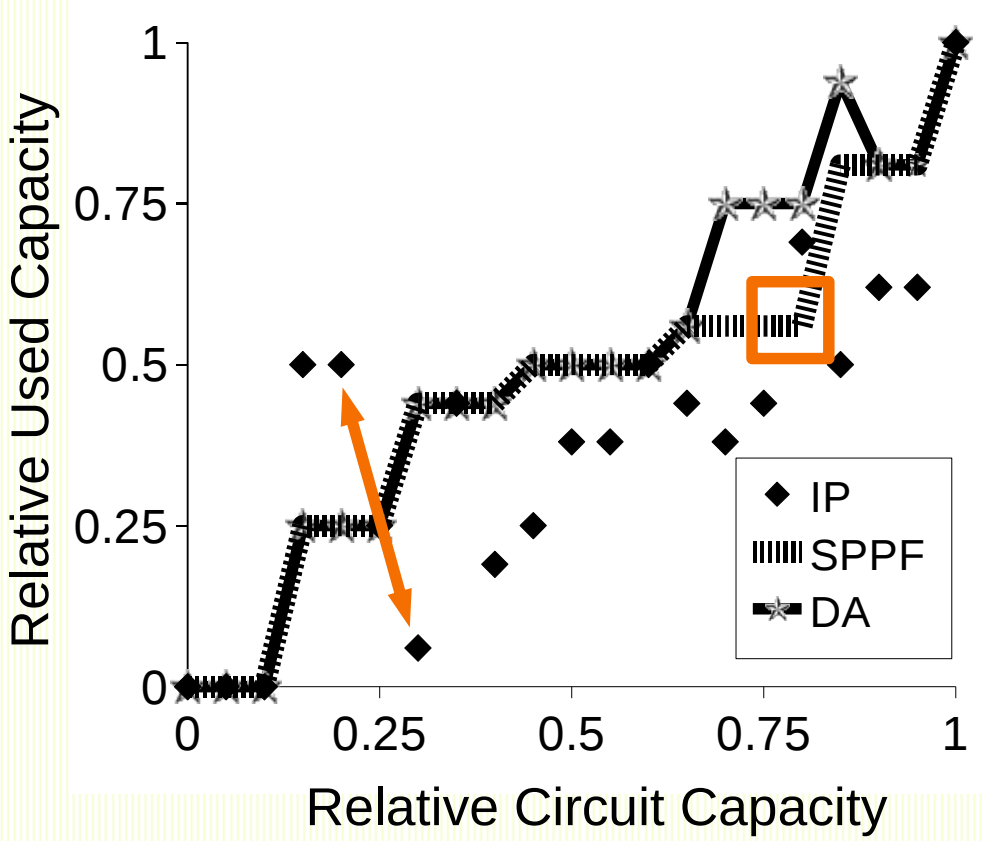
Power Provisioning

1. In Practice
2. Our Approach
3. **Prelim Results**

- Performance metric:
  Fraction of tests where diversity-aware uses the most capacity

  versus

  - *IP: 89%*
  - *SPPF: 98%*
  - *LPPF: 70%*
  - *L-to-S: 93%*
  - *FCFS: 80%*



Future Work

- Provision for multiple circuits with load balancing objectives
- Theoretic bounds on diversity-aware provisioning
- Dynamic power provisioning

Future Work

- Datacenter workloads are diverse, presenting new challenges in power provisioning
- Naive peak-power provisioning makes poor decisions when peak needs do not reflect typical needs
- We propose diversity-aware power provisioning that can provide stable results and use circuit capacity