Thermal Time Shifting: Leveraging Phase Change Materials to Reduce Cooling Costs in Warehouse-Scale Computers

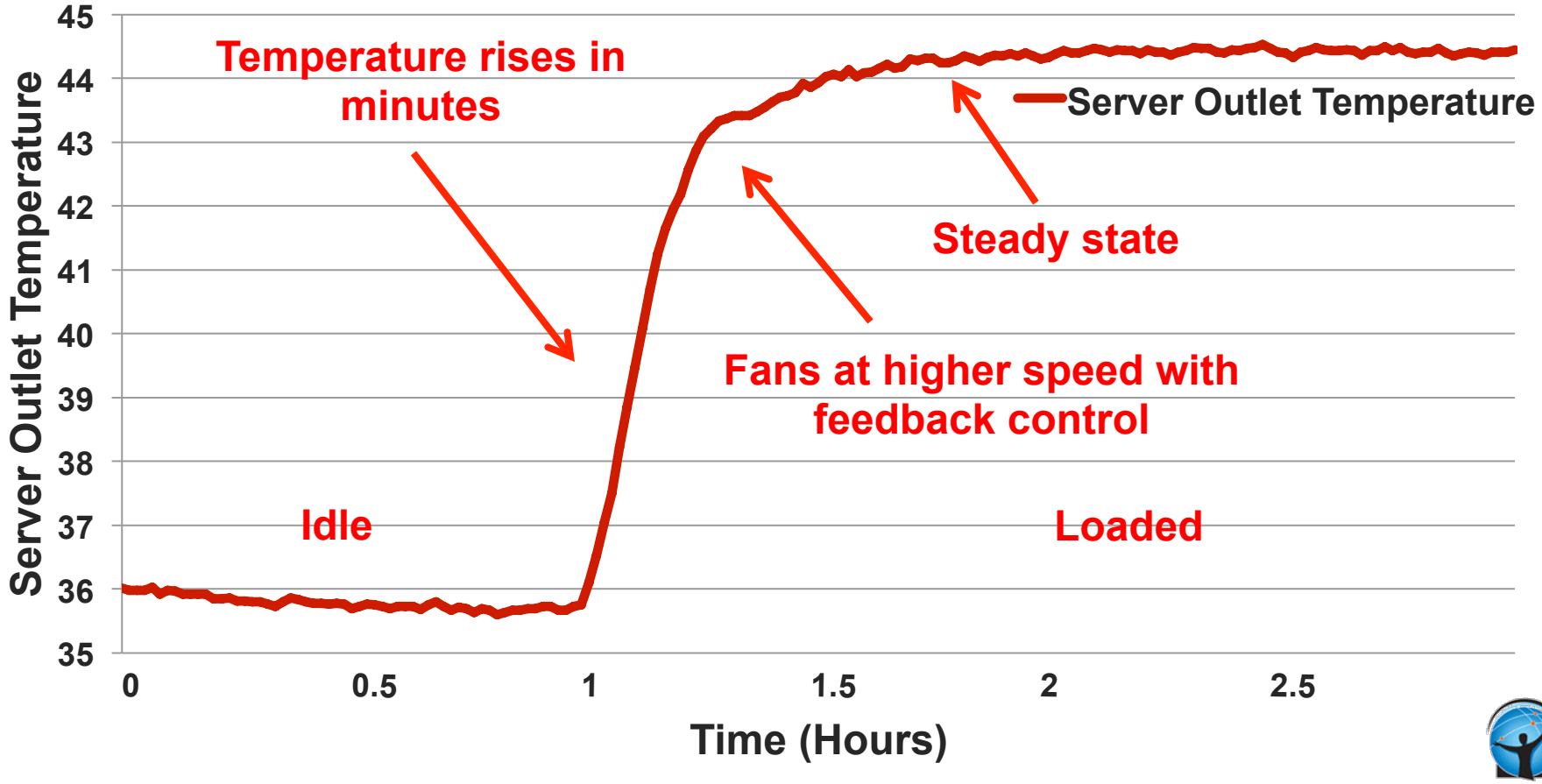
[1] University of Michigan
[2] Advanced Micro Devices, Inc.
[3] University of California, San Diego

MATT SKATCH[1], **MANISH ARORA[2,3]**, CHANG-HONG HSU[1], QI LI[3], DEAN M. TULLSEN[3] LINGJIA TANG[1] & JASON MARS[1]

JUNE 2015

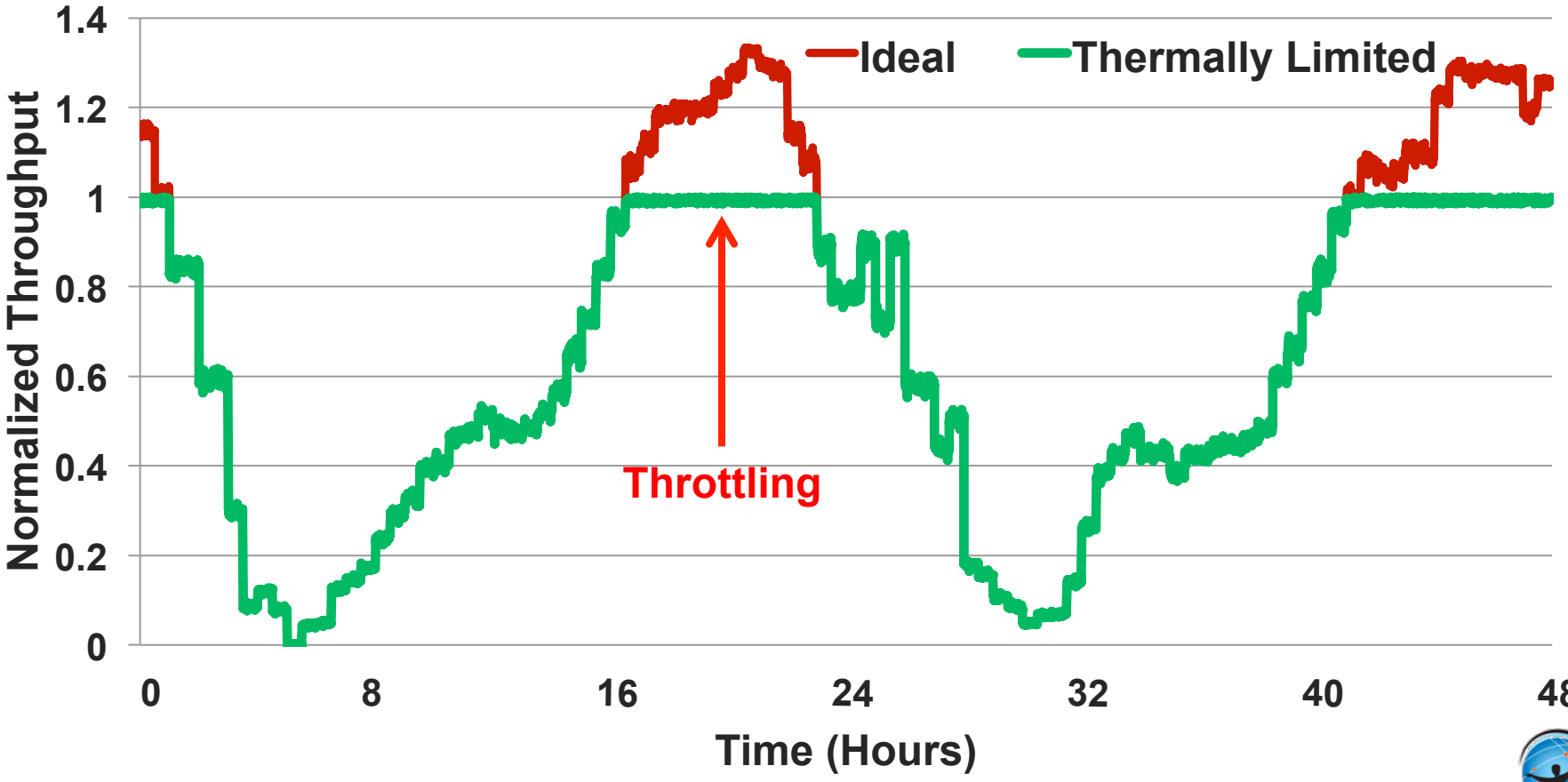
REAL MEASUREMENTS ON TEMPERATURE RISE

- Temperature increases rapidly with power increase
- Compute & Cooling coupling 90W/20W → 185W/45W
- Heat needs immediate expulsion, low "Thermal Capacity"



THE BIG PICTURE – COOLING PROBLEM 1

- Datacenters limited by peak cooling capacity
  - Need bigger cooling, but would rather spend on compute
  - DVFS applied, right when u really need performance

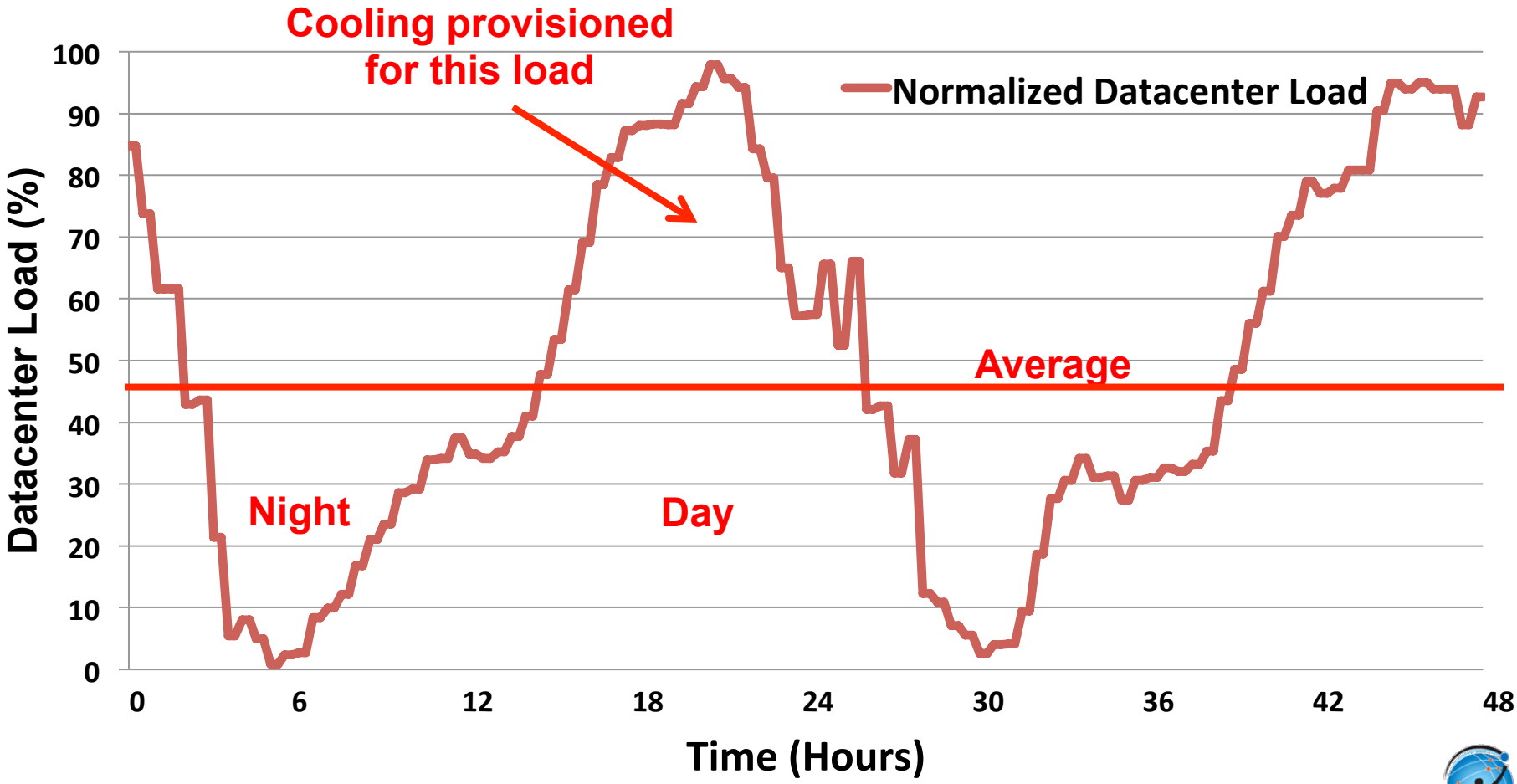


THE BIG PICTURE – COOLING PROBLEM 2

- Cooling capacity provisioned “once” for a datacenter
  - Prevents upgrades to more powerful / denser systems
  - Increasing utilizations / power even for same systems
    - More co-location and better system management
    - Software stacks getting better every few months
- Either not increase compute capacity or suffer more throttling

THE BIG PICTURE – COOLING PROBLEM 3

- Cooling system design is over-provisioned vs. avg. run
  - Designed to support peak load
  - Run at lower efficiency at all other times



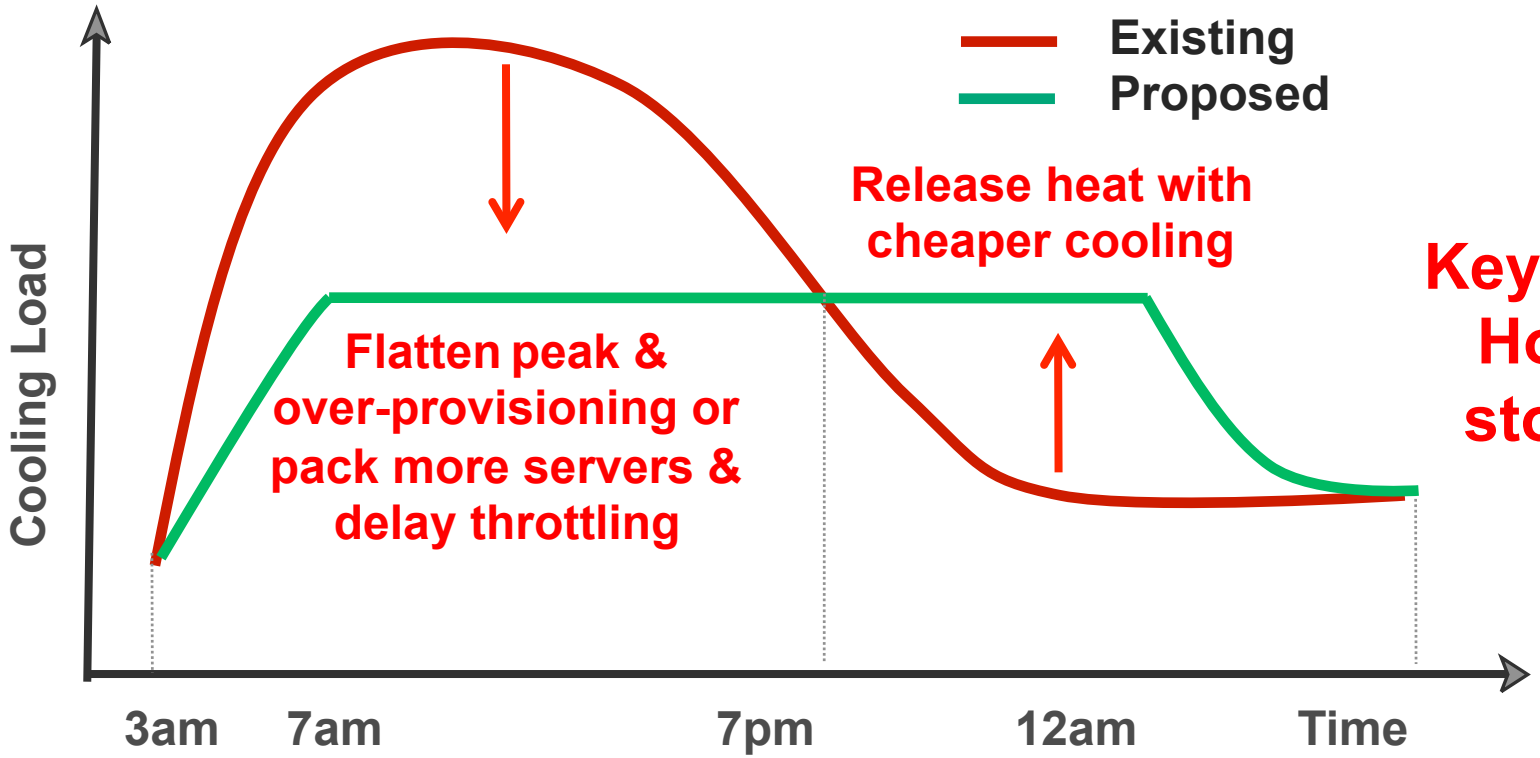
KEY TAKE AWAY

- Summary of datacenter cooling system problems
  - Poor timing, more power needed to cool when power should actually be spent on compute
  - Does not easily allow upgrades / peak power increases
  - Expensive, costly to operate and runs in-efficiently for most of the time

- Need a way to throttle the temperature without throttling compute

THE BIG PICTURE

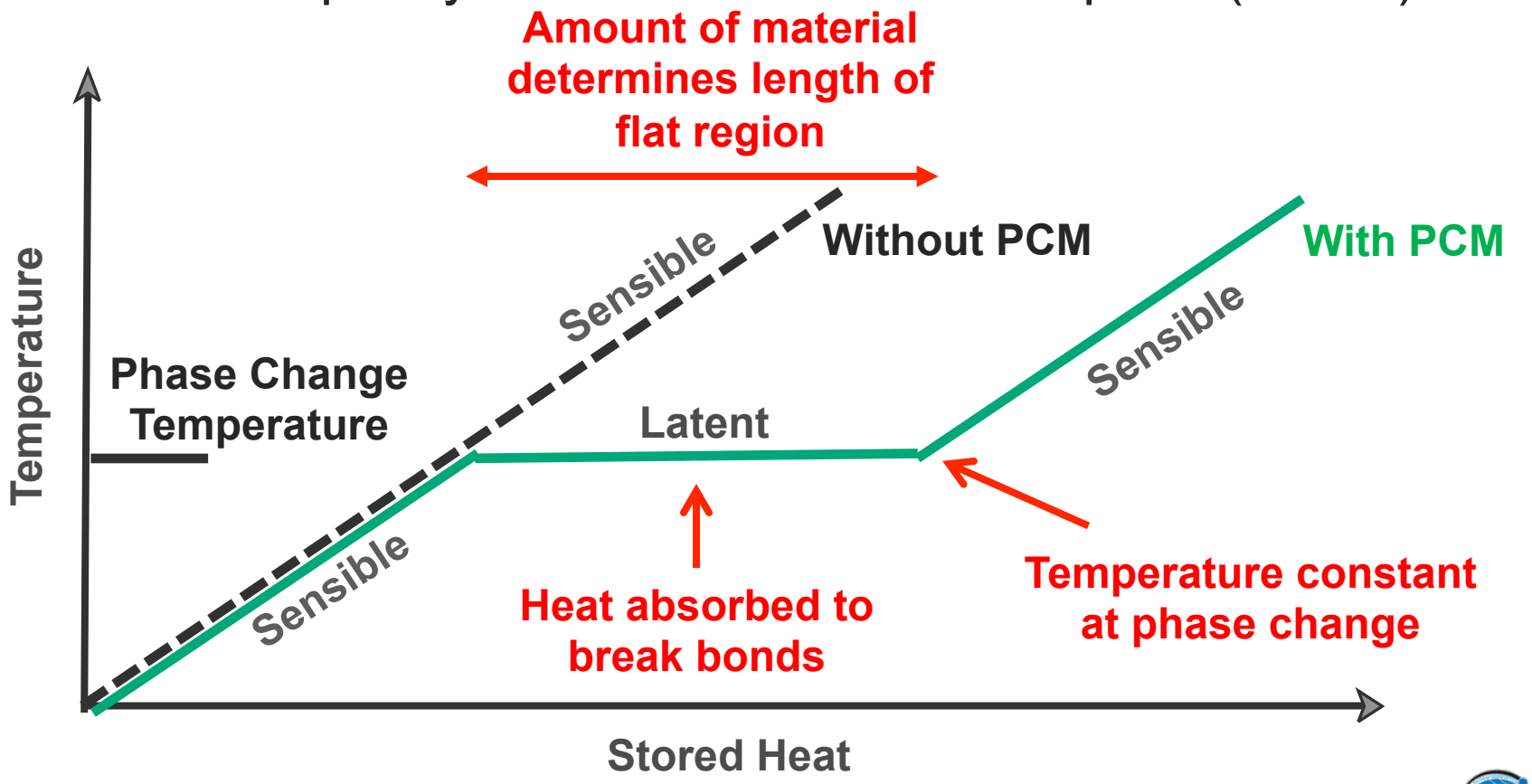
- Proposal: Thermal Time Shifting
  - Store heat, don't cool everything immediately
  - Break compute-cooling coupling
  - Build cooling systems provisioned for common case
  - Or enable more compute



Key question: How to we store heat?

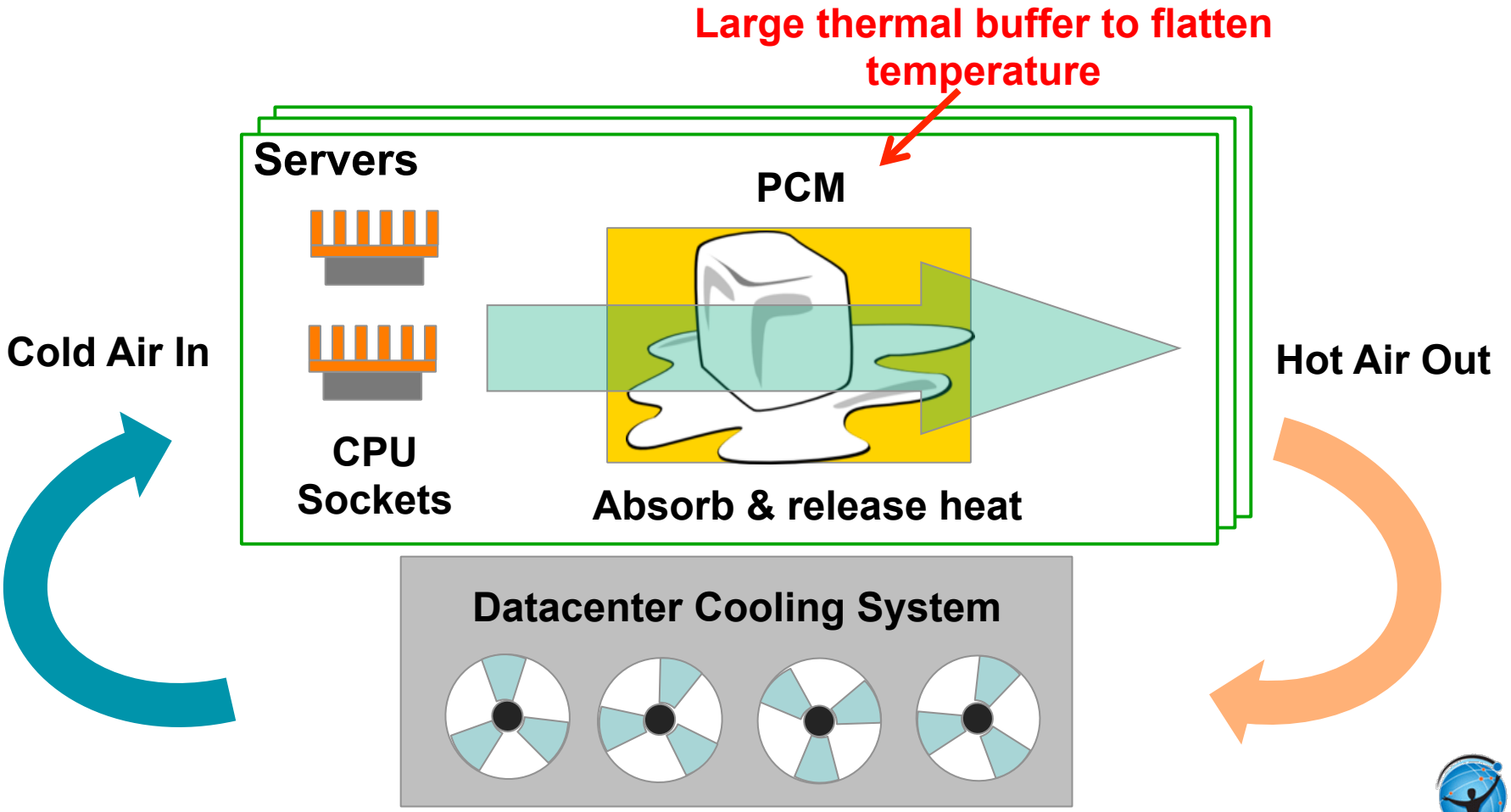
PHASE CHANGE MATERIALS (PCM) 101

- Explored in “Computational Sprinting” work
- Materials have ability to store heat
- Latent heat capacity >> Sensible heat absorption (> 50X)



THE BIG PICTURE

- Phase Change Materials (PCM) on server
  - Latent heat of fusion (~200 Joules/gram)
  - Integrate liters on server (~Mega joule storage range)



GOAL & OUTLINE

- Research objective
  - Study PCM as a thermal time shifting mechanism for servers in a datacenter

- Outline
  - Understanding thermal impact of PCM in servers
  - Is more PCM always better?
  - Evaluation
    - Reducing cooling system size with PCM
    - Improving datacenter throughput with PCM
  - Conclusions

AMD

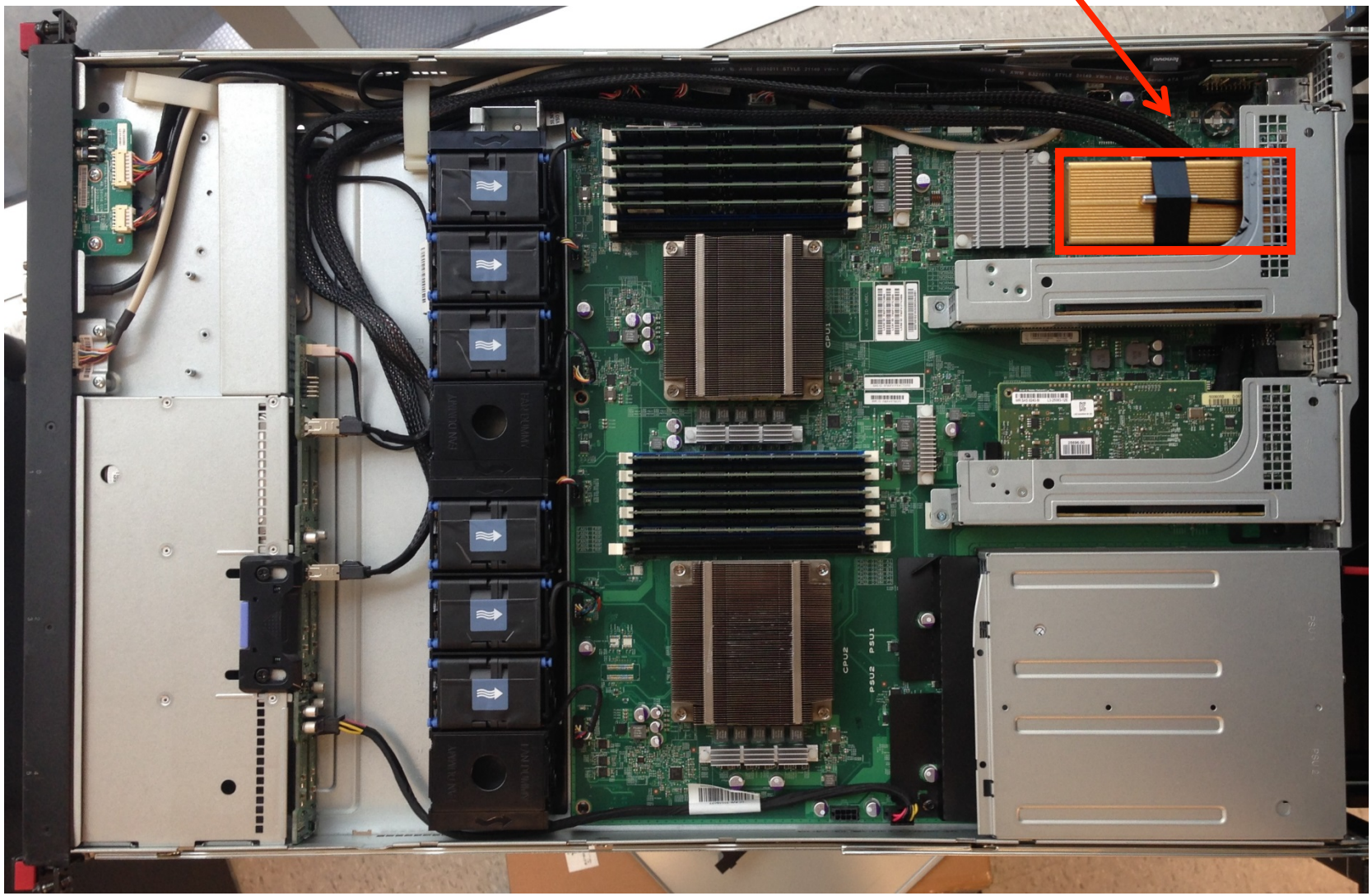
UNDERSTANDING THERMAL IMPACT OF PCM IN SERVERS

UNDERSTANDING THERMAL IMPACT OF PCM

- Built a physical prototype
  - Encapsulated PCM in a container and placed it on the server
  - Studied the temperature impact on a real system

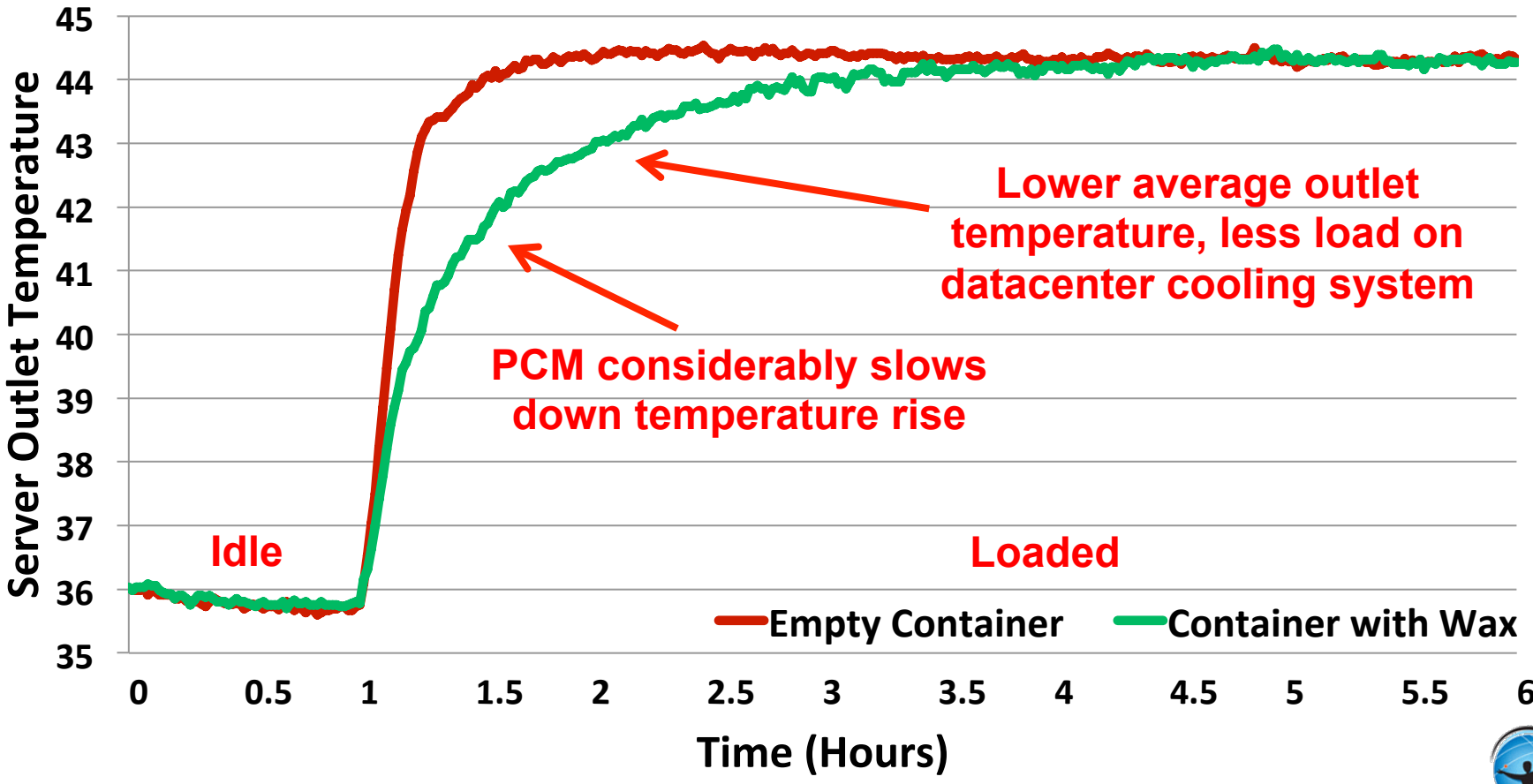
- Built Computational Fluid Dynamics (CFD) models
  - Study the temperature impact with different amounts of PCM
  - Understand tradeoffs in placing PCM

PCM PROTOTYPING WITH IBM RD330



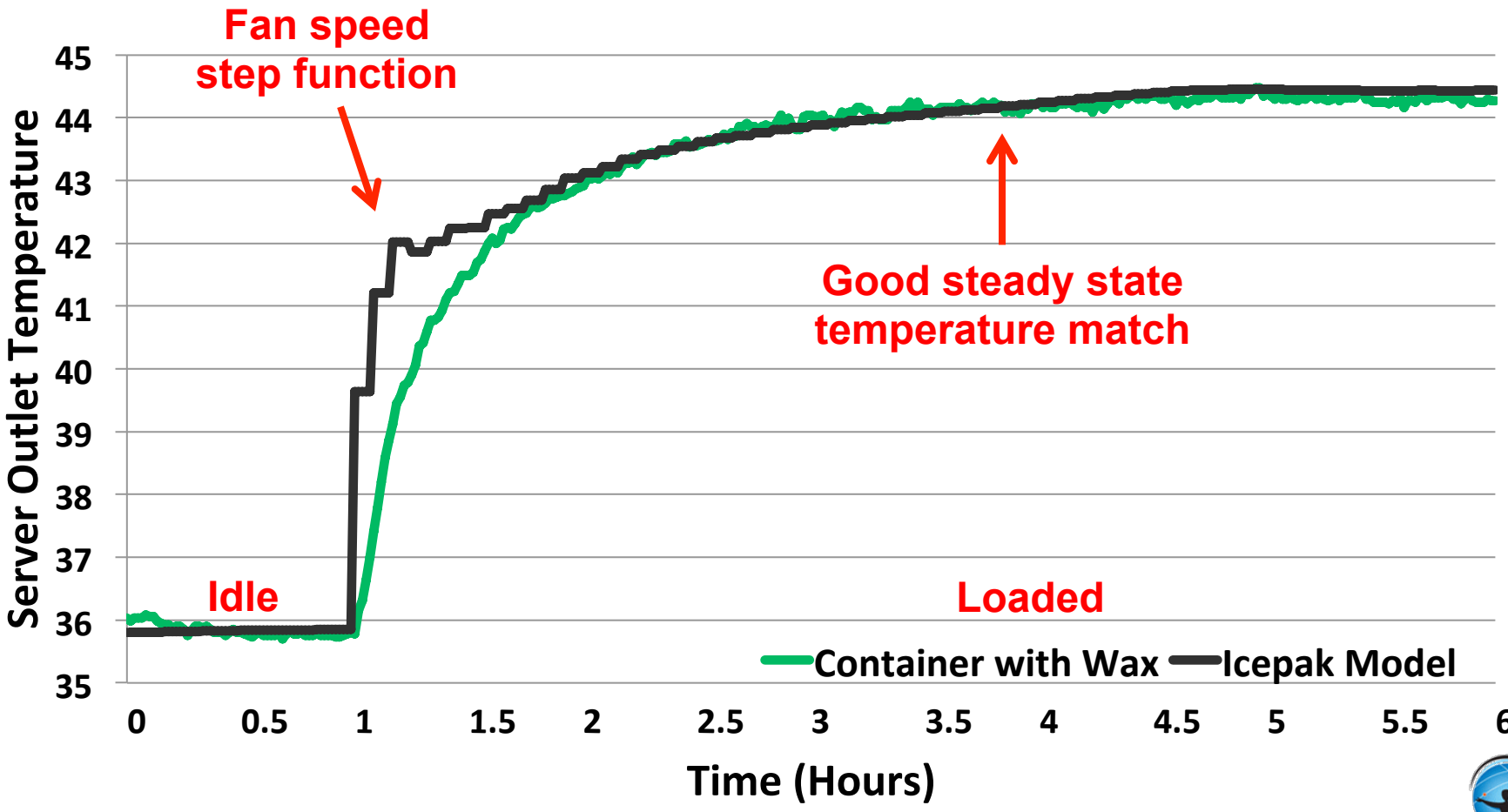
THERMAL IMPACT OF PCM

- Container with 90ml of Paraffin
  - Slower rise of temperature
  - Less load on cooling system / Delayed throttling



THERMAL IMPACT OF PCM

- CFD Model correlation
  - Observe time based fan speed step function
  - Excellent correlation at steady state

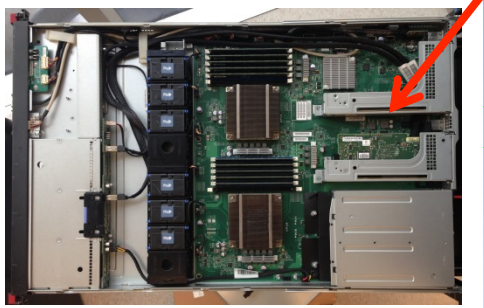


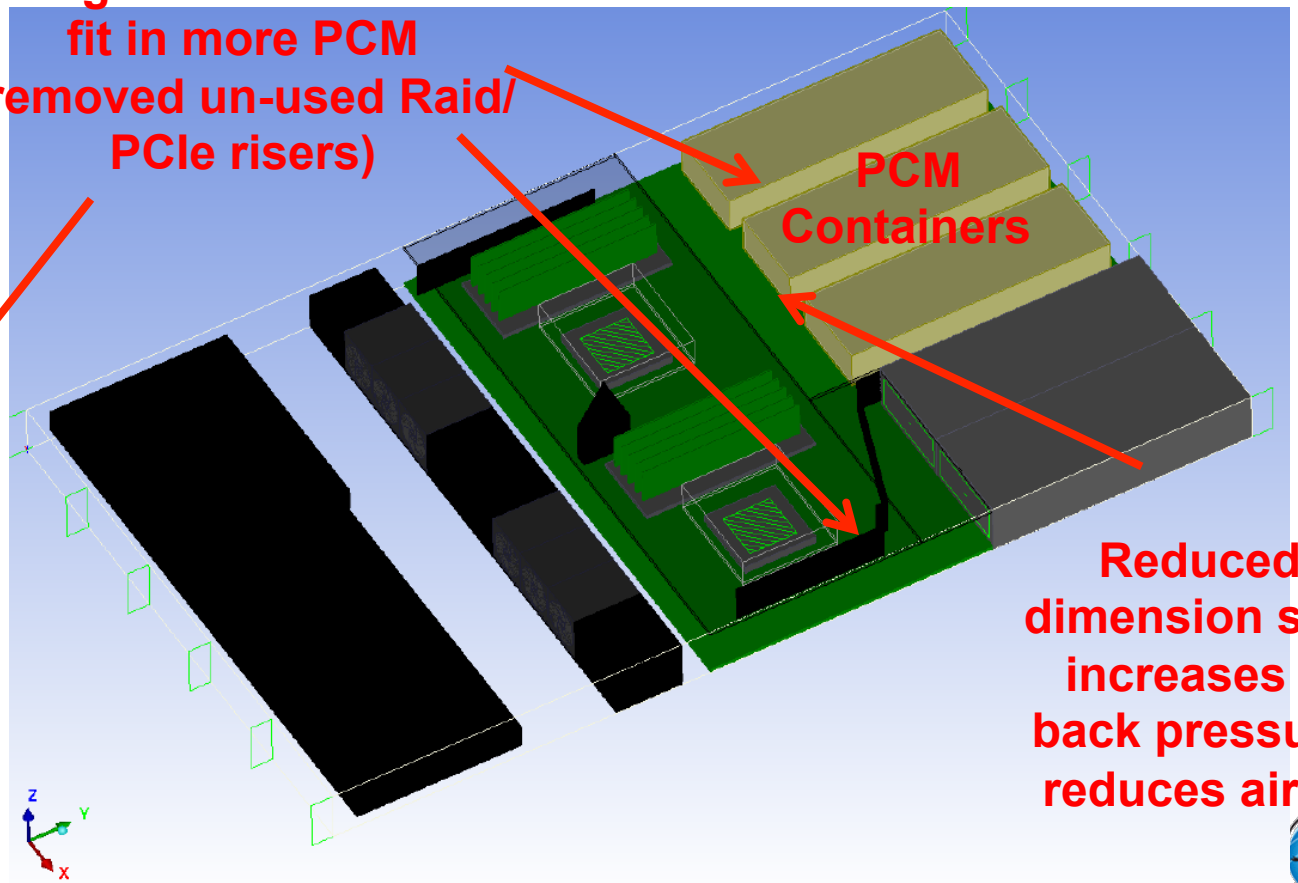
AMD

IS MORE PCM ALWAYS BETTER?

HOW MUCH PCM ON A SERVER?

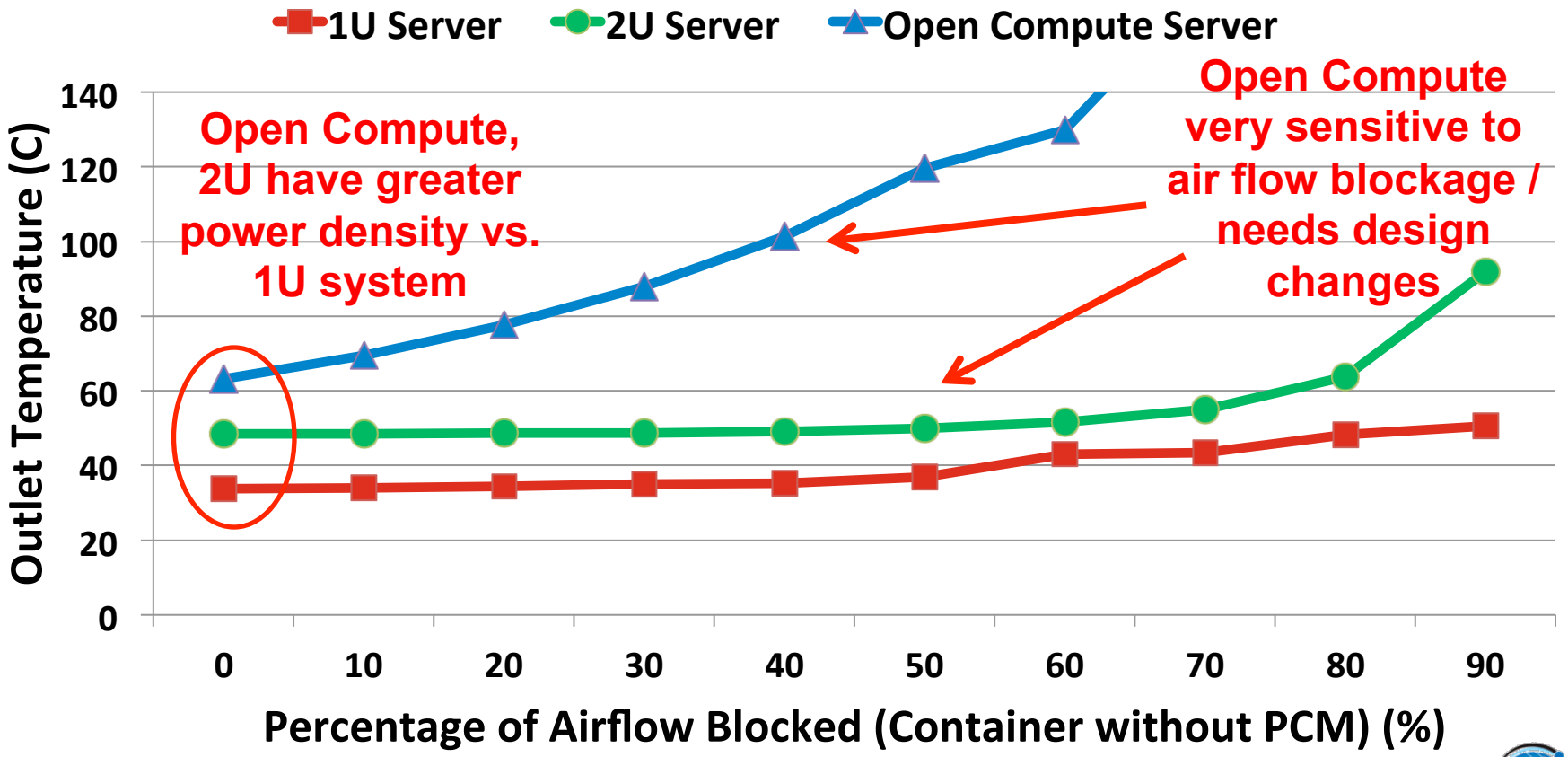
- Not one size fits all
  - Opportunity varies with server design / physical tweaks
  - PCM is downwind, but can block air going out



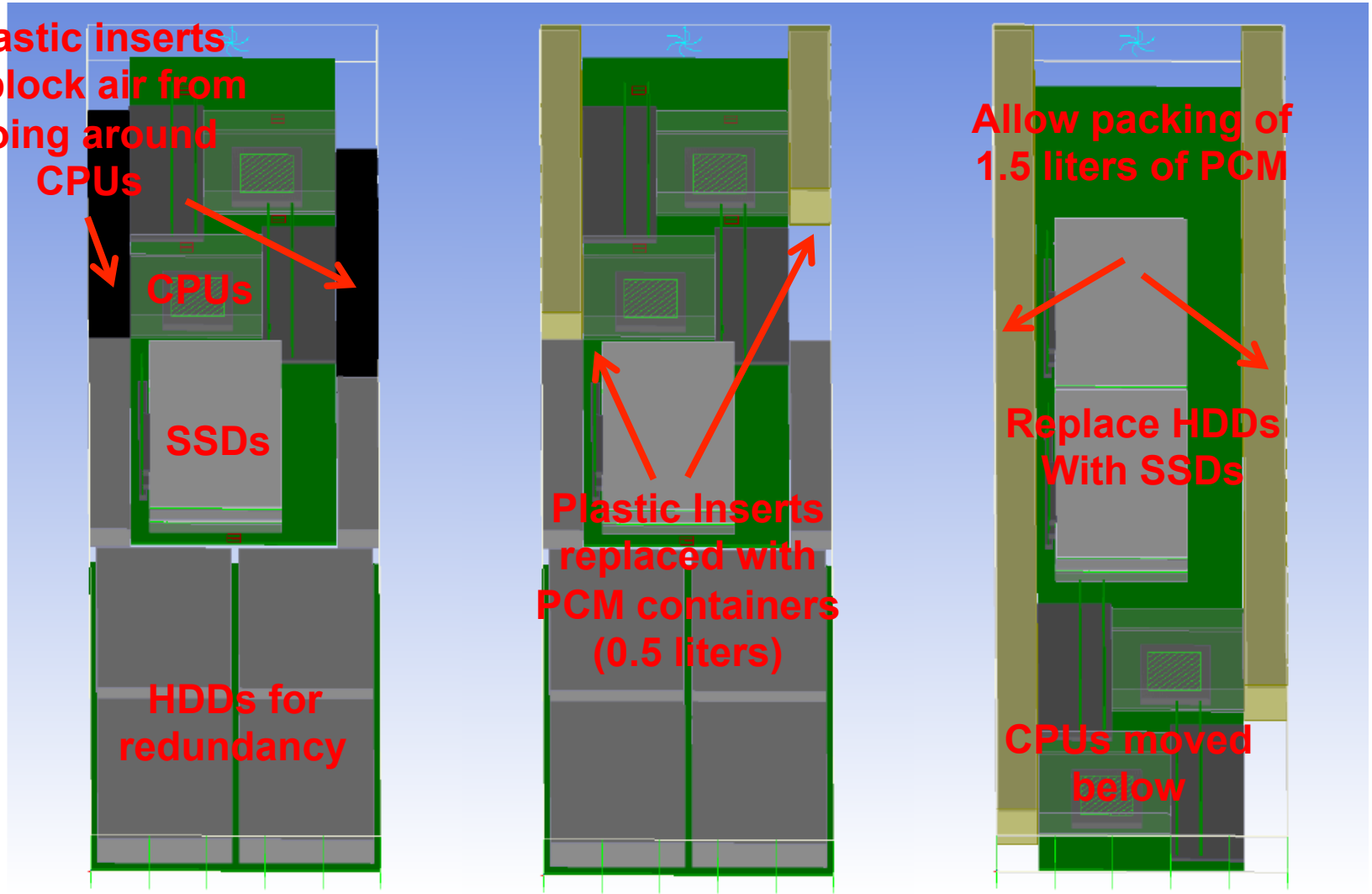


HOW MUCH PCM ON A SERVER?

- Experimentation with 3 different server designs
  - 1U, 2U and Open Compute Blade server
  - Tradeoff between amount of PCM & airflow

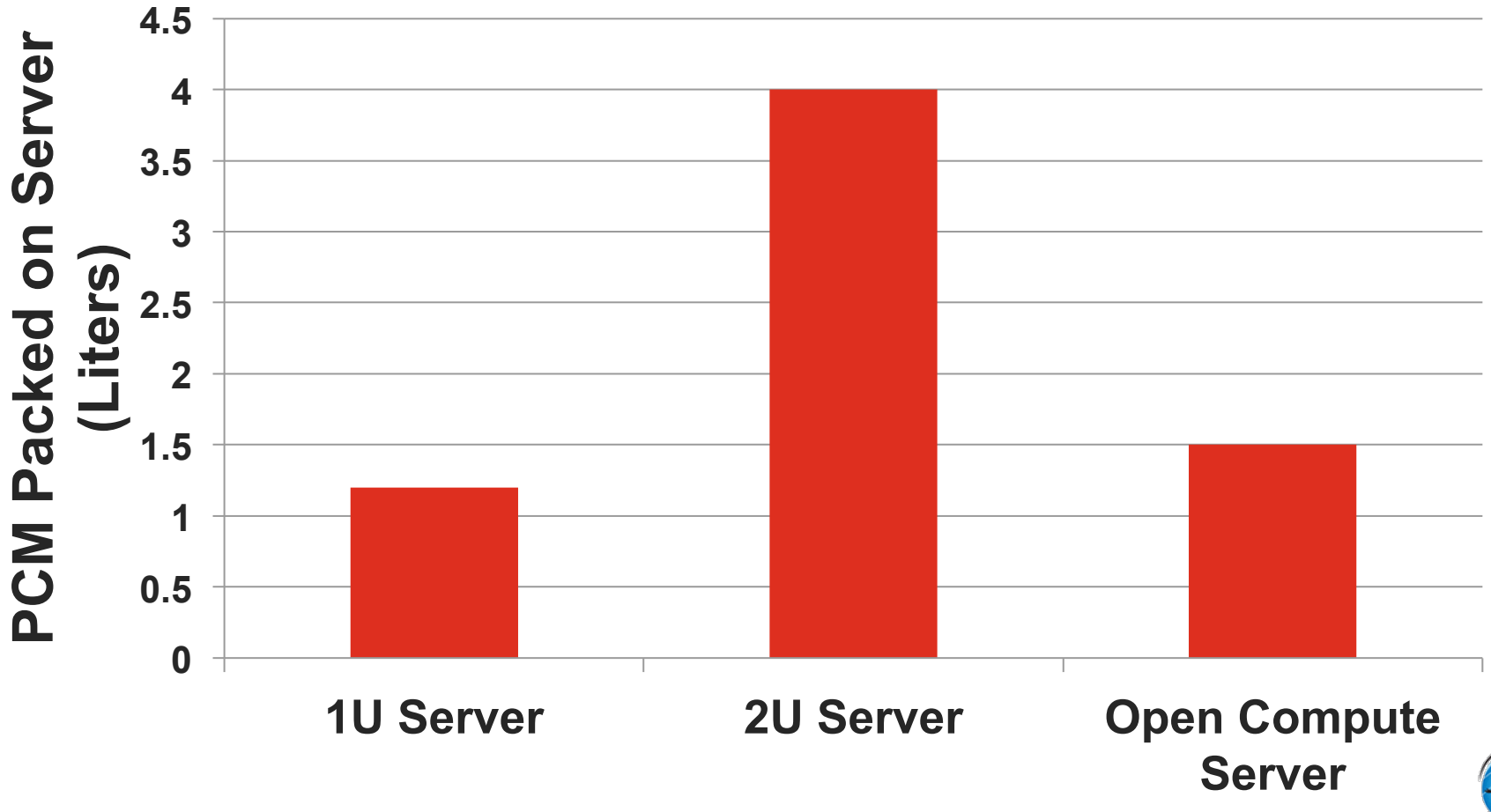


PACKING PCM IN OPEN COMPUTE SERVERS



HOW MUCH PCM FIT IN?

- Packing PCM on servers
  - Server / PCM co-design needed
  - Benefits outweigh effort required to change board layout

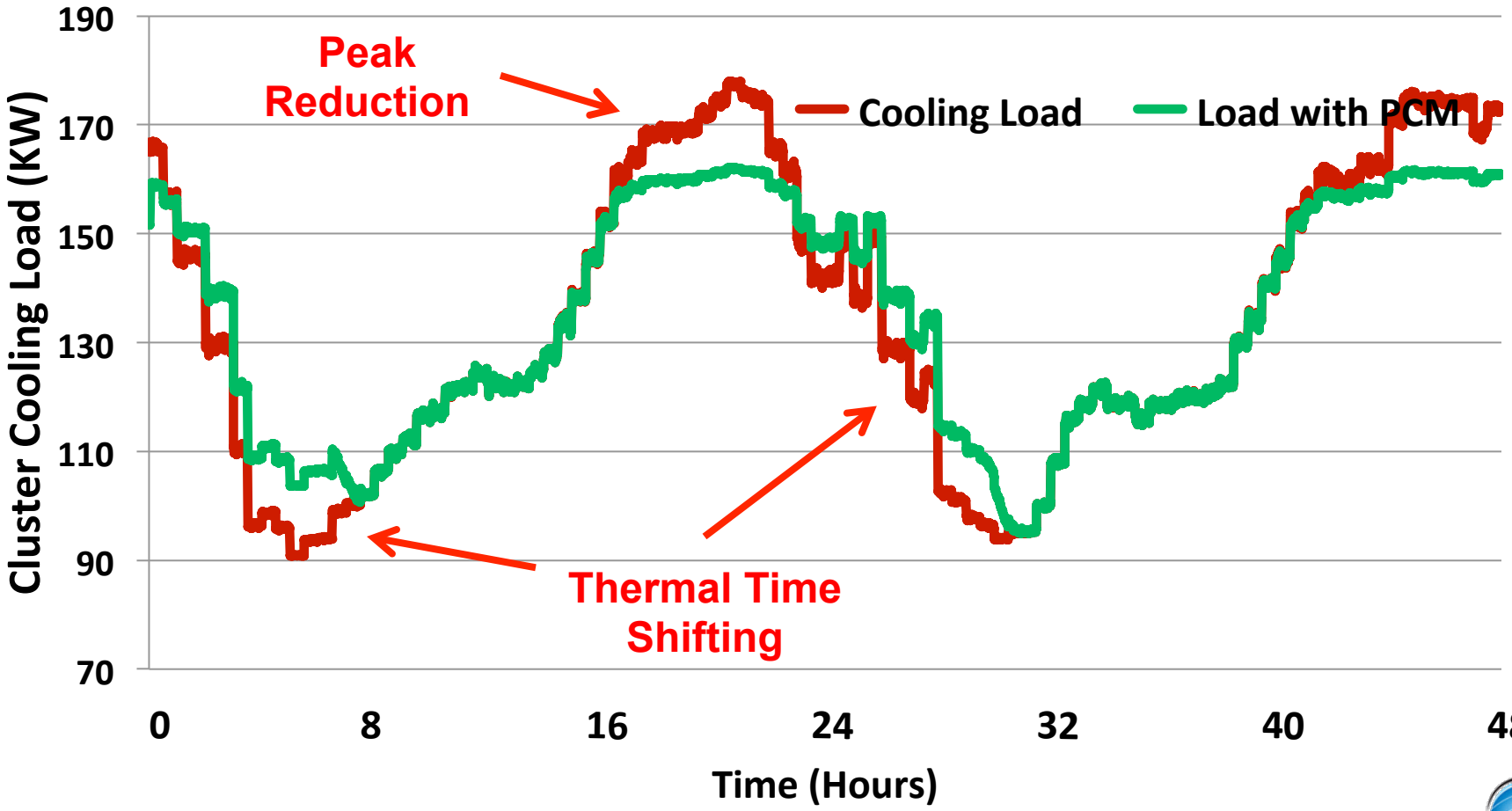


AMD

EVALUATION

PCM TO REDUCE COOLING SYSTEM LOAD

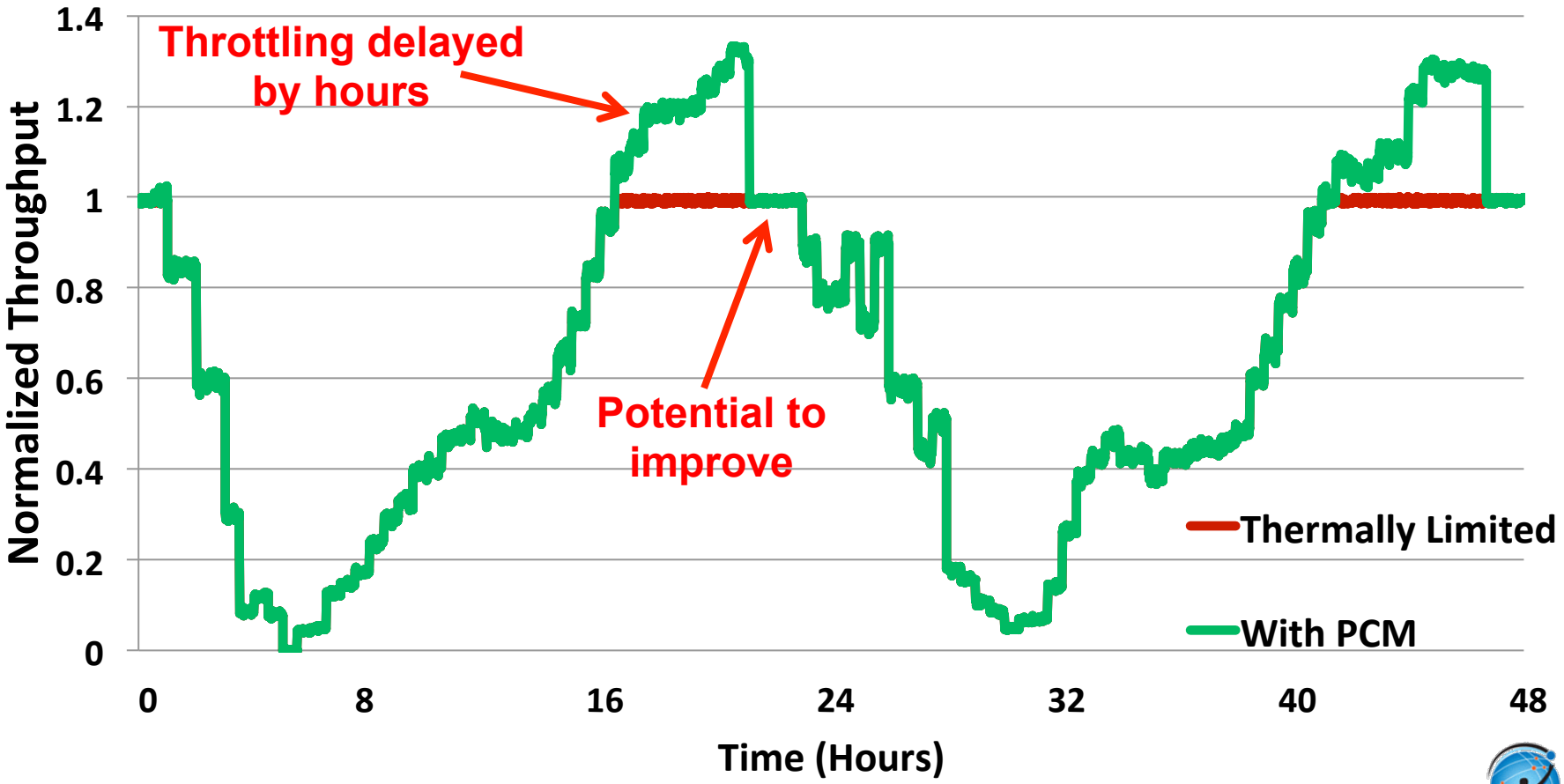
- PCM reduces the peak cooling load by ~10%
  - Using a smaller cooling system saves ~200K$ a year
  - Packing more servers saves ~3M$ in TCO



1U system results, lots more results in the paper

PCM IN OVER-SUBSCRIBED DATACENTERS

- PCM delays onset of thermal constraints
  - Server throttles ~5 hours later
  - 33% high throughput (69% for 3.1 hours best case)



1U system results, lots more results in the paper

AMD

CONCLUSIONS

UCSDCSE
Computer Science and Engineering

CONCLUSIONS

- Tight coupling of compute and cooling power in systems today
- Cooling systems limit compute at peak load, don't allow peak power increase and run in-efficiently for most of the time
- Phase Change Materials (PCM) on server is a mechanism to throttle temperature without throttling compute
- Possible to pack liters of PCM on server via server & PCM placement co-design
- Demonstrate a reduction in peak cooling load and thermally mandated throttling by several hours

DISCLAIMER & ATTRIBUTION

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION

AMD

BACKUP

COMPARISONS WITH EXISTING WORK

- Computational Sprinting vs. Thermal Time Shifting
  - Sprinting reshapes load without impacting thermals vs. reshaping of thermal profile with no change to load
  - Sprinting on chip over seconds vs. on server for hours
  - Changing on-chip / package very difficult vs. changing server design difficult

- Water tanks for energy storage vs. Thermal Time Shifting
  - Active solution vs. passive heat storage
  - Can be used together, reduced load on active solution
  - Need extra space, we fit within server