Building Datacenter Applications with Latency and Confidentiality Guarantees

Adrien Ghosn, Marios Kogias
Microsoft Research

Datacenters as a Systems Challenge I

- The other end of your smartphone
  - 1% world's electricity (Mesanet, Science 2020)
- Infrastructure:
  - Commodity servers
  - Fast Ethernet-based Clos topologies
  - Emerging hardware offloads
- Single administrative domain
  - Vertical integration & cross stack specialization

Datacenters as a Systems Challenge II

Workloads:

- Services e.g. websearch
  - Low **tail**-latency
  - Quality of Service
  - Service-Level-Objectives
- Public Cloud
  - Confidentiality
- General Requirements
  - High Throughput
  - Low Latency
  - Efficiency

Focus of the closing Swiss JRC Project

Focus of the new Swiss JRC Project

Agenda

- TTL-MSR: Taming Tail-Latency for µs-Scale RPCs
  - Funded by Swiss JRC 2019-2021

- Tyche: Confidential Computing on Yesterday's Hardware
  - Funded by Swiss JRC 2022-2024

TTL-MSR: Taming Tail-Latency for µs-Scale RPCs

EPFL: Marios Kogias, Edouard Bugnion
MSR Redmond: Irene Zhang, Dan Ports

Problem Statement

Goal: Build **microsecond-scale** and **tail-tolerant** datacenter systems

Problem 1: Microsecond-scale

- Fast emerging IO devices
- Existing software stack too generic and not build for microseconds

Problem 2: Tail-tolerant

- Complex fan-out/fan-in communication patterns
- Tail-at-Scale problem

Why is it a problem?

- Quality of user experience
- Consolidation and efficiency

Approach Overview

Step 1: Build the right tools

- Lancet: A Self-Correcting Latency-Measuring Tool [Usenix ATC 2019]

Step 2: Design new abstractions

- R2P2: Making RPCs First-Class Datacenter Citizens [Usenix ATC 2019]

Step 3: Unlock new abstractions and use new hardware

- HovercRaft: Achieving Scalability and Fault-Tolerance for microsecond-scale Datacenter Services [Eurosys 2020 Best Student Paper]
- Tail-Tolerance as a Systems Principle not a Metric [APNet 2020]

Step 4: Retrofit

- Bypassing the Load Balancer without Regrets [SoCC 2020]

Summary of Contributions

Conference and Workshop Papers

- Lancet: A Self-Correcting Latency-Measuring Tool [Usenix ATC 2019]
- R2P2: Making RPCs First-Class Datacenter Citizens [Usenix ATC 2019]
- HovercRaft: Achieving Scalability and Fault-Tolerance for microsecond-scale Datacenter Services [Eurosys 2020]
- Tail-Tolerance as a Systems Principle not a Metric [APNet 2020 (Workshop)]
- Bypassing the Load Balancer without Regrets [SoCC 2020]

Awards:

- Eurosys 2020 Best Student Paper
- Dennis Ritchie Doctoral Dissertation Award 2021
- Roger Needham Honorable Mention 2021

Retrospect

- Things that worked well
  - Freedom to work on a variety of topics
  - Access to new emerging hardware
  - New hires
  - Continuing collaborations

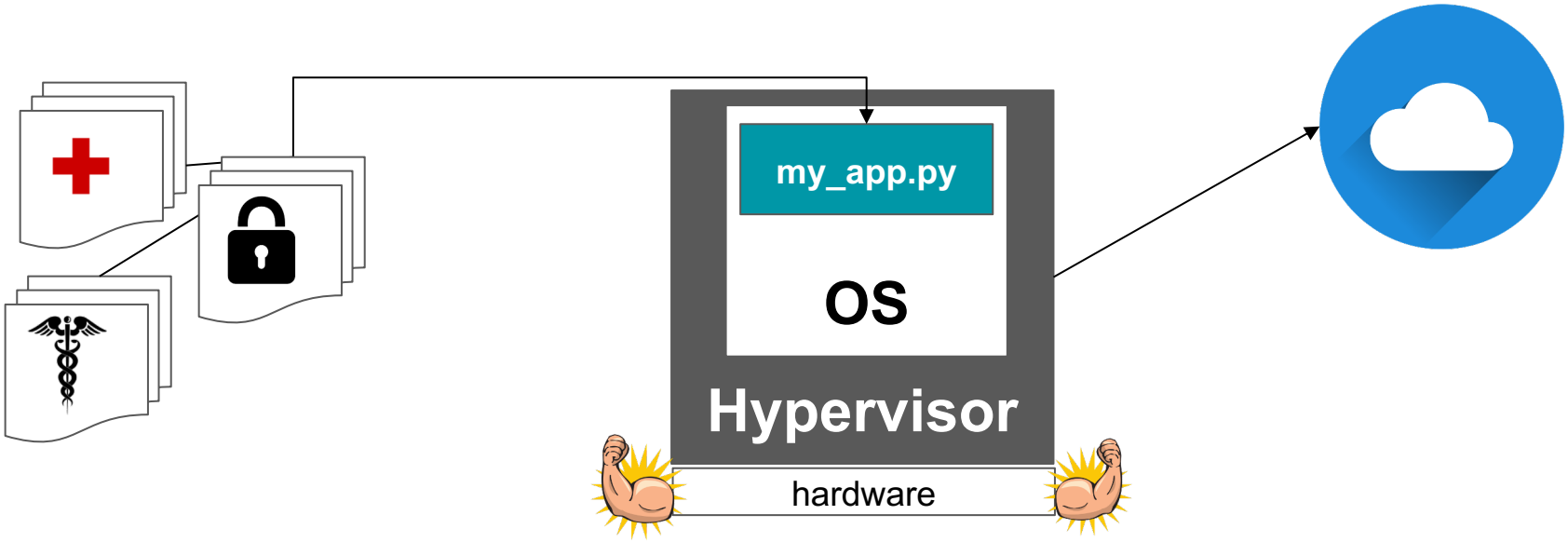
- Things we could improve
  - Concrete research plan from the beginning without much mentor input
  - Not very close collaboration with the mentors
  - No common publication
  - No direct interest by Microsoft to adopt the research output

Tyche

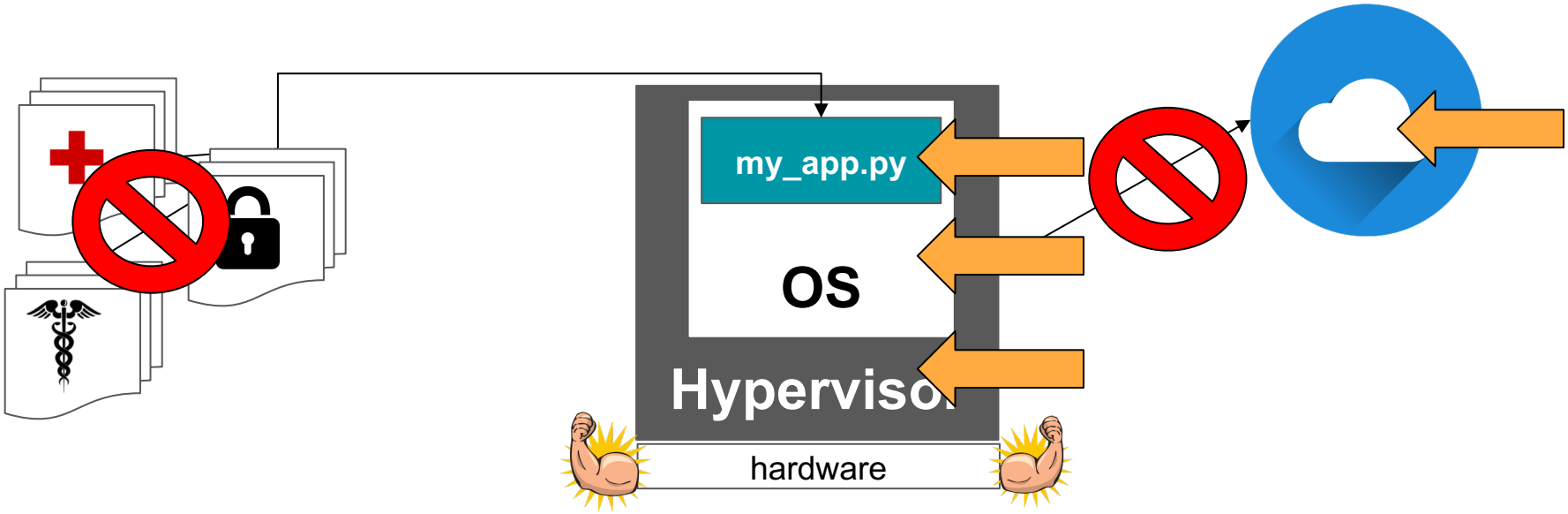
Confidential Computing on Yesterday's Hardware

MSR Cambridge : Adrien Ghosn, Marios Kogias
EPFL: Charly Castes, Edouard Bugnion, Mathias Payer, James Larus

Processing Sensitive Data

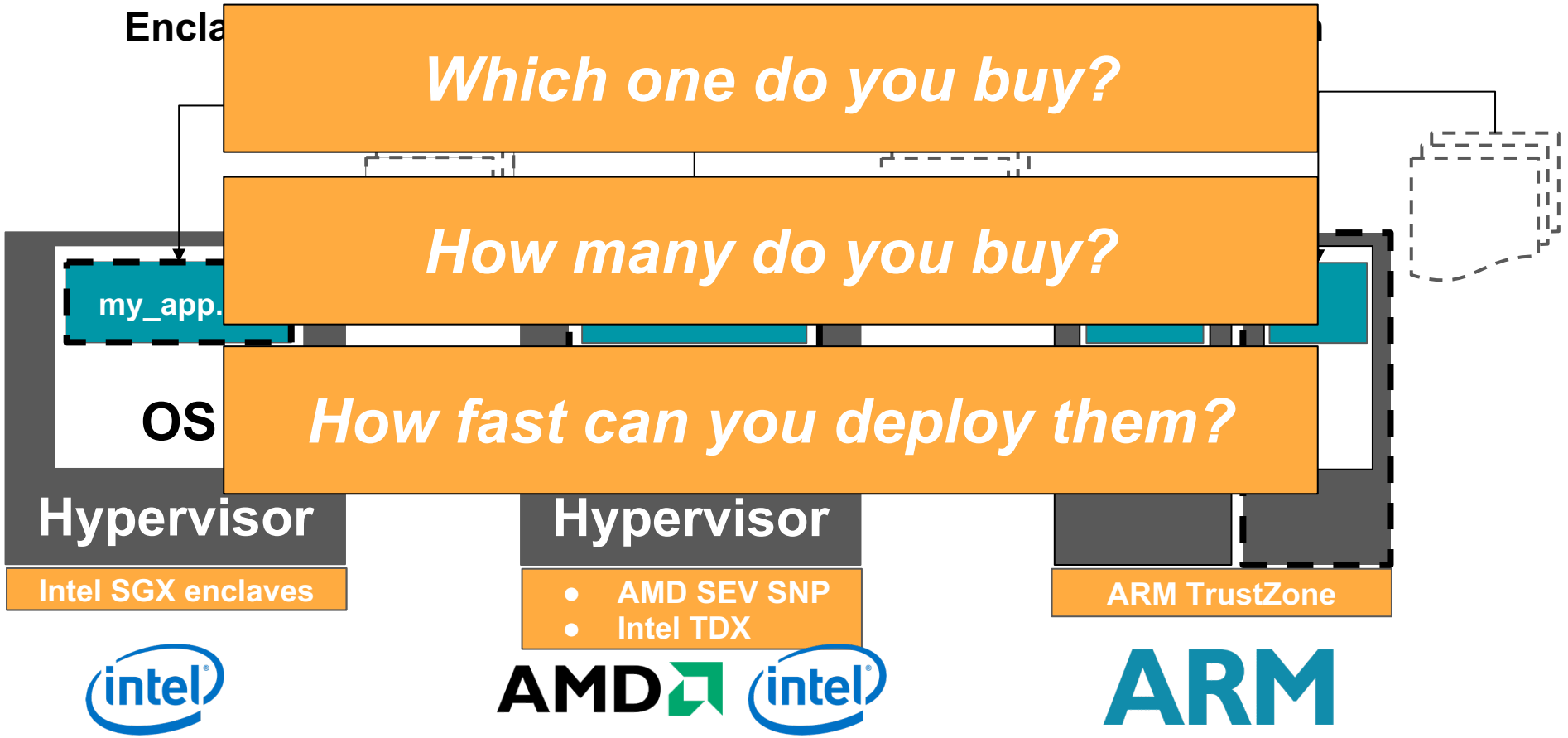


Who can access the data?



Too many people…

Enter Trusted Execution Environments (TEEs)



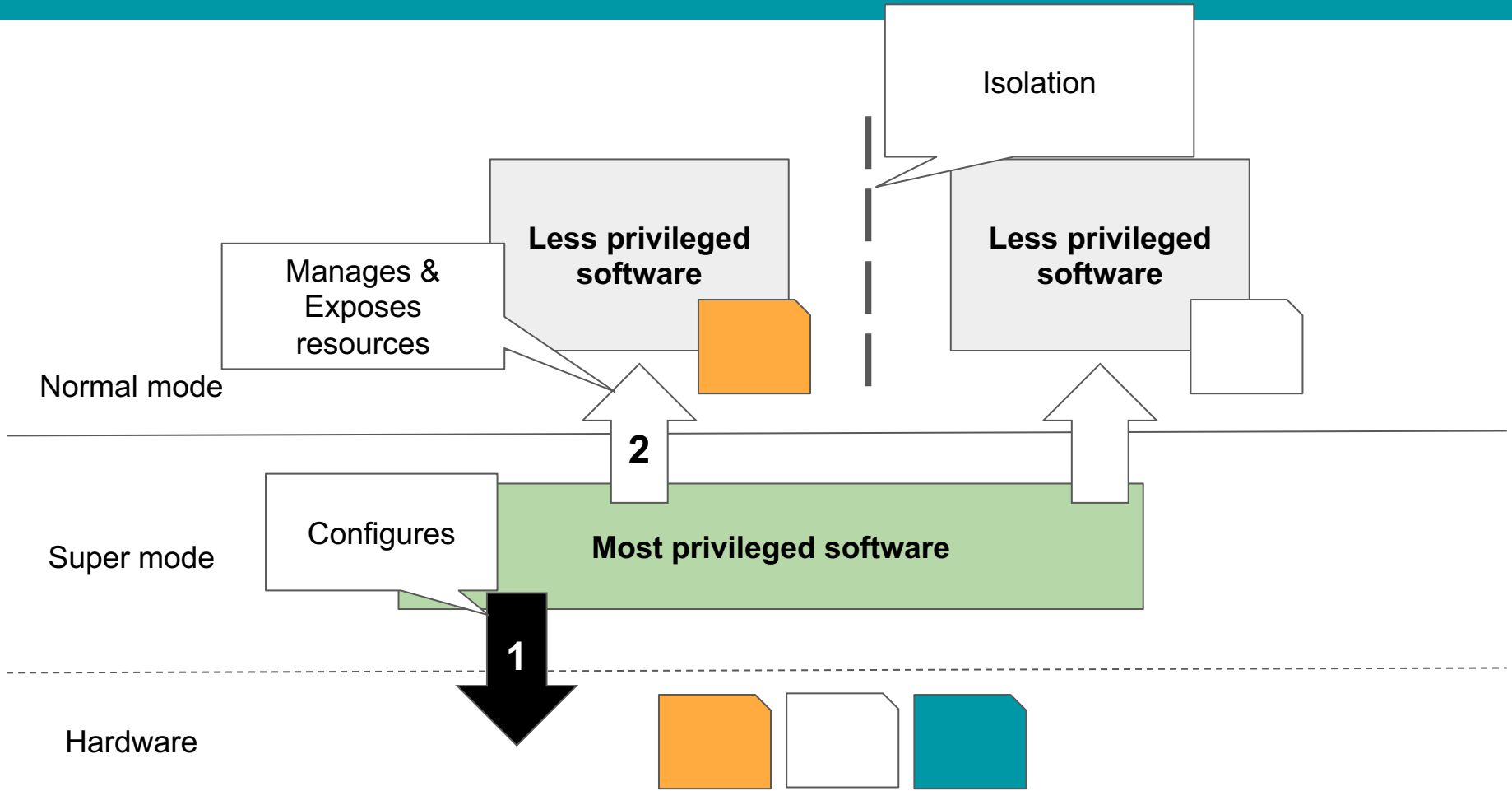
Our answer

- Do not buy any of them.
  - We have everything we need.
  - We can do better than hardware.

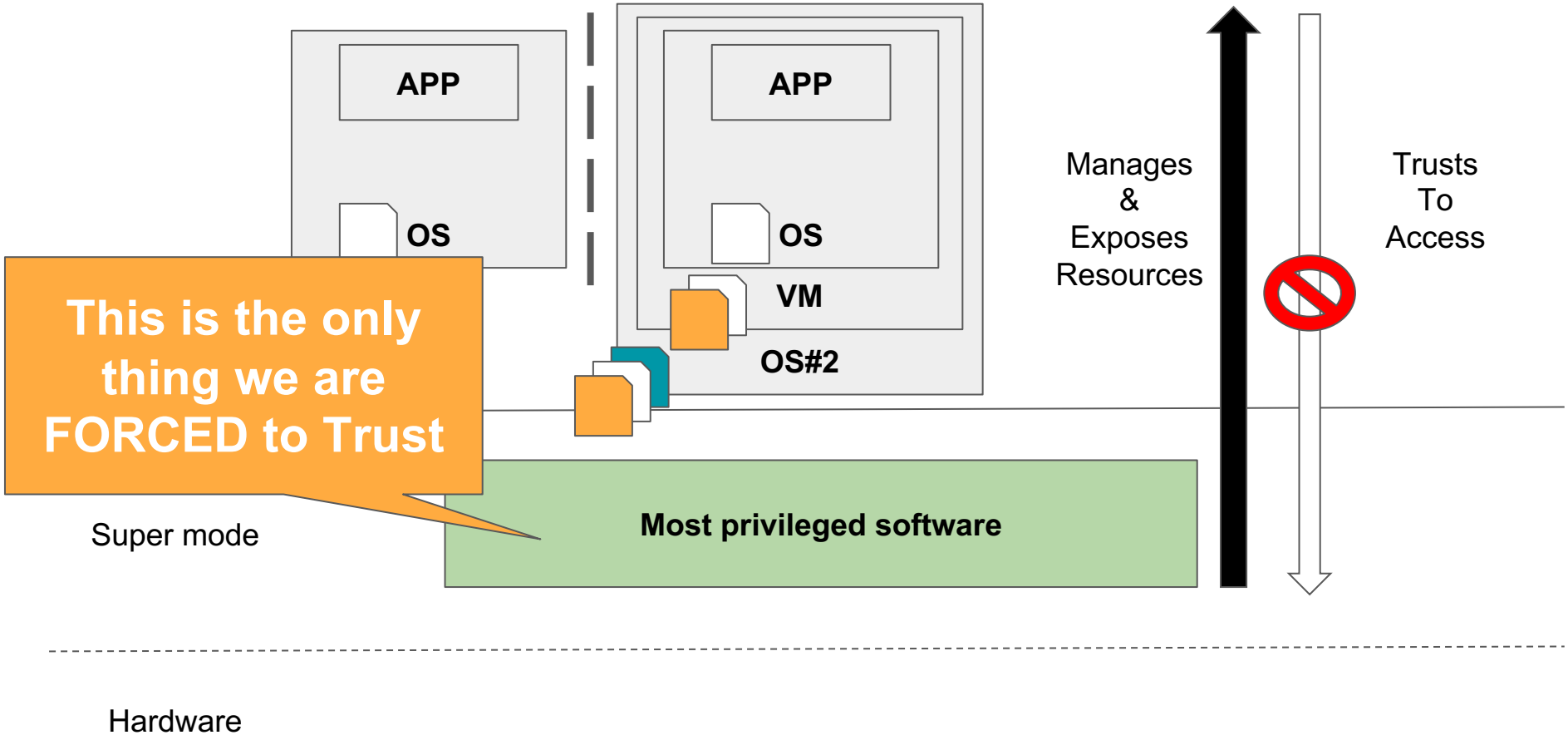
Better as in …

- Subtler notion of trust.
- Any programming abstraction.
- Any combination of programming abstractions.

What the real problem is…



What the real problem is…



Wait Wait Wait…

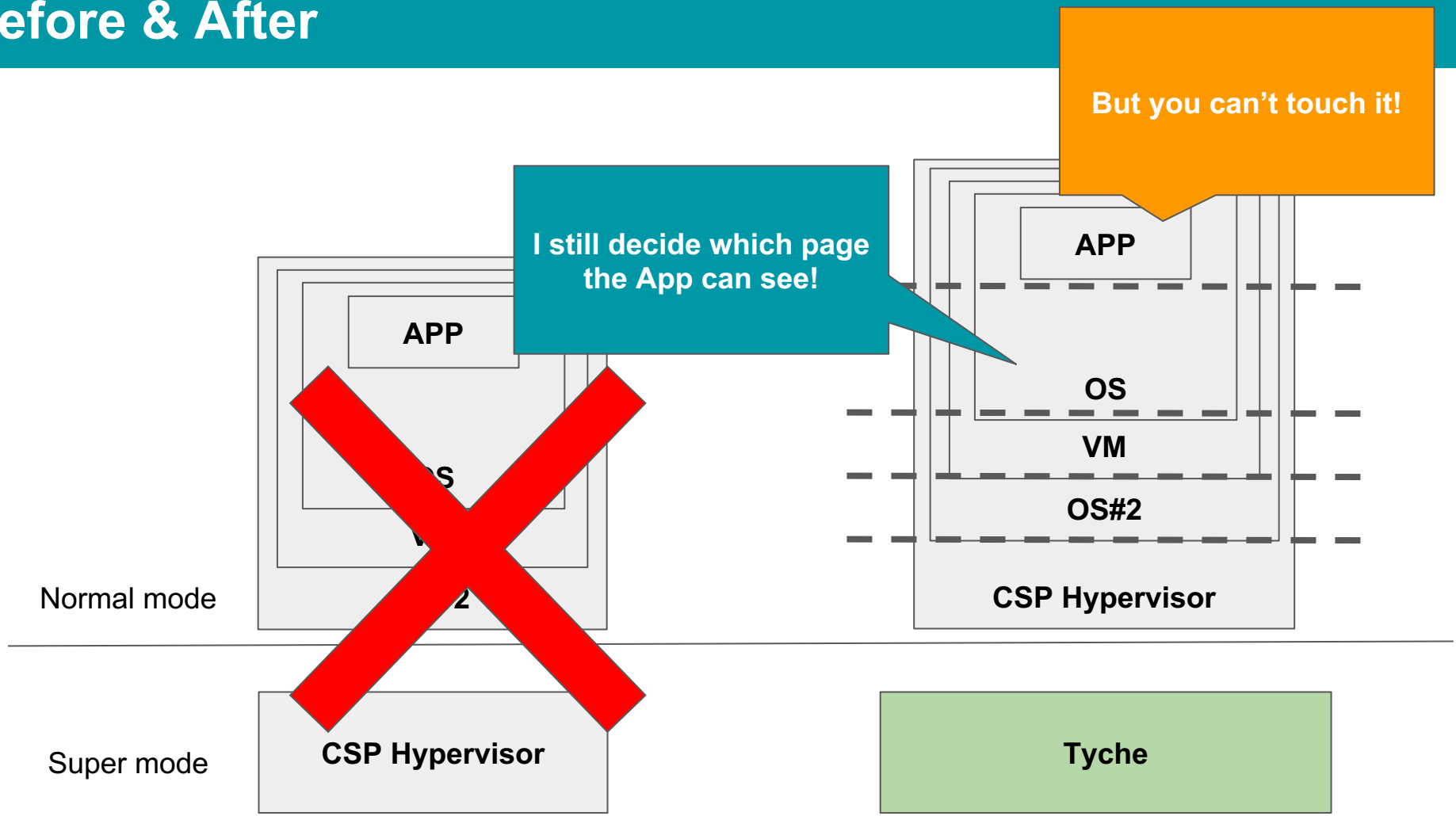
- W
  - … for now.

De-Privilege the Hypervisor

Let's summarize:

- We need to trust the most privileged software.
- The most privileged software is the hypervisor.
- We cannot trust the hypervisor…

Before & After



Tyche

- Tyche
  - Formally verify the monitor.
  - Negotiation protocol between manager & client.
  - Extension to Popek & Goldberg.
- Hardware features
  - TPM attestation.
  - Memory encryption.
  - Hardware accelerators.