ALPS: Agile Power Management For Future Energy-Efficient Servers In Data Centers Running Latency-Critical Applications

Haris Volos, Yiannakis Sazeides, Georgia Antoniou
Department of Computer Science
University of Cyprus

Jawad Haj-Yahya
Rivos Inc

Πανεπιστήμιο Κύπρου
University of Cyprus

January 21, 2025

Rivos

Who we are

Haris Volos
Assistant Professor
University of Cyprus

Yiannakis Sazeides
Professor
University of Cyprus

Georgia Antoniou
PhD Student
University of Cyprus

Jawad Haj-Yahya
Principal Architect
Rivos Inc.

Acknowledgements

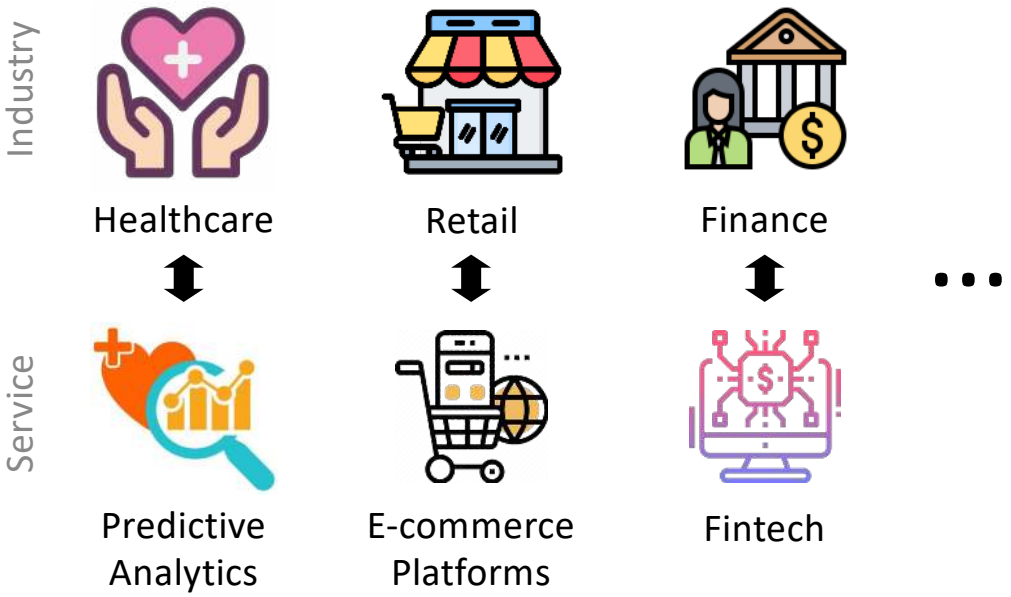
- Collaborators
  - Davide B. Bartolini
  - Zhirui Chen
  - Ye Geng
  - Kleovoulos Kalaitzidis
  - Jeremie S. Kim
  - Marios Kleanthous
  - Ziwei Li
  - Onur Mutlu
  - Markos Othonos
  - Tom Rollet
  - Zhe Wang

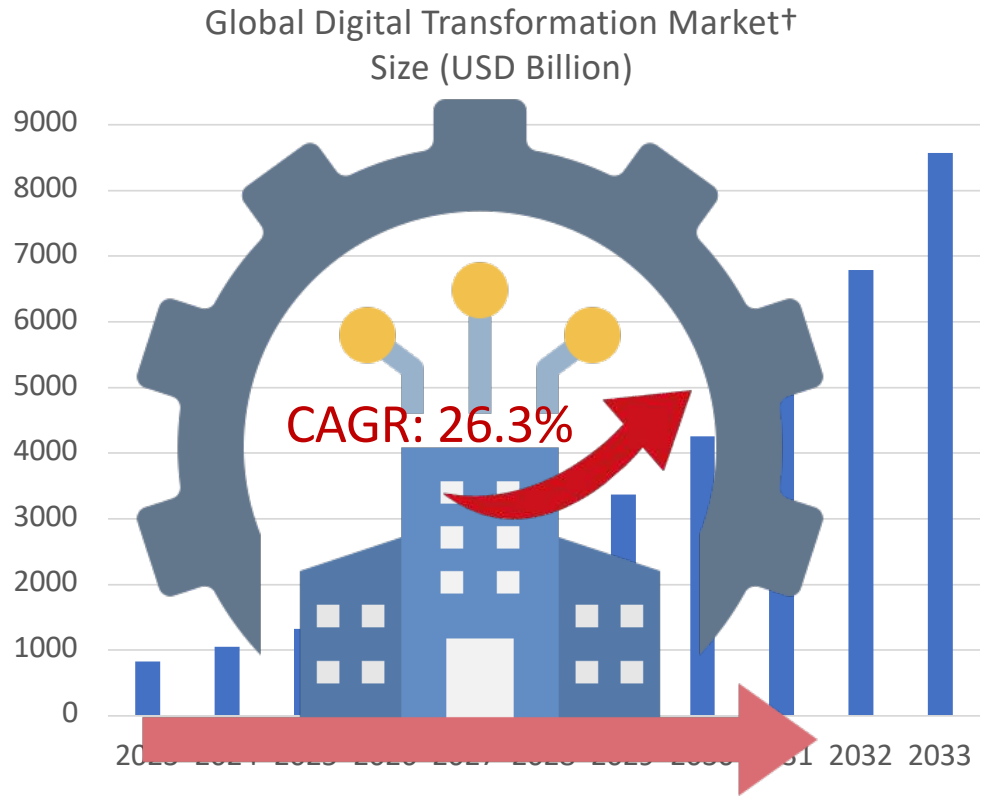
- Funding

- Infrastructure

Digital Transformation

Many industries increasingly transformed by digital services





Exponential demand for digital services

† Source: market.us

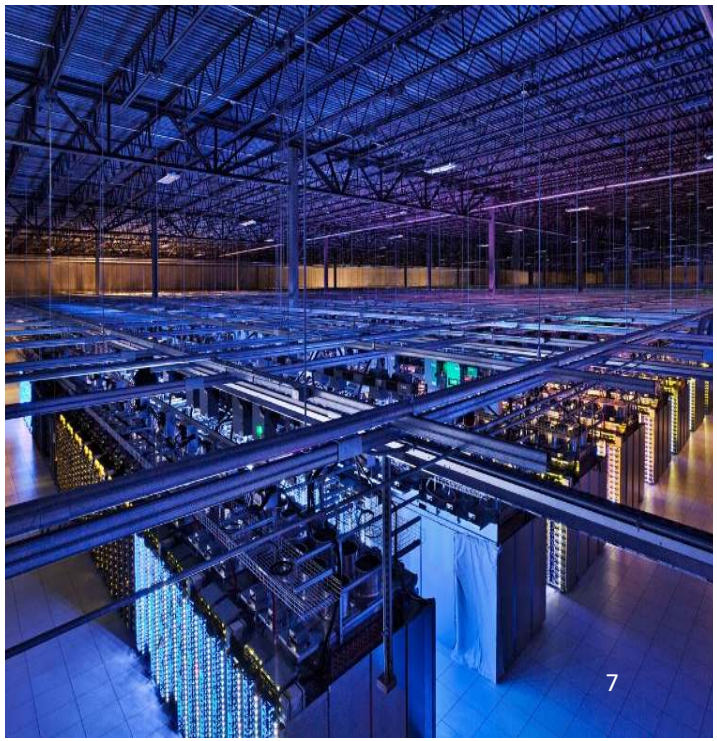
Datacenters are the backbone of Modern Applications

Datacenters are the backbone of Modern Applications

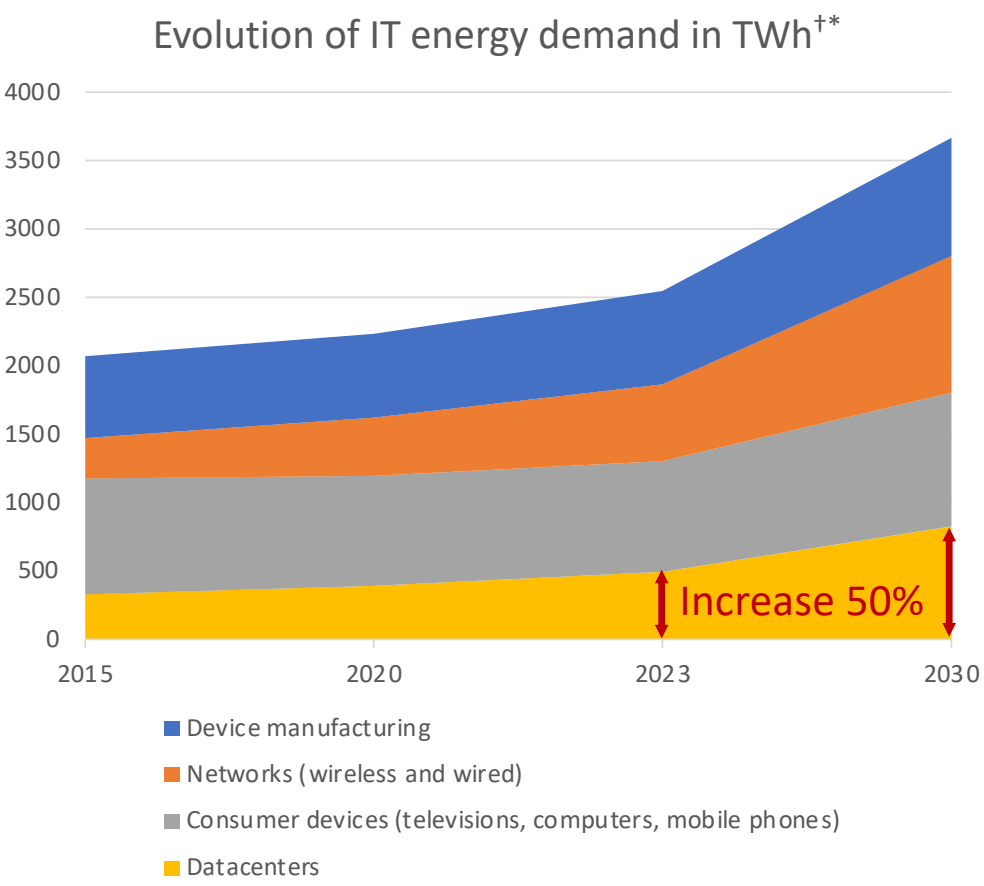
- Comprise 10,000 to 100,000 servers connected through a high-bandwidth network

Datacenters are the backbone of Modern Applications

- Comprise 10,000 to 100,000 servers connected through a high-bandwidth network
- Provide user-facing applications
  - Search, social networking, online shopping, video sharing
- Emphasize cost-efficiency
  - Attention to power and energy efficiency



Datacenters are becoming more power hungry



† Source: Babak Falsafi, What's hot? Post-Moore datacenter architecture, HotInfra, 2023
* Source: Digital economy & climate impact, Schneider Electric May 2022

NuclearNewswire

Amazon buys nuclear-powered data center from Talen

Thu, Mar 7, 2024, 3:01PM | Nuclear News

Susquehanna nuclear plant in Salem Township, Penn., along with the data center in foreground. (Photo: Talen Energy)

DCD > Channels > Media > Live Events > Academy > Industry Awards

Google CEO: We're working on 1GW data centers, seeing money going into SMRs

September 23, 2024 By: Sebastian Moss Have your say

npr

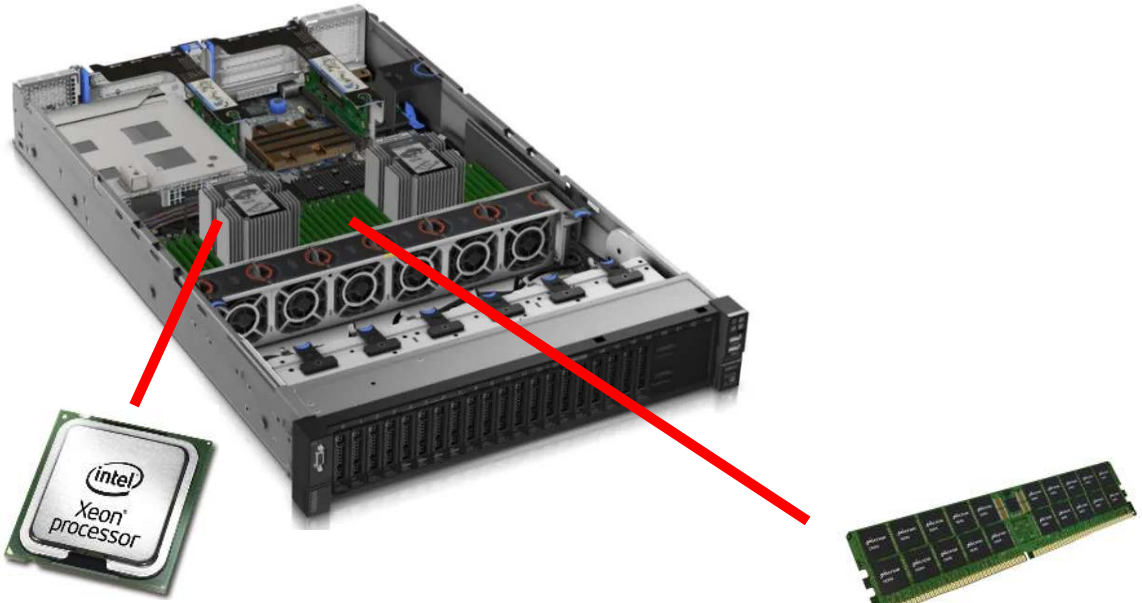
NATIONAL

Three Mile Island nuclear plant will reopen to power Microsoft data centers

SEPTEMBER 20, 2024 · 1:40 PM ET

By C Mandler

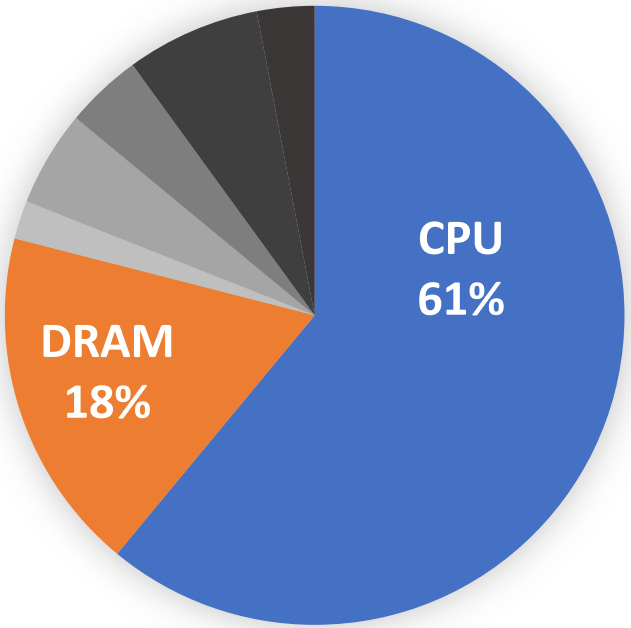
Servers are a critical factor in datacenter power consumption



CPU for General Purpose Workloads

Main memory for fast data access

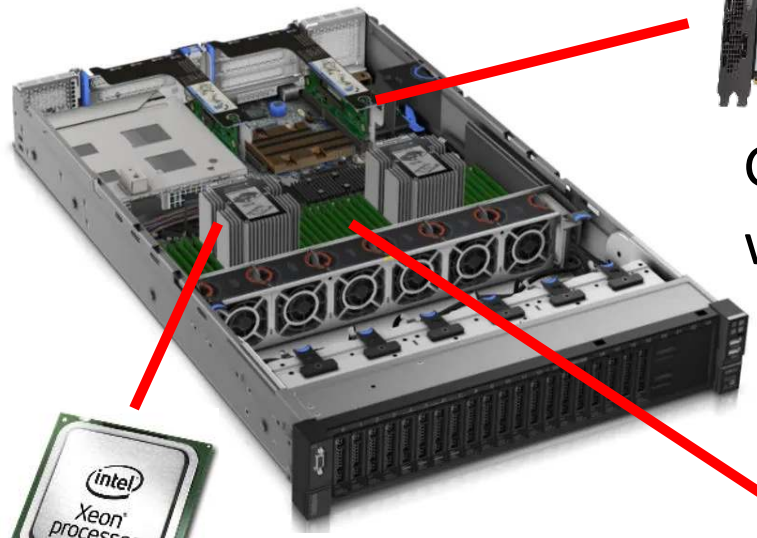
Datacenter Power Consumption



c. 2018 †

† Source: Barroso et al. The Datacenter as a Computer: Designing Warehouse-Scale Machines, Third Edition, Springer Nature, 2019

Servers are a critical factor in datacenter power consumption

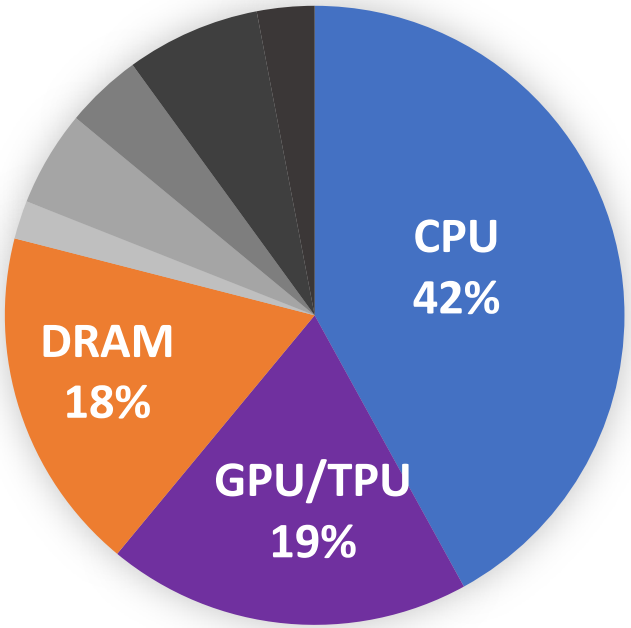


GPU/TPU for AI workloads

CPU for General Purpose Workloads

Main memory for fast data access

Datacenter Power Consumption

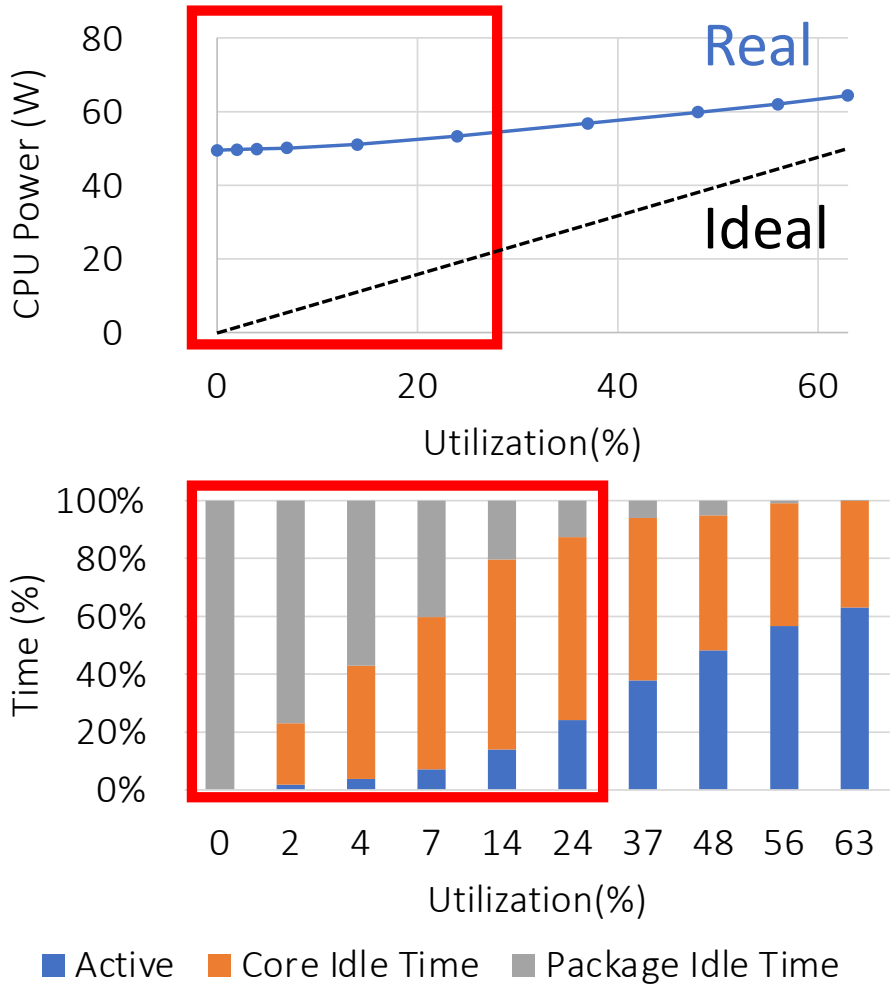


c. 2028 † (Projection)

† Disclaimer: Projections are based on analyst expectations that AI will account for 19% of data center power demand (Goldman Sachs, 'AI is poised to drive 160% increase in data center power demand', 2024). Assuming AI is handled by GPU/TPU, CPU power consumption is projected to decrease from 61% to 42% .

Modern CPUs are not Energy Proportional

- Energy proportionality: energy usage scales with load
- High power consumption at low utilization, typical for interactive workloads
- But ample opportunity for power savings from idle resources



Motivating Workload: E-commerce Platform

Our Focus

Interactive (user-facing) component

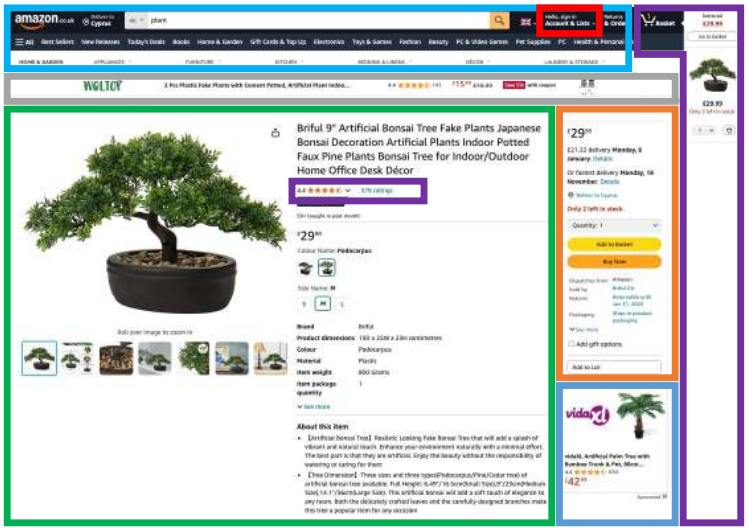
- Product catalog, shopping cart, etc
- O(ms) request latency (tail and avg)
- Low Server Utilization (5-25%) for handling bursty workload

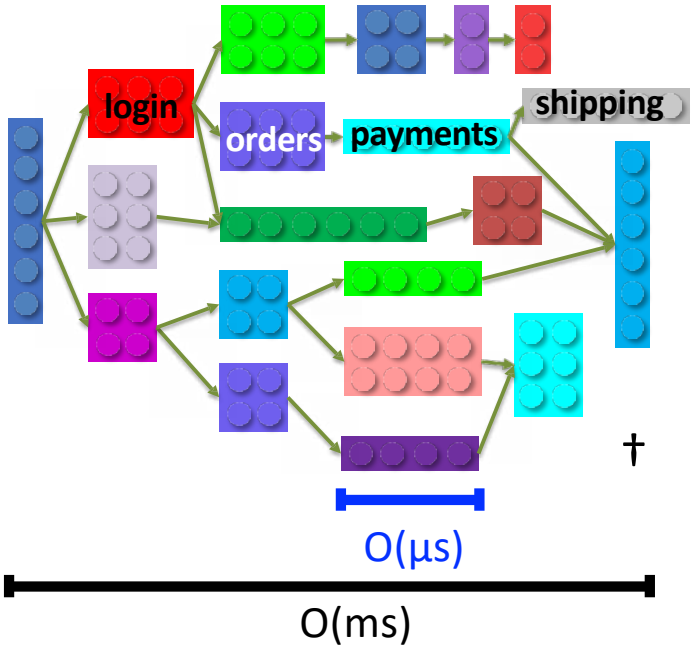
Batch (background) component

- Data analytics, ML model training, etc
- High throughput
- High Server Utilization (50-80%) for efficiency

Microservices amplify latency challenges

Applications comprise multiple microservices for modularity and scalability





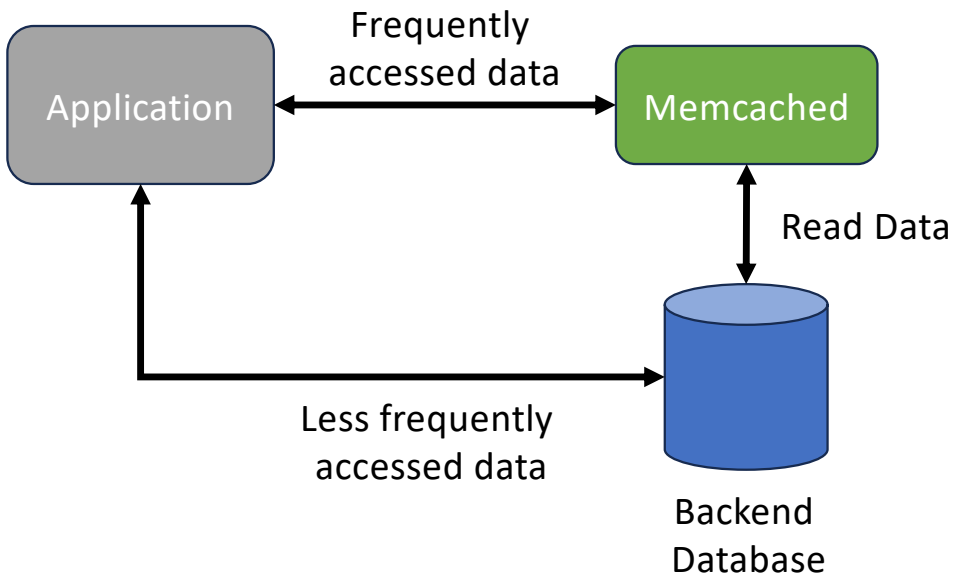
Microservice service time ranges from 10 to 100 µs,
making them highly sensitive to killer-microsecond overhead

† Source: Gan et al., An Open-Source Benchmark Suite for Microservices and Their Hardware-Software Implications for Cloud & Edge Systems, ASPLOS, 2019

Microservice Example: Memcached

- In-memory key-value store

- Deployed as a low-latency (~100us) caching service
  - Meta (memcache)
  - Twitter (twemcache)
  - …



Deep idles states cause killer-microsecond overheads

- Idle states are power saving states that enable the core to reduce its power consumption during idle periods
- Intel's Skylake architecture idle core states:
  - C0 – Active – no power saving
  - C1 – Shallow
  - C1E – Medium
  - C6 – Deep – wakeup latency = 133us

Performance overhead of deep C-states is in the same order as the latency target of microservices

Service providers may often disable deep C-states to maintain latency target

Goal

- Improve the energy proportionality of servers running user-facing latency-critical applications

- AgileWatts:
  - C6A: Deep idle Core C-state with a nanosecond-scale transition latency
  - C6AWarm: Deep idle Core C-state with a nanosecond-scale transition latency and no cold-start latency

- AgilePkgC:
  - PC1A: Deep idle Package C-state with a nanosecond-scale transition latency

Tutorial Outline

Time	Topic
14:00 – 14:15	Welcome, Motivation, and Tutorial Outline
14:15 – 14:45	Background: Server Power Management
14:45 – 15:30	Agile Core C-state Architecture
15:30 – 16:00	**Coffee break**
16:00 – 16:30	Agile Package C-state Architecture
16:30 – 16:45	Memory Energy Efficiency
16:45 – 17:15	Tools Demonstration
17:15 – 17:30	Future Directions, Open Discussion and Wrap-up