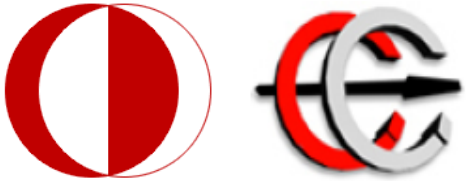


Cloud Computing and Hardware Accelerated Clouds

Ece Güran Schmidt

METU Electrical and Electronics Engineering

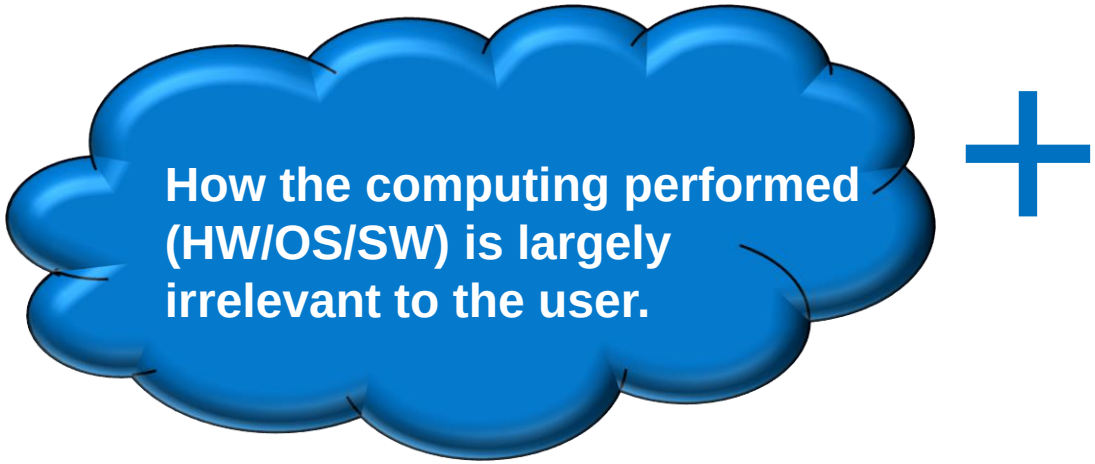
Hello ☺

- I work in the field of computers in Electrical and Electronics Engineering in METU
- Teaching: Logic Design, Data Structures, Computer Networks, Computer Architecture
- (Some) Research Interests: Computer Networks, Real-time and Embedded Systems, Hardware Accelerated Cloud Computing

Overview of the talk

- Part I: Cloud Computing
- Part II: Hardware Acceleration
- Part III: Hardware Accelerated Clouds
- Part IV: ACCLOUD Research Project

Part I: Cloud Computing



Legacy Definitions of Cloud Computing

"A model for enabling, ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources"

- Resources: (networks, servers, storage, applications, and services)
- Can be rapidly provisioned and released with minimal management effort or service provider interaction.

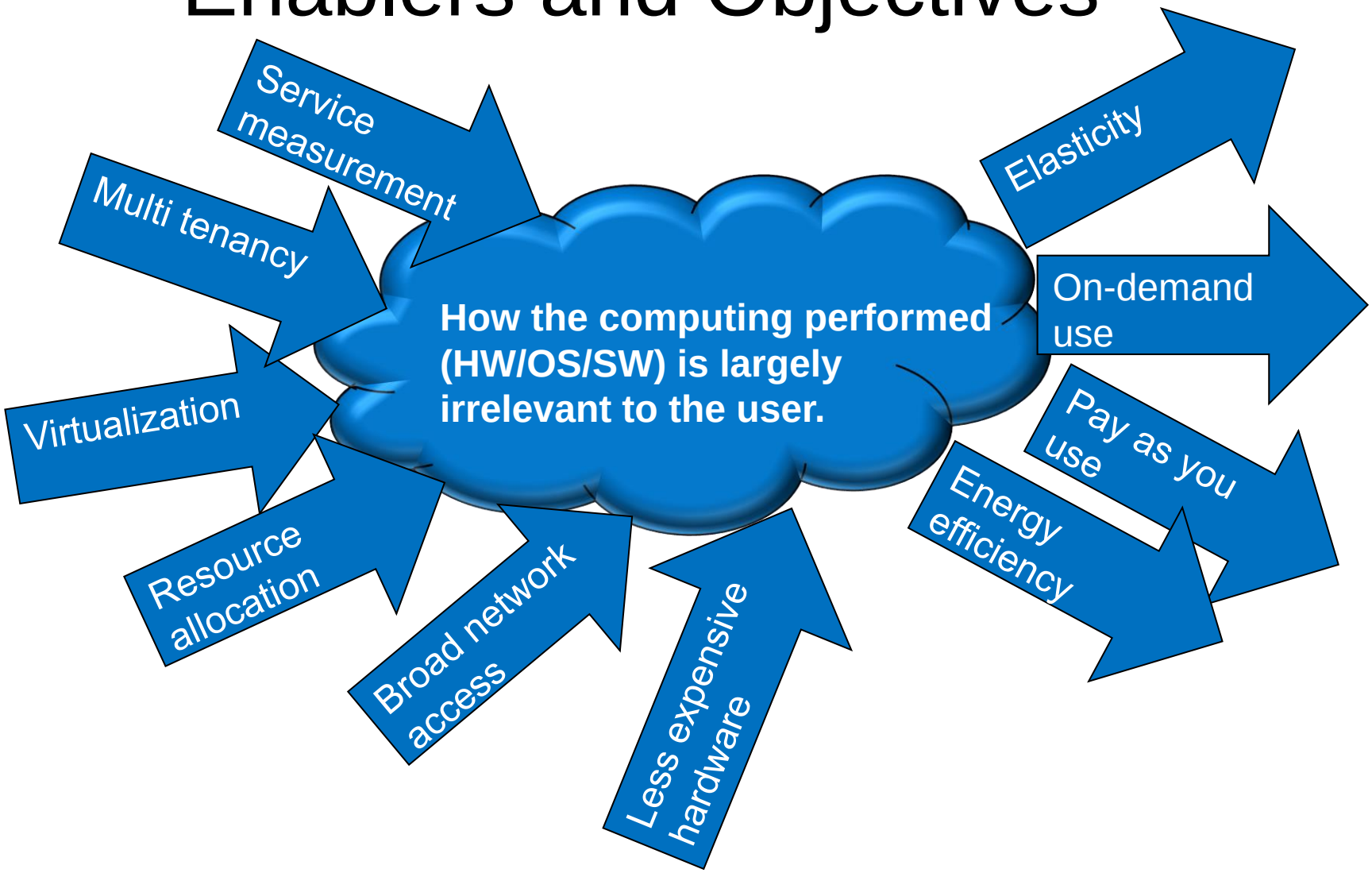
https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

"The applications delivered as services over the Internet and the hardware and systems software in the data centers that provide those services."

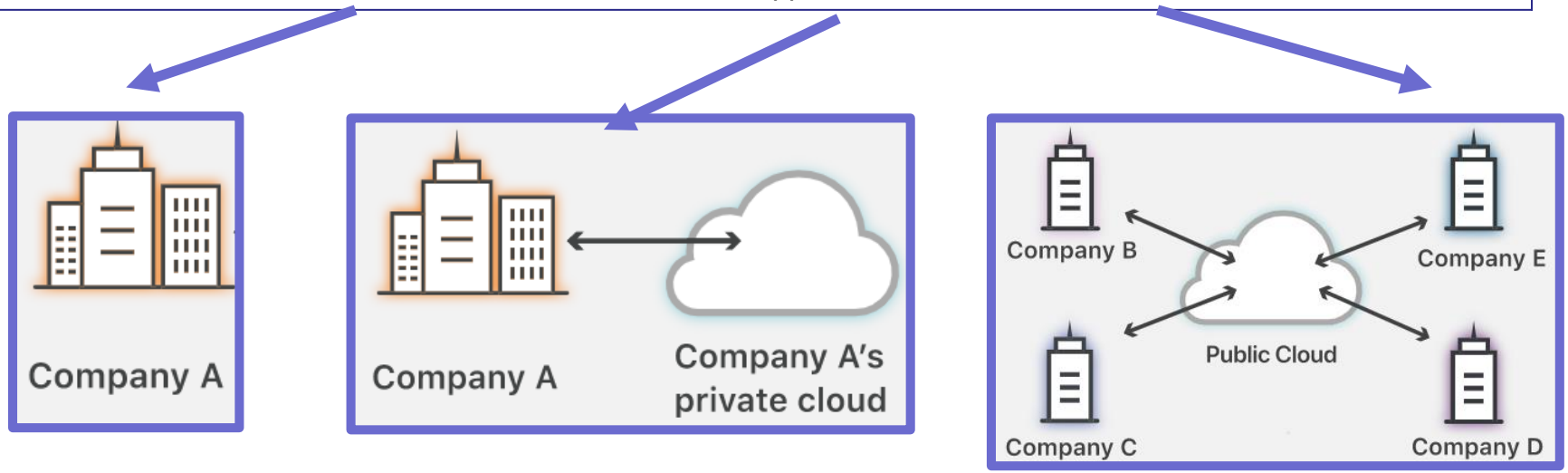
Armbrust, Michael, et al. "A view of cloud computing." *Communications of the ACM* 53.4 (2010): 50-58. (11846 citations)

Enablers and Objectives

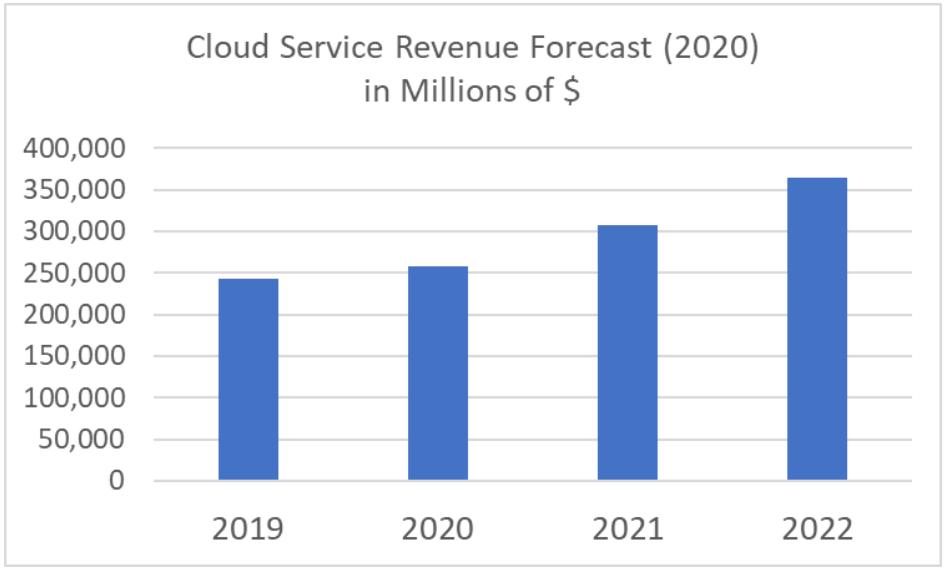


Computing Models

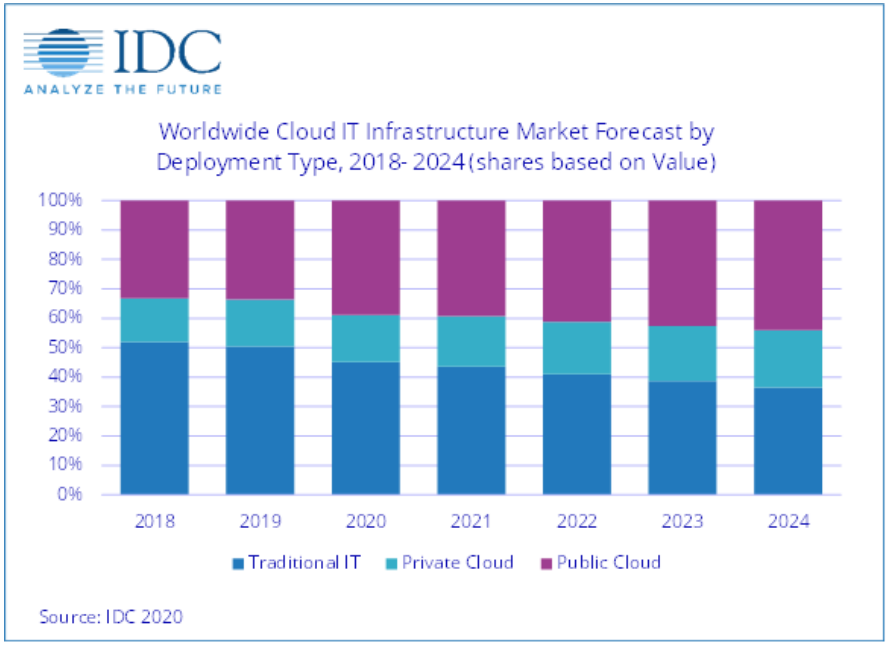
	On-Premise (Traditional IT)	Private Cloud	Public Cloud
HW Infrastructure Ownership	Company	Company	Service Provider
Software Ownership	Company	Company	Service Provider/Company
Maintenance	Company	Company	Service Provider/Company
Resource allocation	Fixed allocation to Company and applications	Fixed allocation to Company Elastic allocation to the applications.	Elastic allocation to Company and applications.



Economy

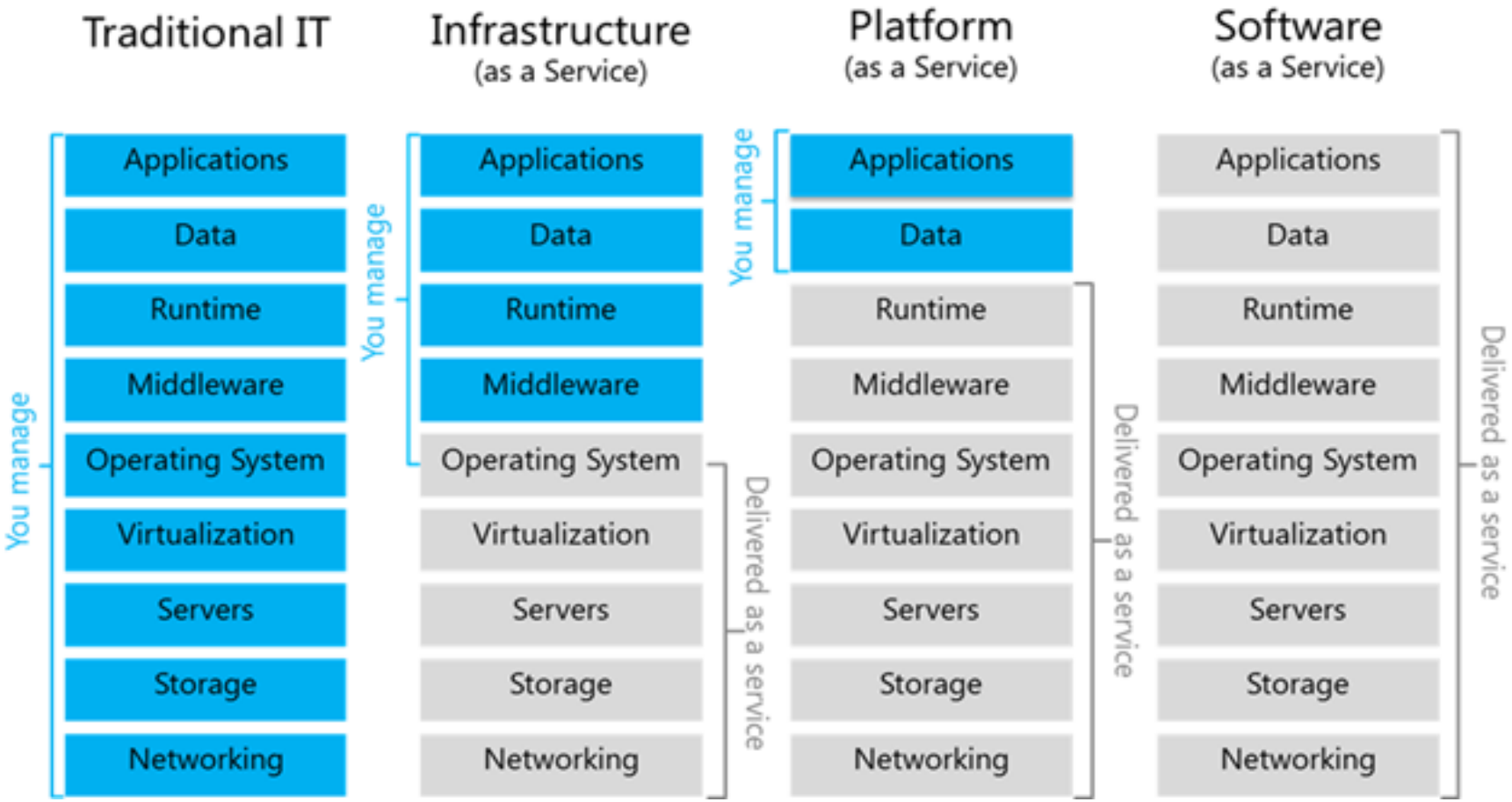


Worldwide Public Cloud Service Revenue Forecast (Millions of U.S. Dollars)
https://www.gartner.com/
Date: 2020-07-23

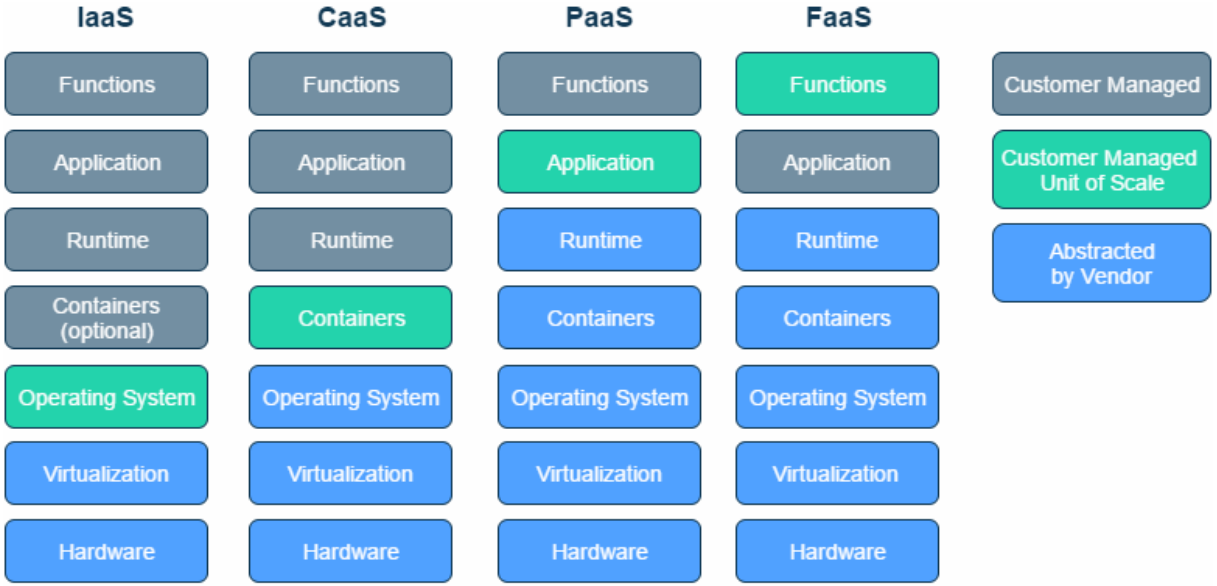


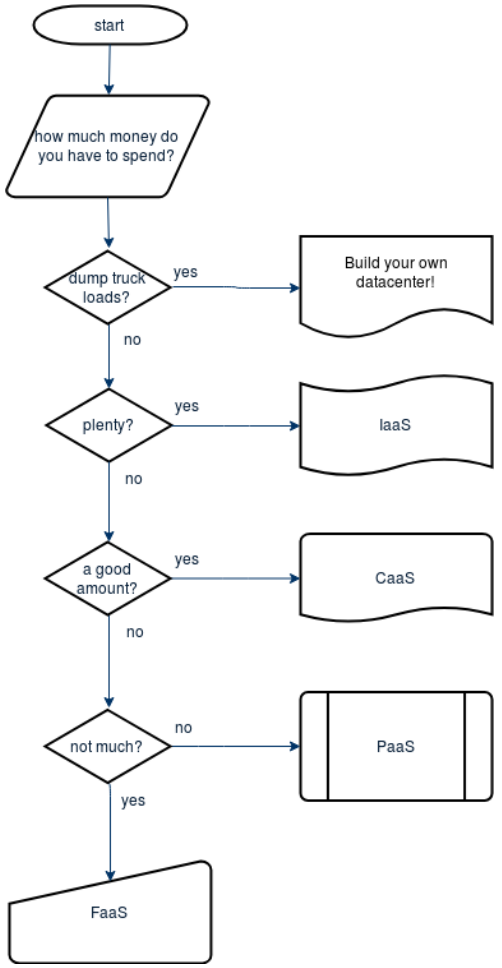
https://www.idc.com/getdoc.jsp?containerId=prUS46895020

Cloud Services (Legacy Breakdown)



Cloud Services (More Contemporary)





https://serverless.zone/abstracting-the-back-end-with-faas-e5e80e837362

https://developer.ibm.com/articles/when-to-use-iaas-faas-paas-and-caas//

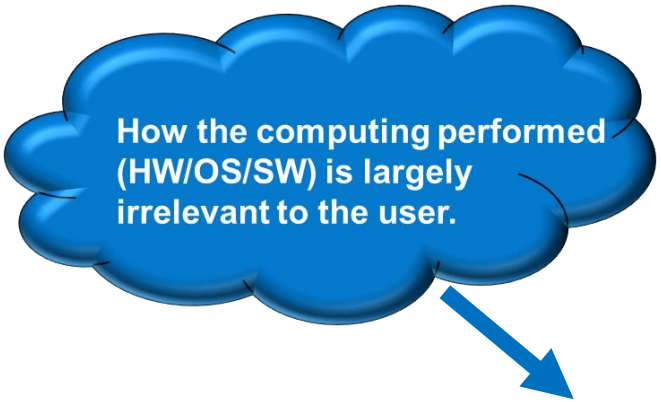
Cloud Data Center

- Traditional Data Center
  - a single physical facility with all hardware infrastructure and equipment
  - Houses all data and applications

https://www.cisco.com/c/en/us/solutions/data-center-virtualization/what-is-a-data-center.html#~types-of-data-centers

- Cloud Data Center
  - it's all online!
  - **cloud servers host data and applications**
  - Data automatically gets fragmented and duplicated across various locations for secure storage.

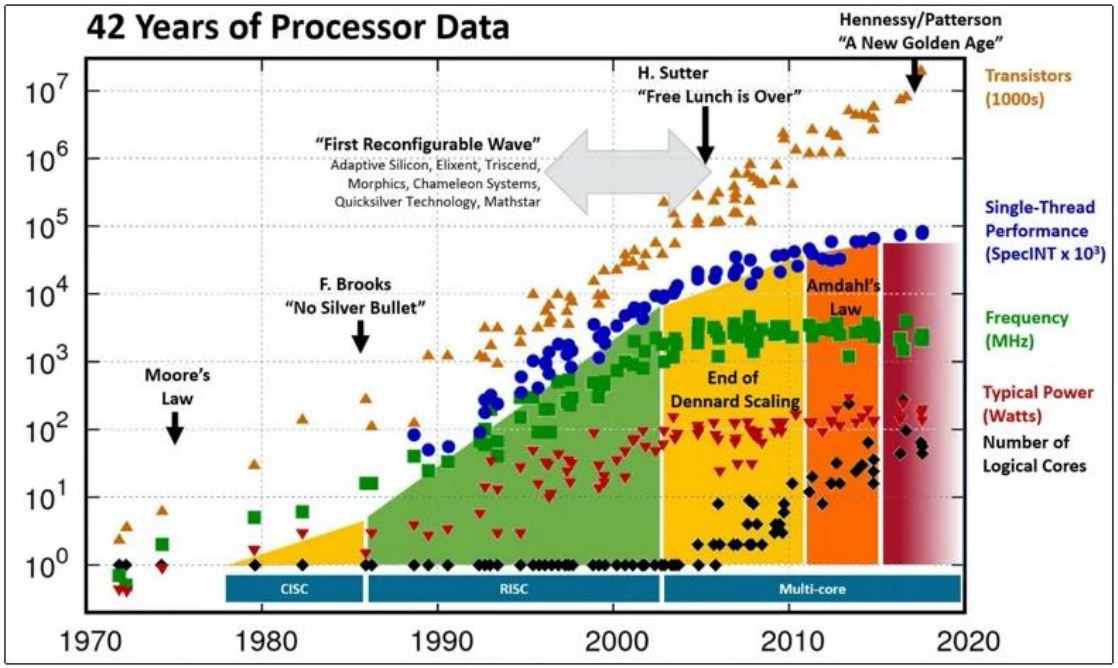
Part II: Hardware Accelerators



Xilinx Versal ACAP Adaptive Compute Acceleration Platform

https://www.xilinx.com/products/silicon-devices/acap/versal.html

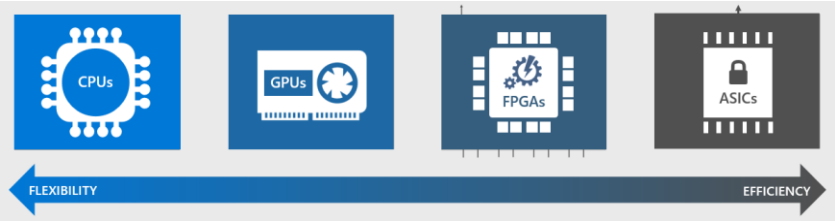
State of Computing

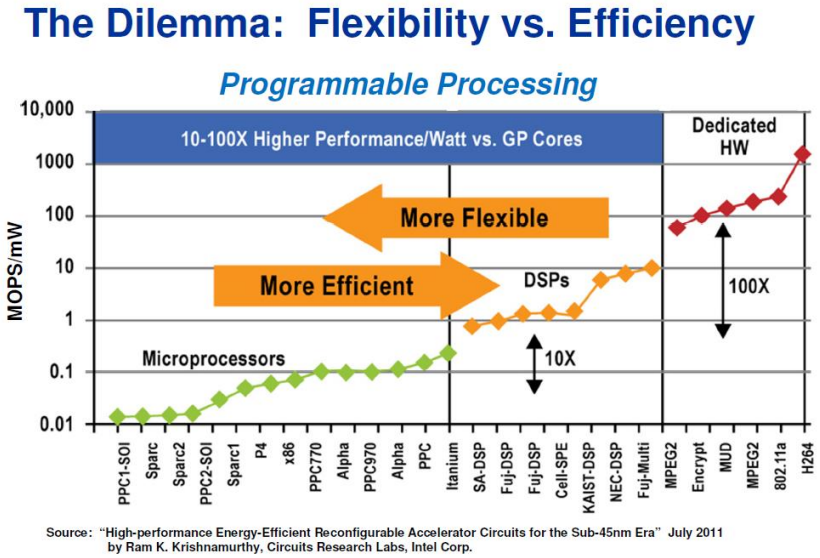


https://iscaconf.org/isca2018/docs/HennessyPattersonTuringLectureISCA4June2018.pdf

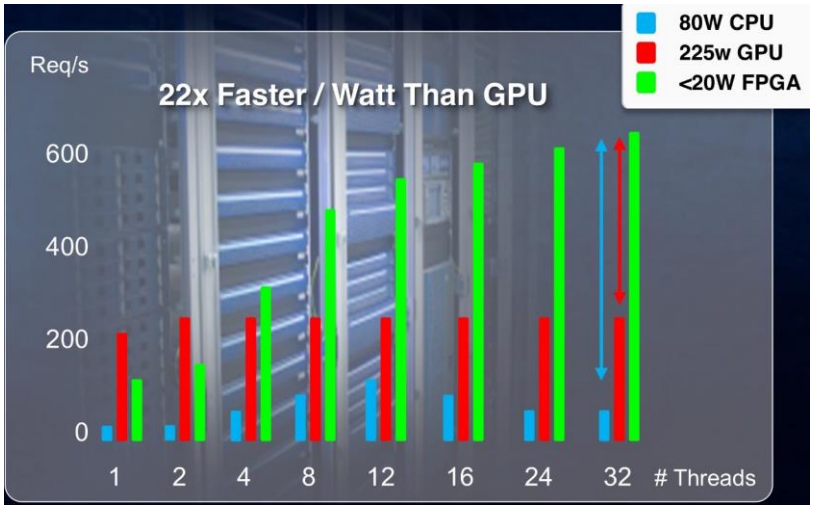
- Moore's law ends➔ Thermal constraints
- Dennard's scaling ends➔ Gains from multiprocessor architectures slow down
- Henessy &Patterson 2018 Turing Lecture's solution: Domain Specific Architectures➔ Hardware Accelerators

Hardware Accelerators





- Specialized hardware instead of general purpose hardware
- Performance and energy-efficiency improvements

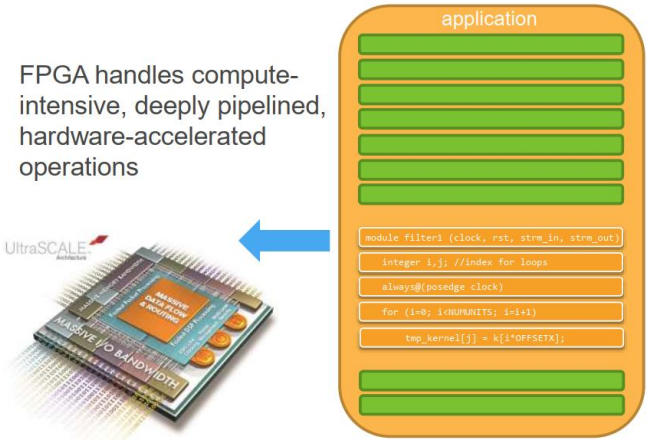


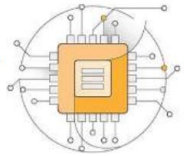
[Source: Xilinx, 2016]

Kachris, Christoforos, and Dimitrios Soudris. "A survey on reconfigurable accelerators for cloud computing." *2016 26th International conference on field programmable logic and applications (FPL)*. IEEE, 2016.

Hardware Accelerators: Development and Use

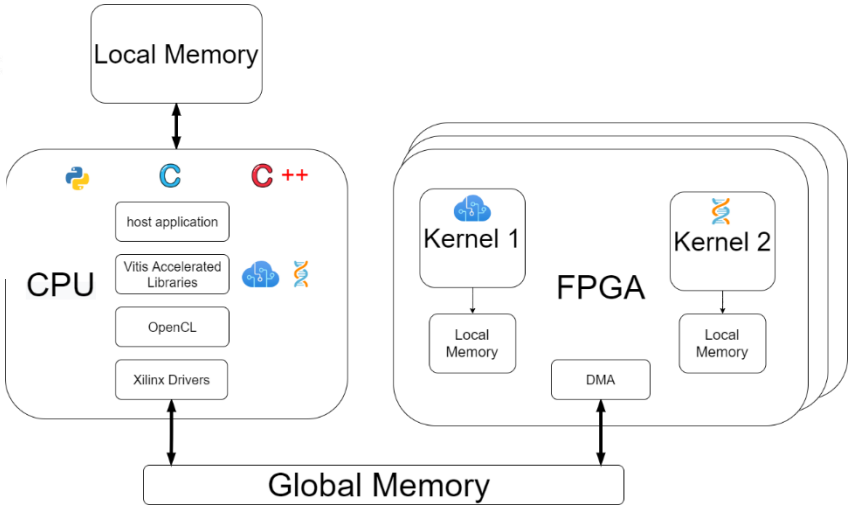
FPGA handles compute-intensive, deeply pipelined, hardware-accelerated operations





CPU handles the rest

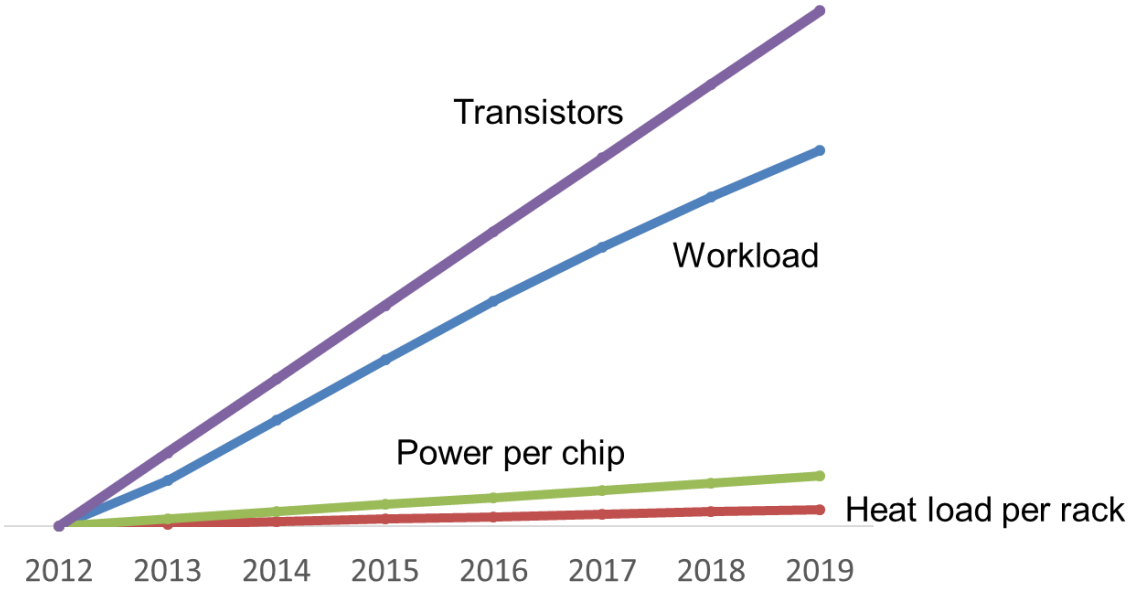
https://aws.amazon.com/



Anıl Tırlıoğlu
M.Sc. Thesis-ACCLOUD Project

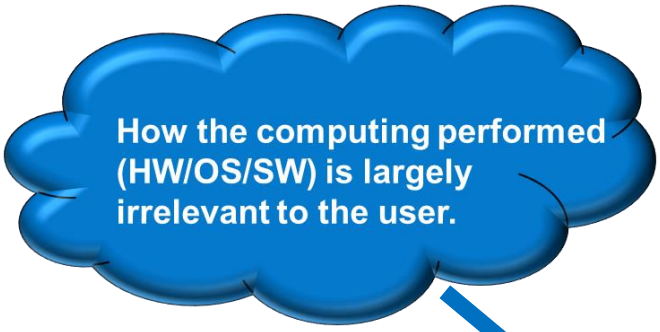
State of Cloud Data Center

- Workload and Transistors increase fast
- Power and heat budget stay the same



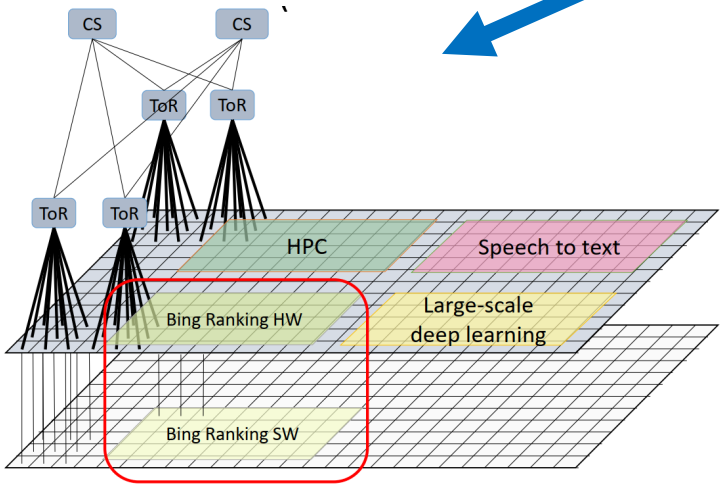
Kachris, Christoforos, and Dimitrios Soudris. "A survey on reconfigurable accelerators for cloud computing." *2016 26th International conference on field programmable logic and applications (FPL)*. IEEE, 2016.

Part III: Hardware Accelerated Cloud Data Centers



Xilinx Versal ACAP Adaptive Compute Acceleration Platform

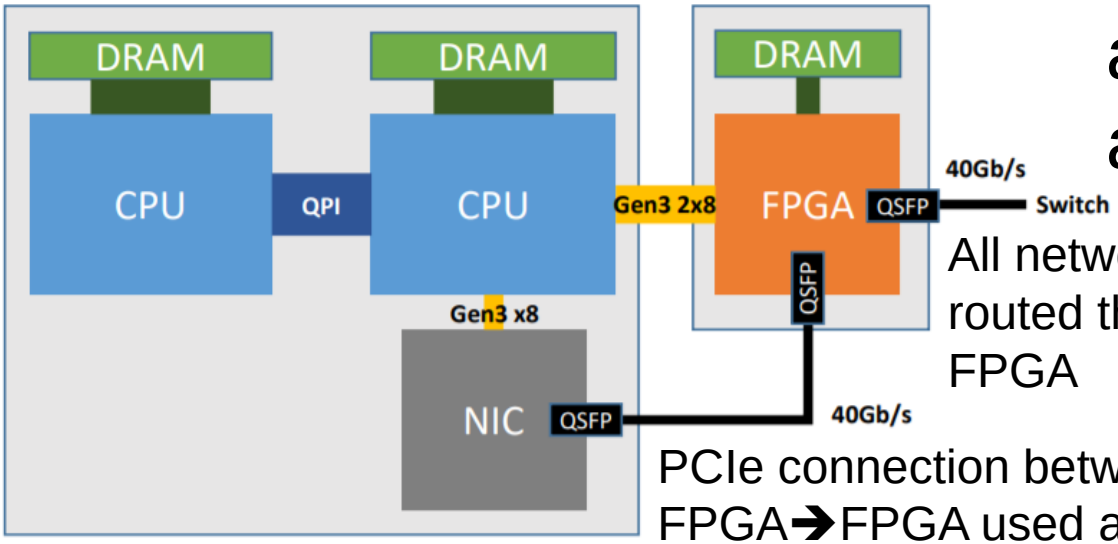
https://www.xilinx.com/products/silicon-devices/acap/versal.html



Hardware Accelerated Cloud Data Center

Hardware Accelerated Cloud Data Centers

- Microsoft Catapult Project
- FPGA is both the accelerator platform and the NIC



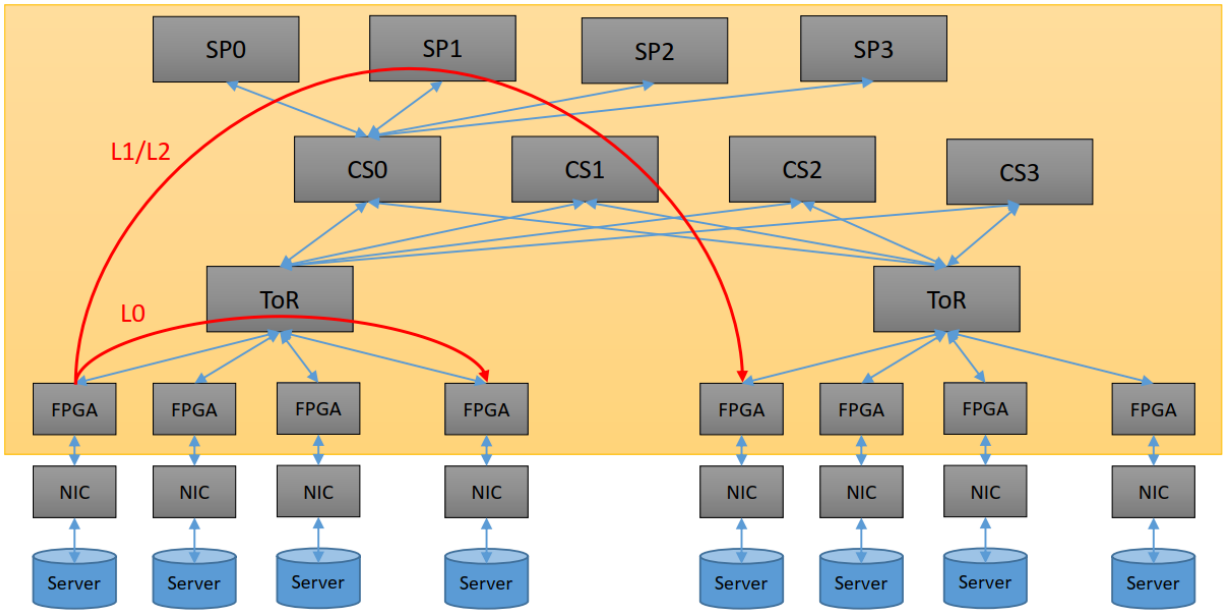
All network traffic is routed through the FPGA

PCIe connection between the host CPU and FPGA➔FPGA used as a local compute accelerator.

Caulfield, Adrian M., et al. "A cloud-scale acceleration architecture." *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2016.

Hardware Accelerated Cloud Data Centers

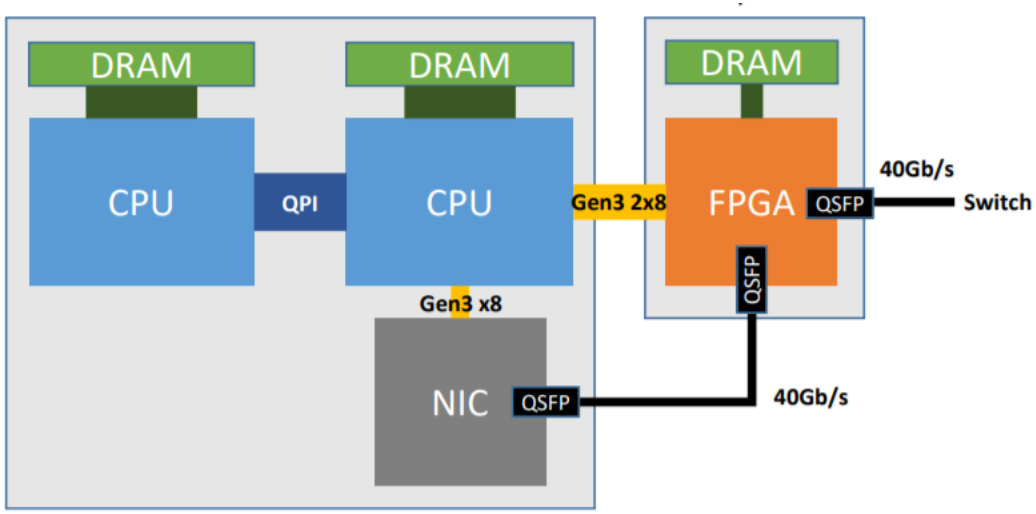
- Microsoft Catapult Project



- Low-latency inter-FPGA communication (Light Transport Layer)

Hardware Accelerated Cloud Data Centers

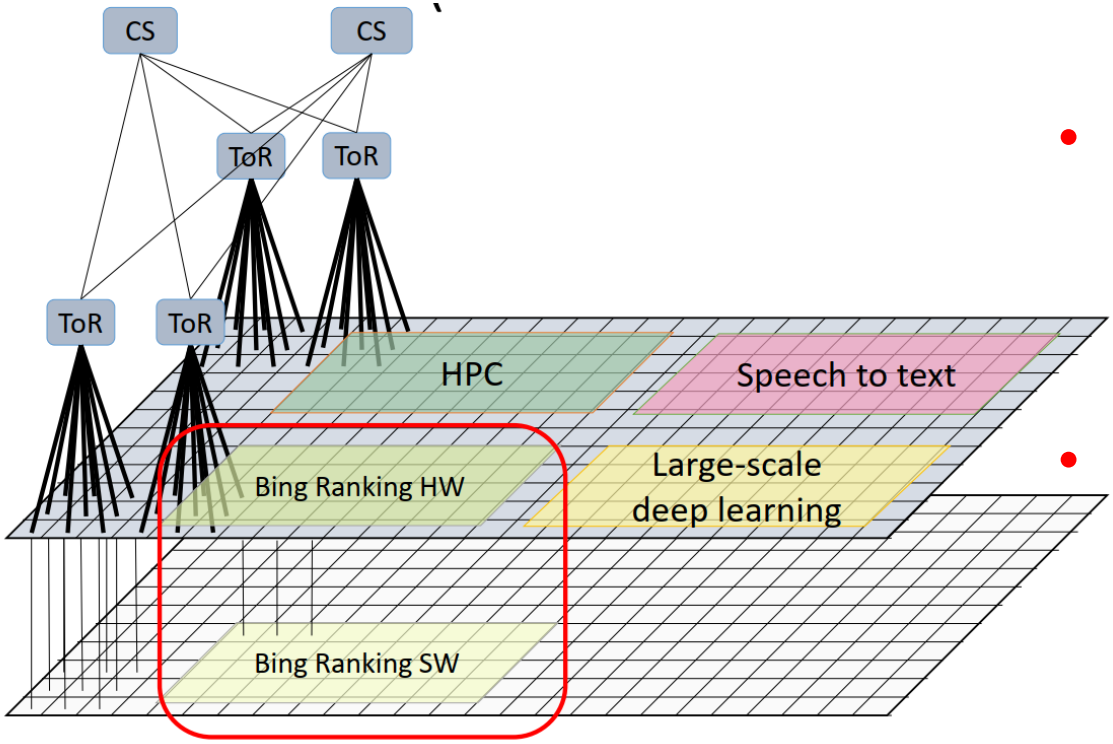
- Microsoft Catapult Project



- Local compute accelerator
- Network/storage accelerator
- Remote compute accelerator

Hardware Accelerated Cloud Data Centers

- Microsoft Catapult Project



- Hardware Acceleration as a Service Across Data Center (or even across Internet)
- FPGA is independent of the server

Part IV: ACCLOUD (Accelerated Cloud): A New Cloud Architecture with FPGA Acceleration

- TUBITAK Funded 1003 Research Project (to finish in April 2021)
- METU and Aselsan are partners
- Participation of many graduate students

http://accloud.eee.metu.edu.tr/

ACCLOUD (Accelerated Cloud): A New Cloud Architecture with FPGA Acceleration

Accelerator as a Service

Optimal, accelerator aware resource allocation

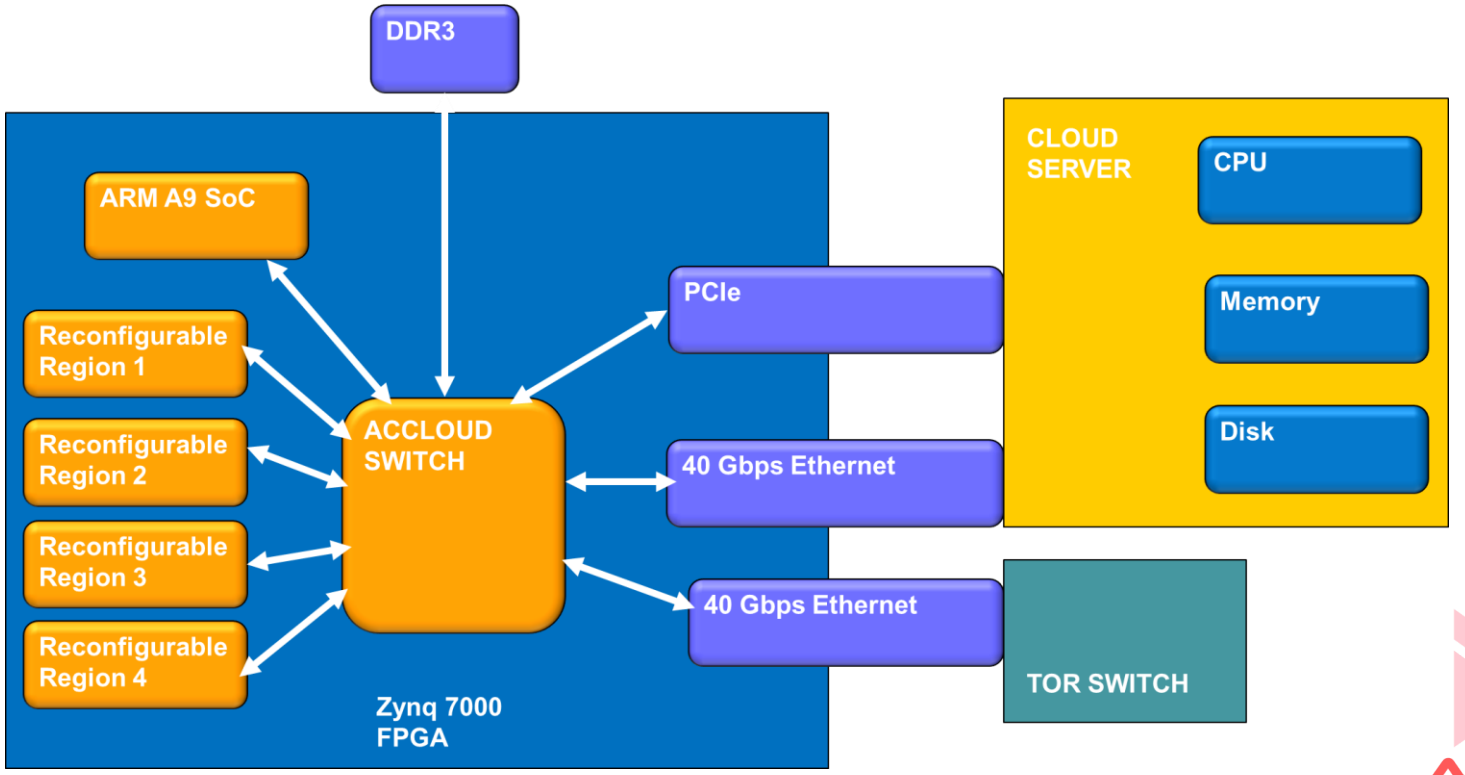
Accelerator implementation on FPGA reconfigurable regions

Transparent allocation of accelerators as Virtual Machine parameters

On-chip switch architecture for interconnecting hardware modules.

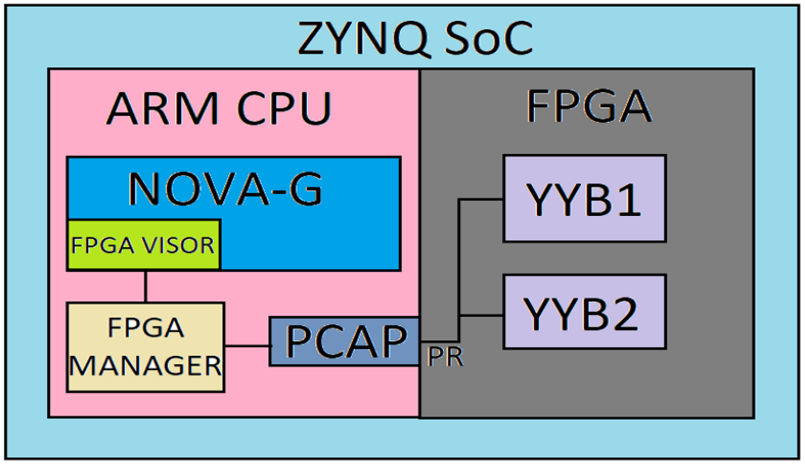
ACCLOUD (Accelerated Cloud): A New Cloud Architecture with FPGA Acceleration

ACCLOUD FPGA Accelerator and Cloud Server Layout



ACCLOUD (Accelerated Cloud): A New Cloud Architecture with FPGA Acceleration

Transparent allocation of accelerators as Virtual Machine parameters



openstack

- Cloud Resource Management Framework for VM Creation
- Modification of Nova Compute component to allocate accelerators
- Similar to allocatin RAM, Disk

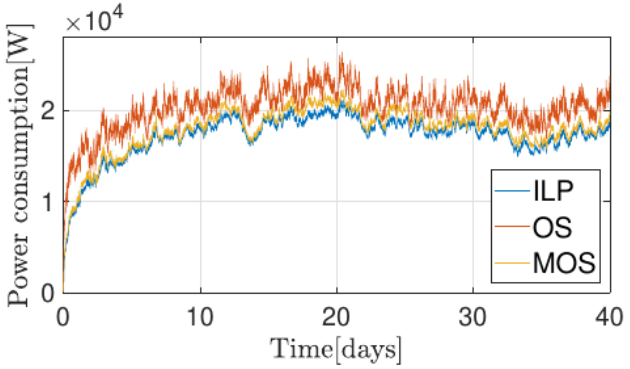
ACCLOUD

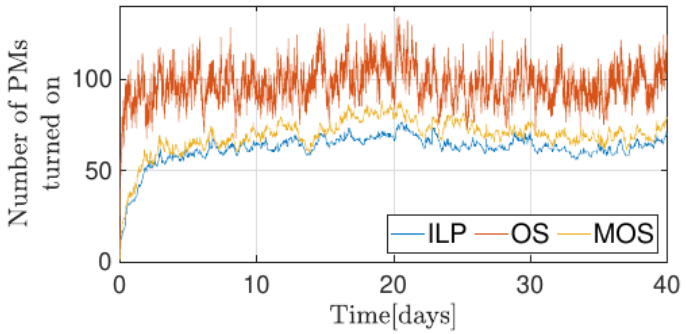
A. Erol, A. Yazar and E. G. Schmidt, "OpenStack Generalization for Hardware Accelerated Clouds," *2019 28th International Conference on Computer Communication and Networks (ICCCN)*, Valencia, Spain, 2019.

ACCLOUD (Accelerated Cloud): A New Cloud Architecture with FPGA Acceleration

Optimal, accelerator aware resource allocation: ACCLOUD-MAN

- Defining alternatives for SaaS requests
- Example: Video processing
  - 4 CPU cores
  - or 2 FPGA regions
  - or 2 CPU cores and 1 FPGA regior
- Resource Allocation with minimum num Physical Machines, minimum power consumption





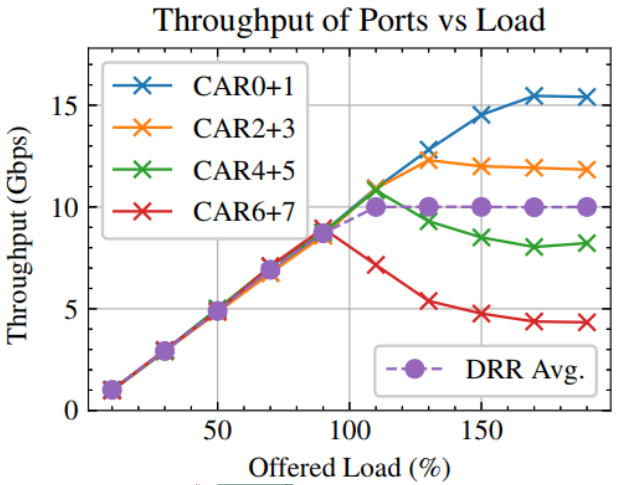
ILP: ACCLOUD-MAN
OS: Legacy OpenStack

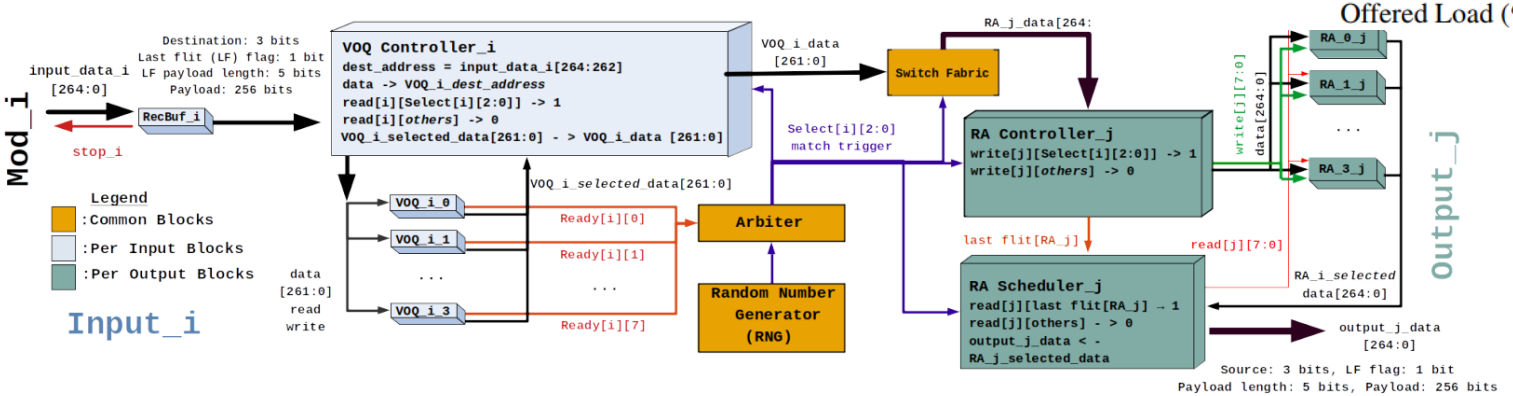
N. U. Ekici, K. W. Schmidt, A. Yazar and E. G. Schmidt, "Resource Allocation for Minimized Power Consumption in Hardware Accelerated Clouds," *2019 28th International Conference on Computer Communication and Networks (ICCCN)*, Valencia, Spain, 2019.

ACCLOUD (Accelerated Cloud): A New Cloud Architecture with FPGA Acceleration

On-chip switch architecture for interconnecting hardware modules: ACCLOUD-SWITCH

- 40 Gbps crossbar fabric
- 256 bit payload/flits
- Novel fabric arbitration with QoS

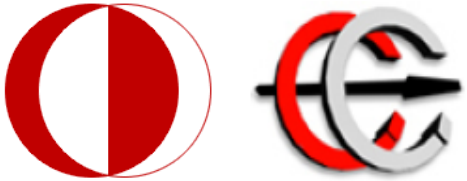




F. Yazıcı, A. S. Yıldız, A. Yazar, E. G. Schmidt, "A Novel Scalable On-chip Switch Architecture with Quality of Service Support for Hardware Accelerated Cloud Data Centers," *IEEE International Conference on Cloud Networking*, 2020.

Concluding Remarks

- Exploiting the *golden age* of hardware acceleration (as put by Henessy and Patterson)
- Seamlessly offering hardware resources to achieve more power efficient and higher performance services
- Wonderful research opportunities with many interesting problems!



Cloud Computing and Hardware Accelerated Clouds

Ece Güran Schmidt

METU Electrical and Electronics Engineering