CompSci 401: Cloud Computing

What is the Cloud?

Prof. Ítalo Cunha

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

What is cloud computing?

- The transformation of IT from product to service

What is cloud computing?

- The transformation of IT from product to service

What is cloud computing?

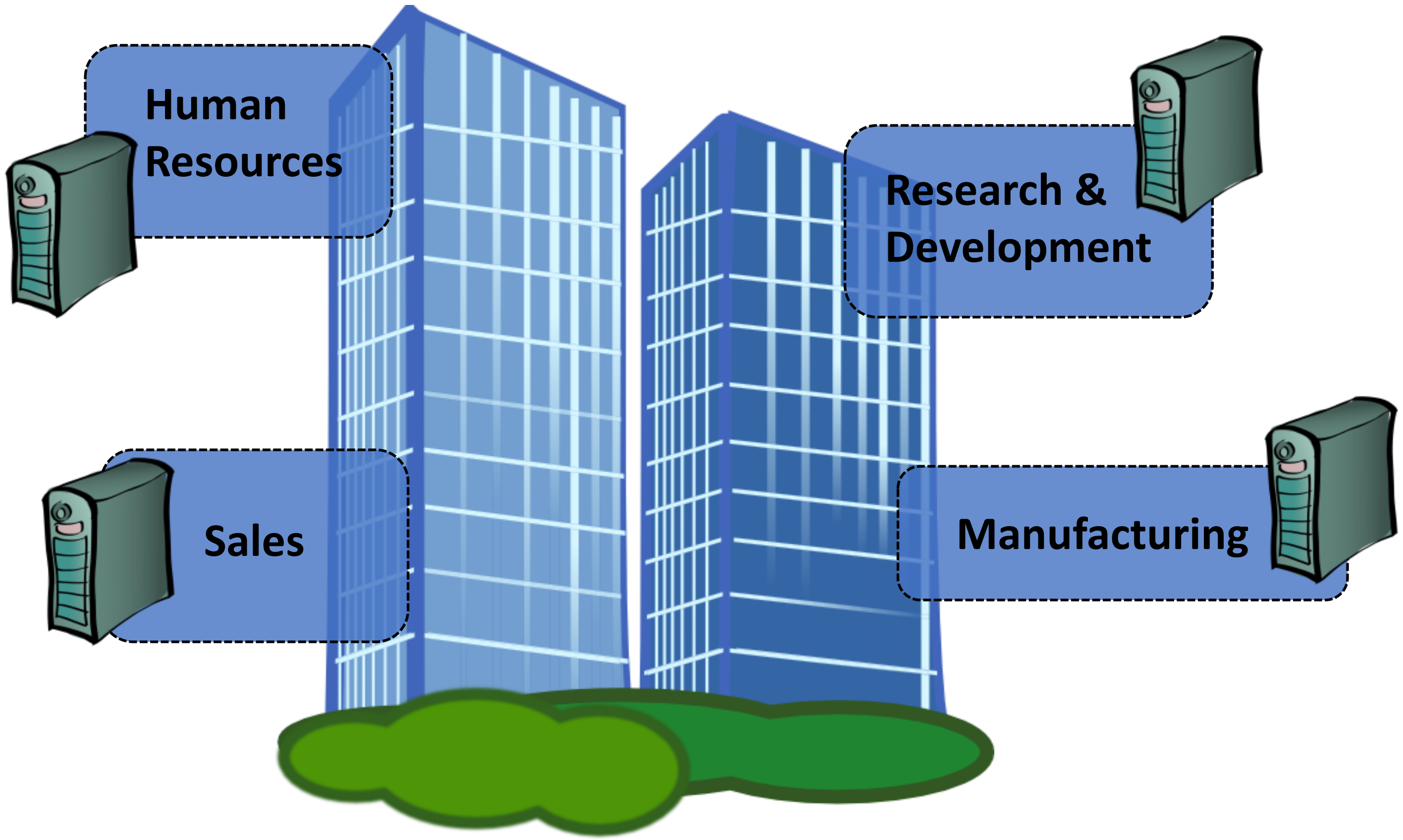
- The transformation of IT from product to service
- Paradigm where applications or services run (partially or completely) on third-party infrastructure, cloud tenants pay for what they use

Examples of interactions with the cloud

- Duke University stores videos on Panopto
- An individual checks her calendar on a smartphone
- Someone watches a video on Netflix
- A developer commits code to GitHub or chat over Slack
- A cyclist’s wearable watch sends information to a health tracking app
- Coworkers edit a document on Google Docs or Office 360
- Enterprise leases servers where it runs intranet services
- A sales company leases additional servers during peak periods like black Friday to keep up with demand

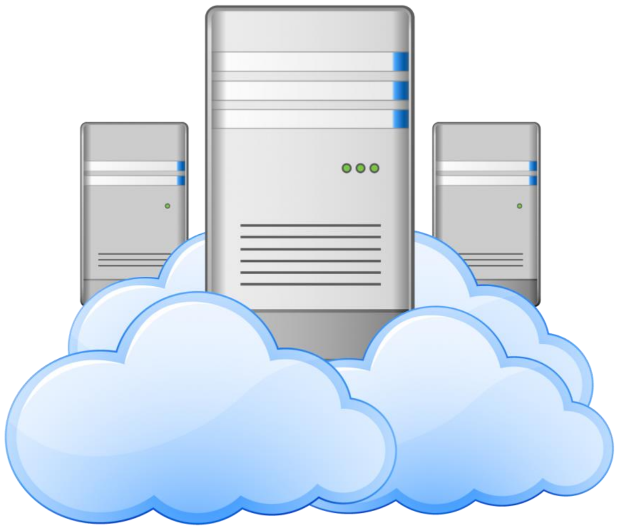
Enterprise computing in the past

- Enterprises ran applications on several servers, possibly spread across multiple departments

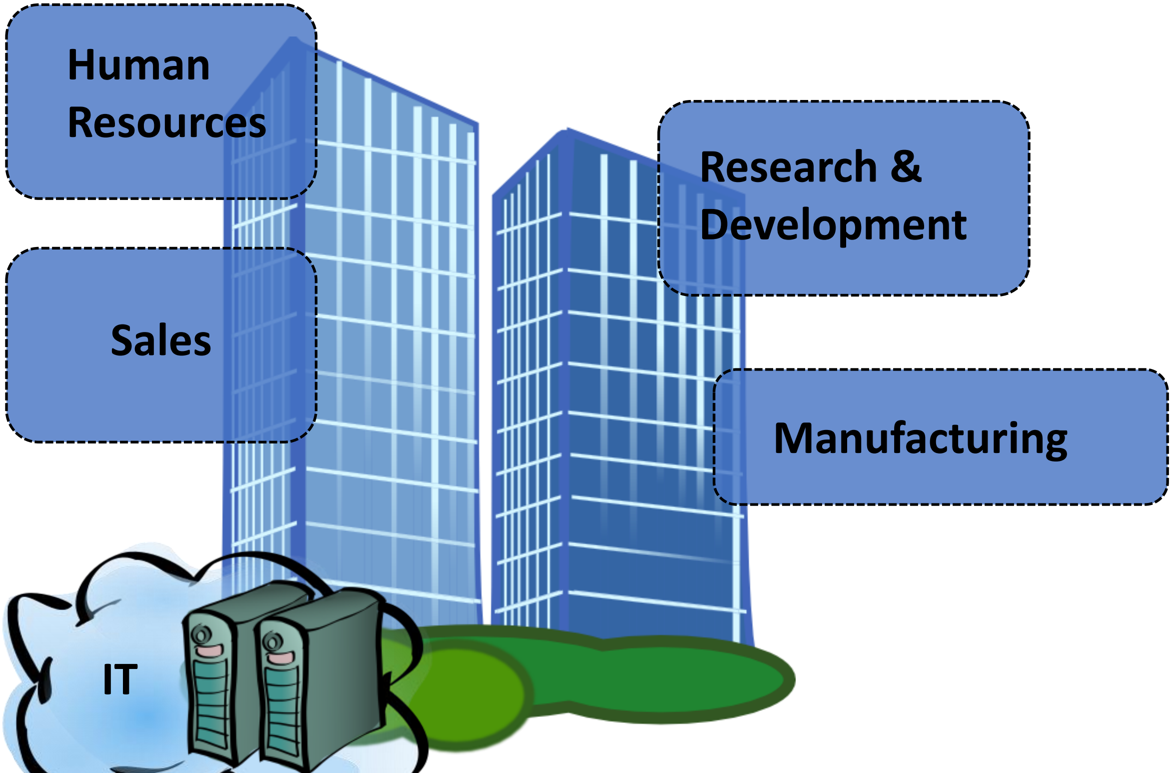


Enterprise computing today

- Some enterprises move *all* IT to the cloud
- Enterprises that keep local compute usually concentrate hardware on a small local datacenter



Third-party infrastructure

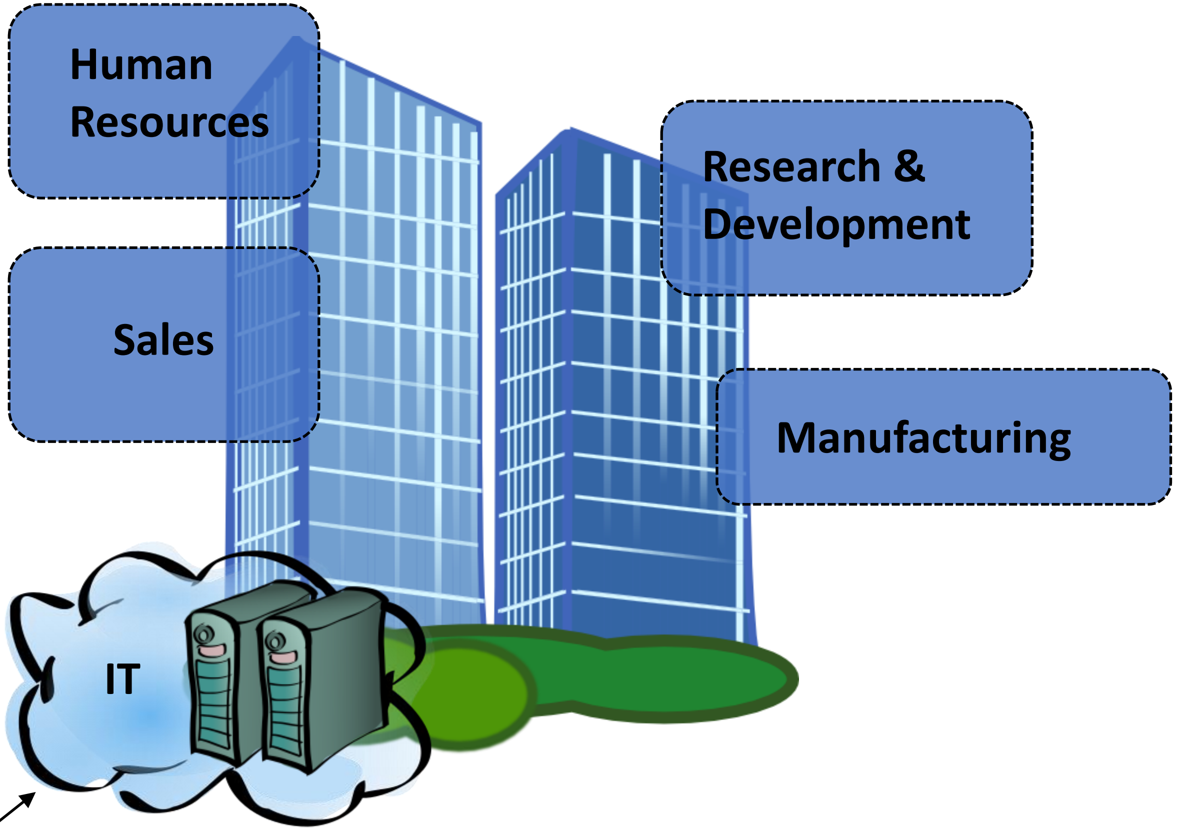


Local infrastructure

Private and public clouds

Third-party infrastructure

Public cloud



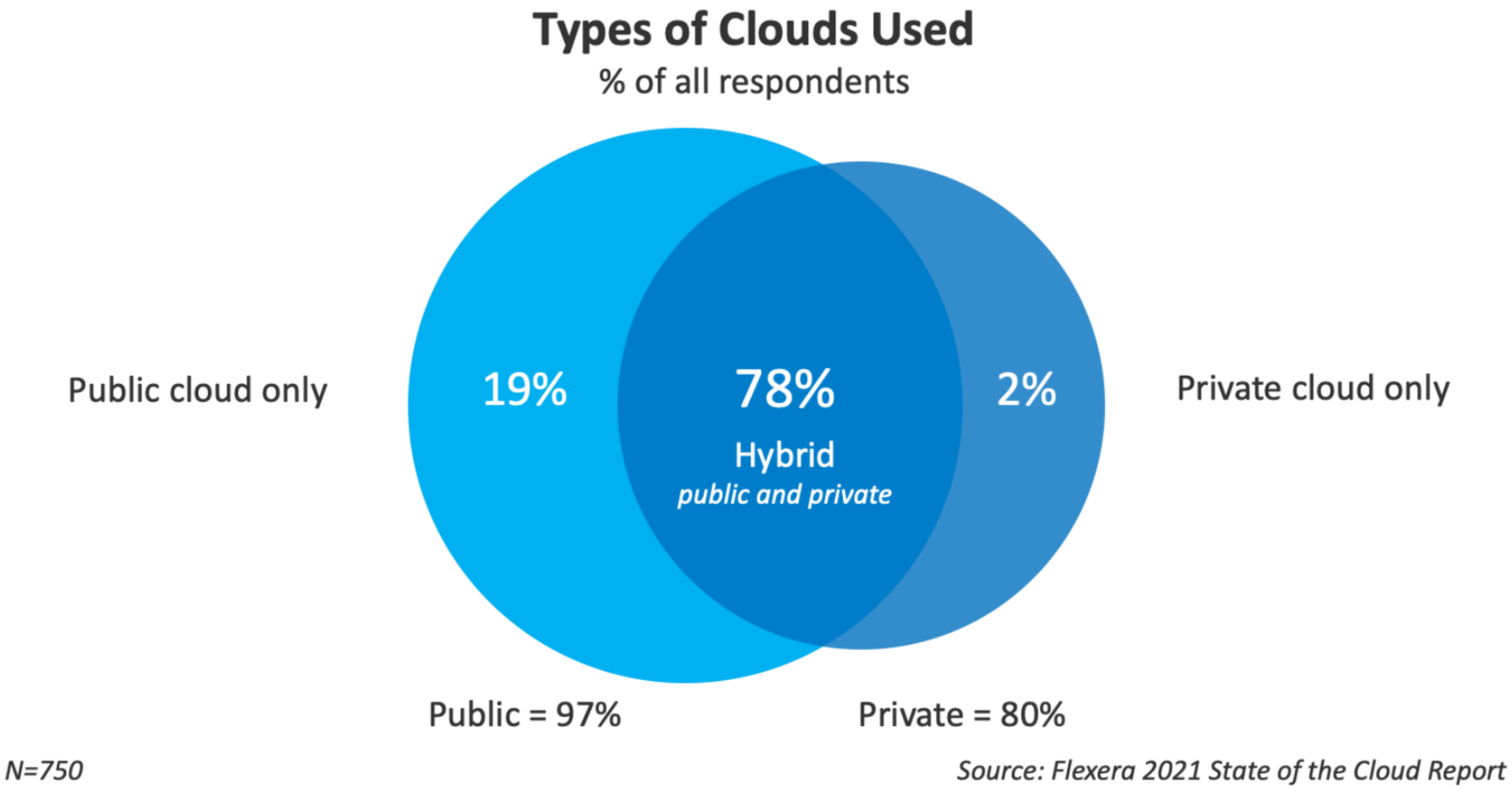
Local infrastructure

Private cloud

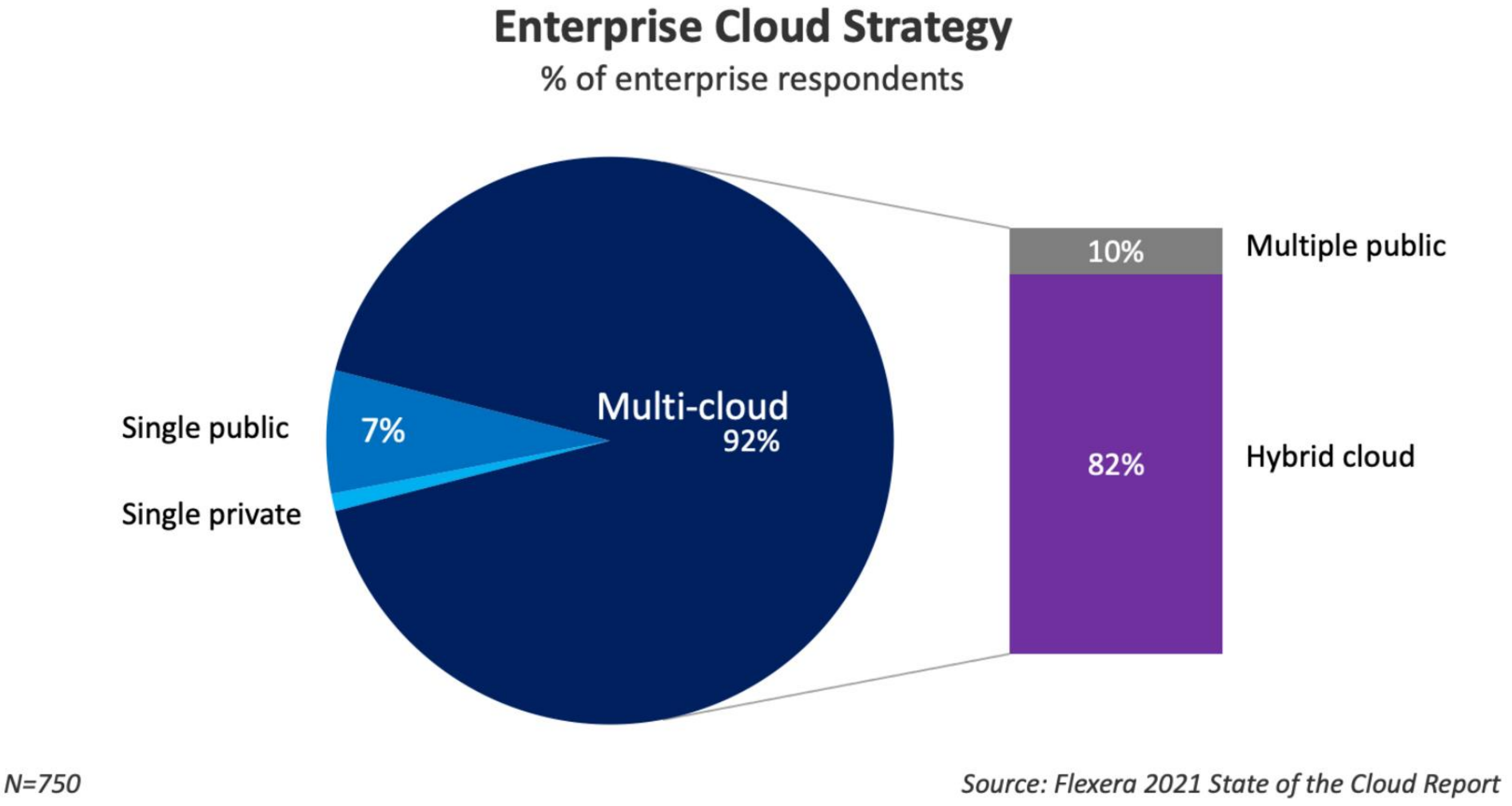
Cloud Types

- Private cloud
  - Internal cloud operated and used by one organization
- Public cloud
  - Commercial service operated by a third-party organization
  - Hosts multiple customers
- Hybrid-cloud
  - Organizations that use both private and public clouds
- Multi-cloud
  - Organizations that use multiple private/public clouds

Most enterprises use at least one cloud



Most enterprises use multiple clouds



Mobile applications

- Cell phones have limited resources
  - CPU, memory, and storage
  - Battery
- Most applications rely heavily on the cloud
  - Store data
  - Run heavy computations
    - Searching e-mail
    - Computing traffic directions

CompSci 401: Cloud Computing

Drivers of Adoption

Prof. Ítalo Cunha

昆山杜克大学
DUKE KUNSHAN
UNIVERSITY

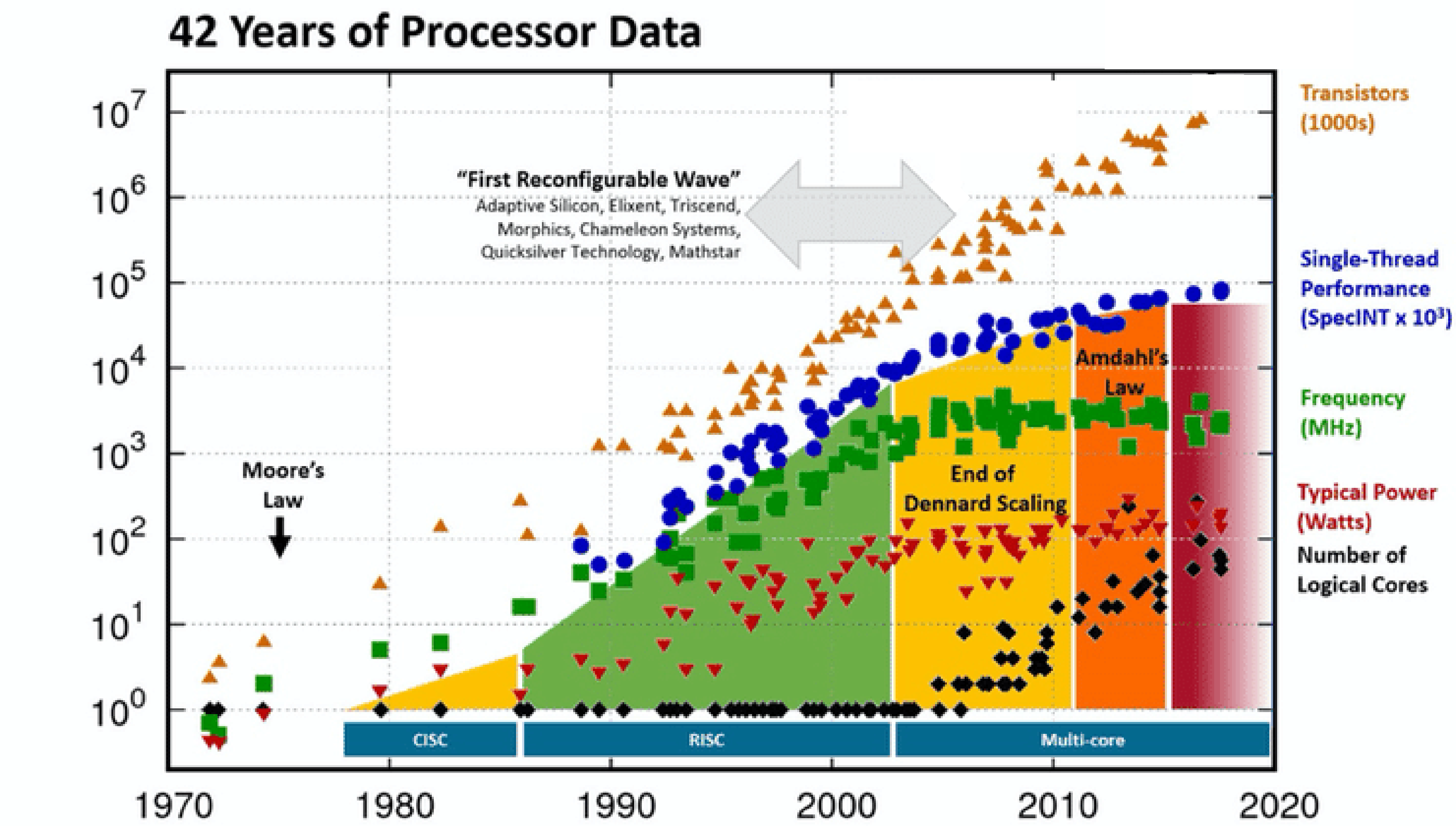
Cloud computing allows control over resource allocation

- Cloud tenants can choose how much resources to use dynamically
- Incremental growth
  - Start small and grow as business picks up
  - Start with a simple deployment and add complexity later
- Dynamic scaling
  - Tenants can not only increase their footprint, but also decrease
  - Absorb bursts of demand (e.g., Black Friday, Super Bowl, World Cup)
  - Pay only for resources used

Two key drivers to cloud adoption

- Flattening of single-thread performance and move to parallelism
- Changes to infrastructure operation and maintenance costs

Power constraints and multiple cores



Hennessy and Patterson, Turing Lecture 2018, overlaid over "42 Years of Processors Data"
https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/;"First Wave" added by Les Wilson, Frank Schirrmeister
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

Scaling behind a single server

Cray Y-MP
1st between 1988-1989

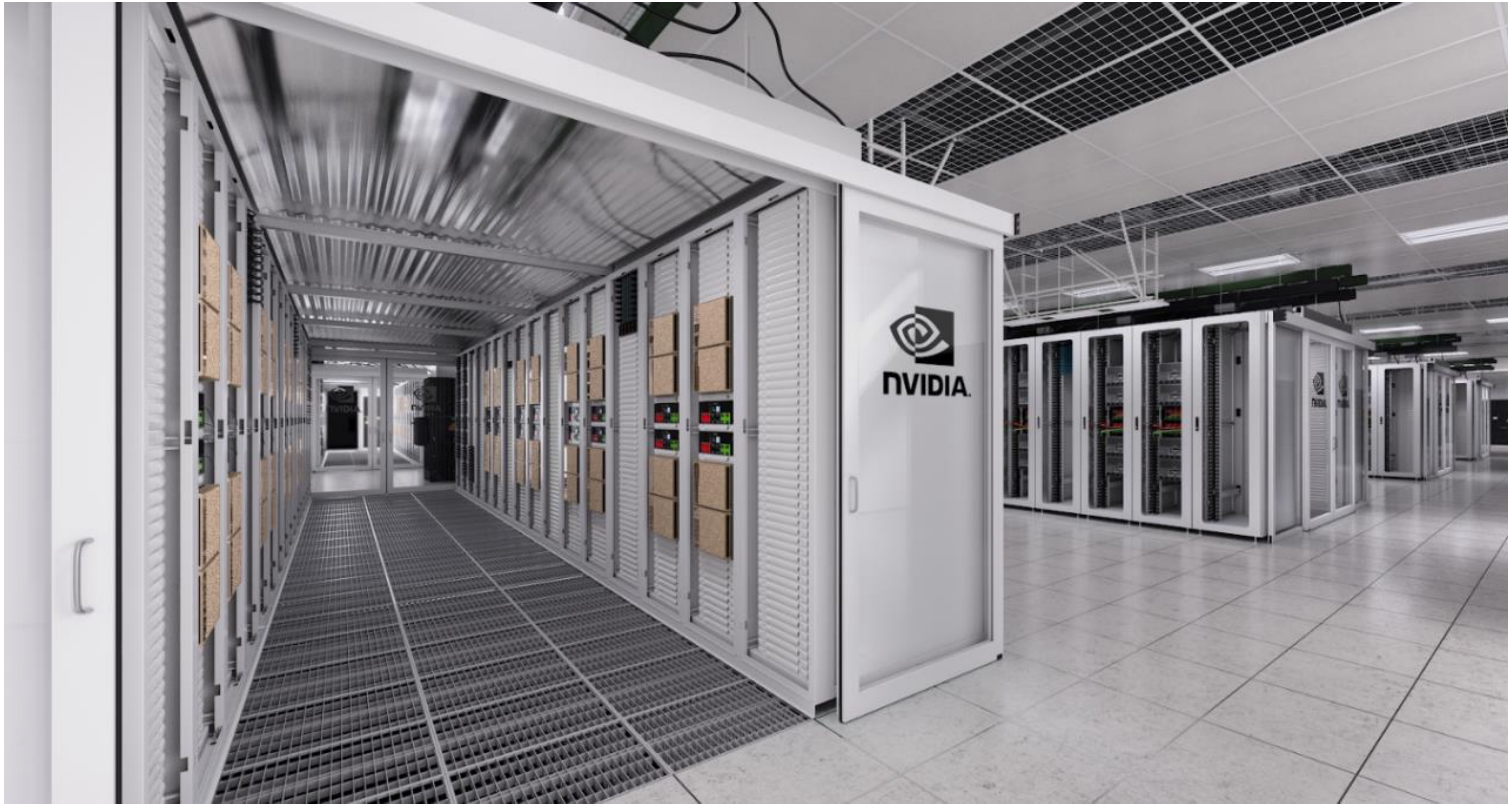
2, 4, or 8 vector processors

Scaling behind a single server

NEC Earth Simulator
1st between 2002-2004

640x SX-6 nodes with:
- 8 vector processors
- 16 GiB of RAM

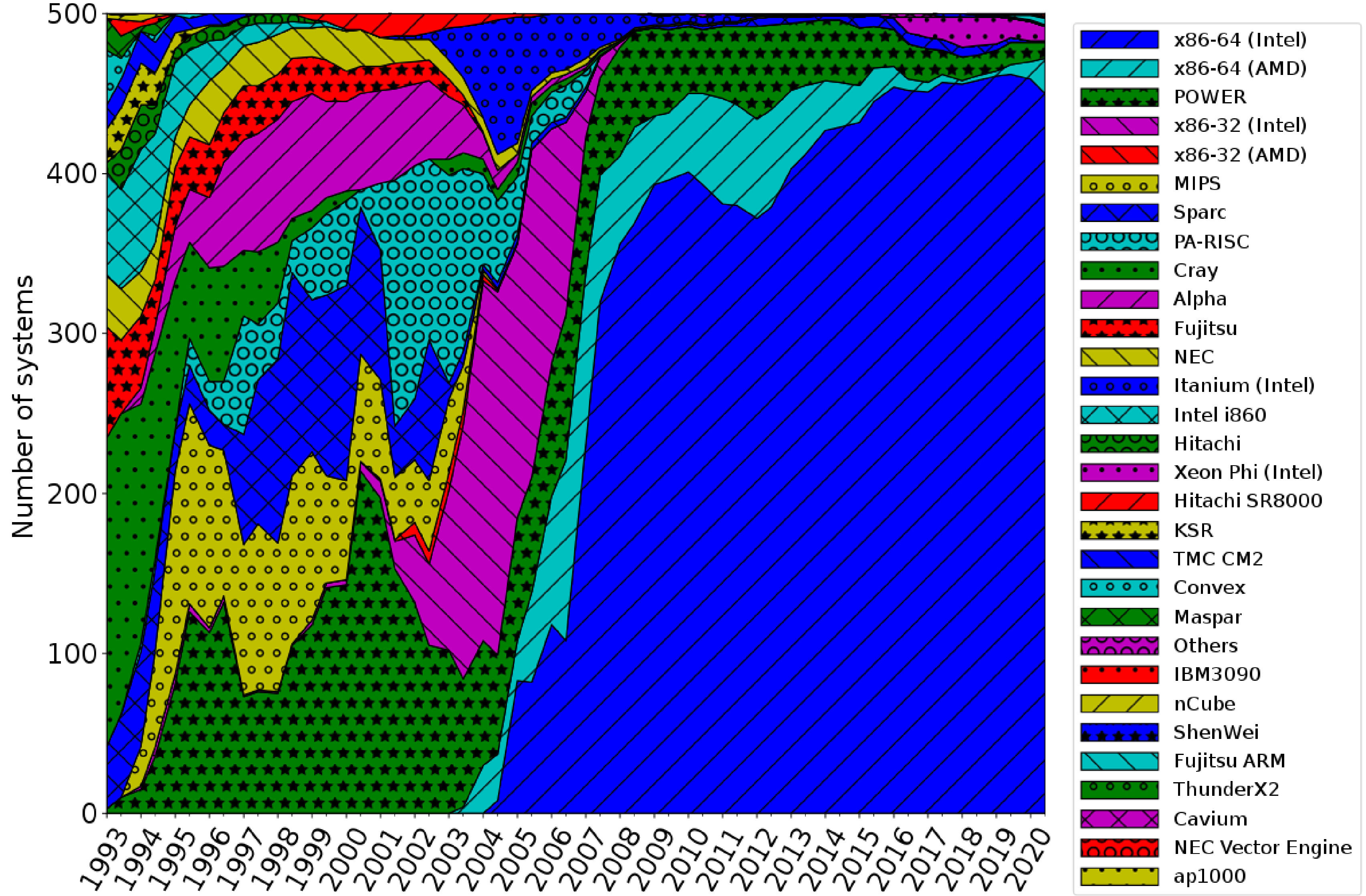
Scaling behind a single server



Selene
5th fastest @ 2021

AMD Epyc and
NVIDIA GPUs

Processor families of top500 supercomputers



CompSci 401: Cloud Computing

Computing Cluster

Prof. Ítalo Cunha

昆山杜克大学
DUKE KUNSHAN
UNIVERSITY

Limitless scaling with off-the-shelf equipment

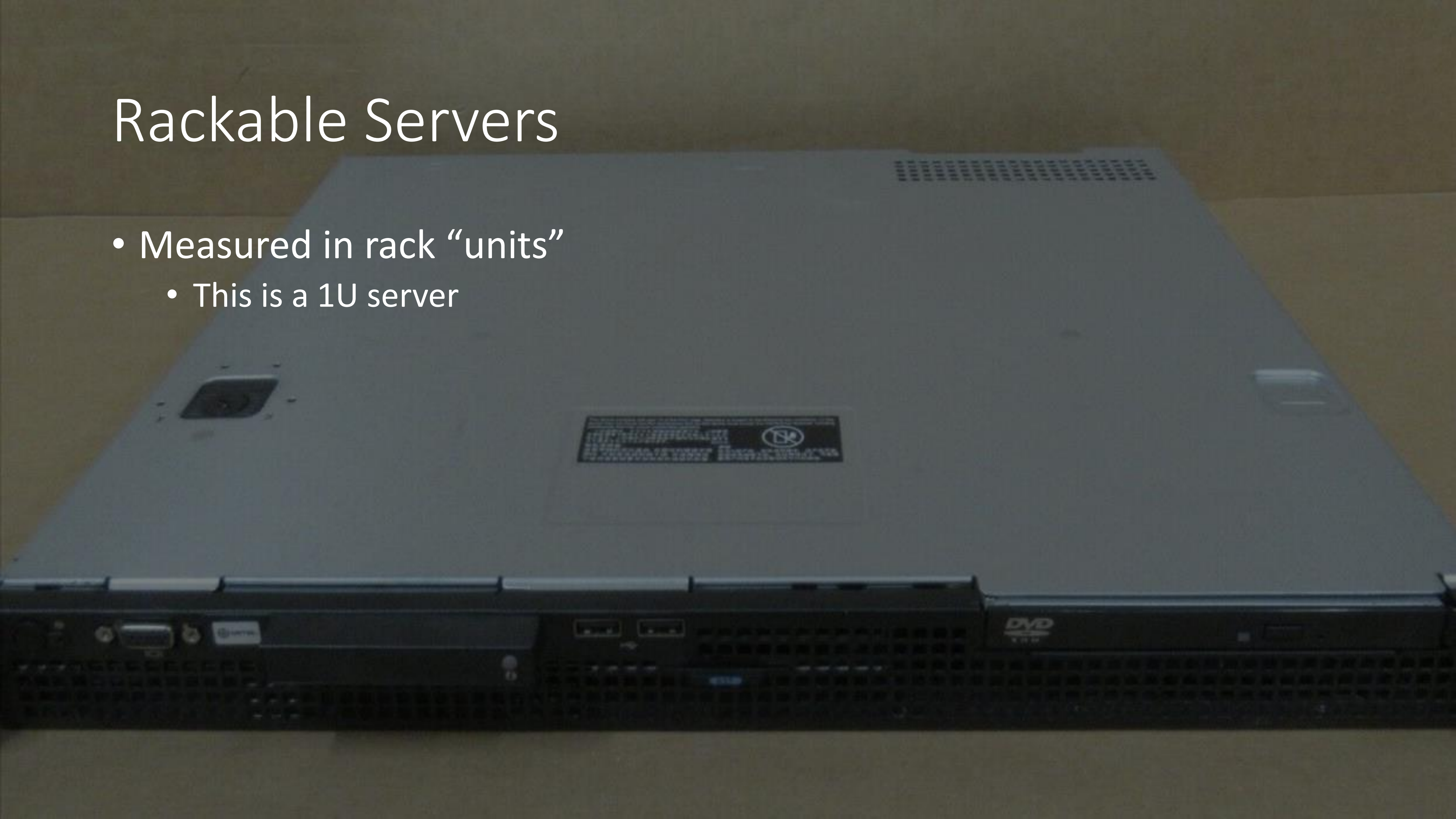
- Build *clusters* of computers with commodity components
- Good performance-to-cost trade-offs
  - Including replacement and maintenance
- Good performance-to-power trade-offs
  - GFLOPS per Watt
- Easier to program and use compared to custom architectures

Limitless scaling with off-the-shelf equipment

- Build *clusters* of computers with commodity components
- Good performance-to-cost trade-offs
  - Including replacement and maintenance
- Good performance-to-power trade-offs
  - GFLOPS per Watt
- Easier to program and use compared to custom architectures
  - But applications need to fit into commodity computers

Rackable Servers

- Measured in rack “units”
  - This is a 1U server



Server Rack

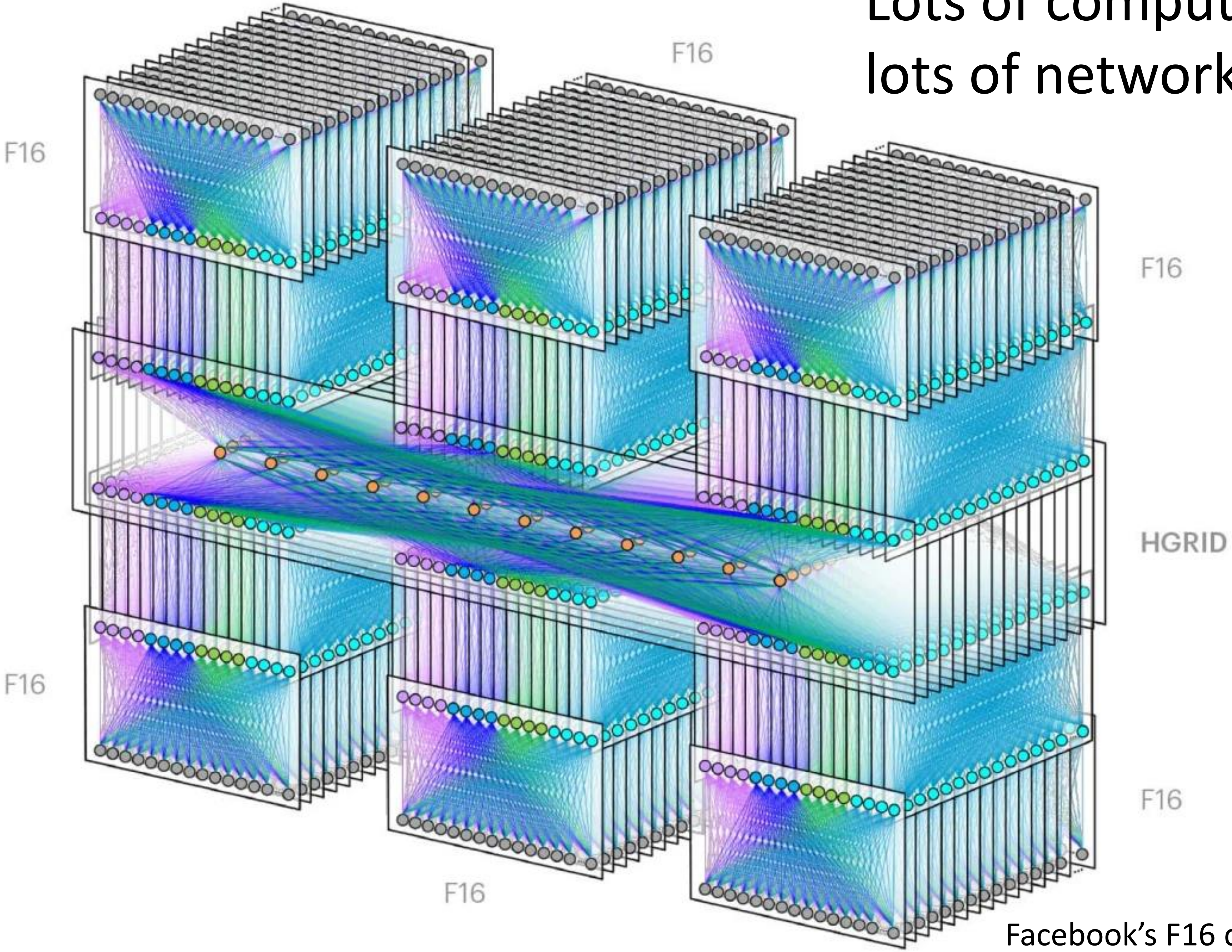
An ad-hoc server rack

(including some non-rackable machines)

Structured racks

NVIDIA DGX SuperPod

Lots of computers need lots of network bandwidth



Facebook’s F16 design

Economic motivation for clustering computers

- Clustering computers reduces costs
- Operating expenses (OPEX)
  - System administrators needed to manage computing infrastructure
    - Installing and configuring software, network cabling
    - Replacing damaged or broken components
    - Expanding server resources (e.g., adding RAM or disks)
  - Having multiple sysadmin teams on multiple departments managing scattered computer resources gets expensive
  - Scattered infrastructure complicates knowledge-sharing among sysadmin teams
    - Some issues happen infrequently, better to have one person know how to handle each
  - Clustering resources drives adoption of automation, reducing costs incurred on repetitive tasks

Economic motivation for clustering computers

- Clustering computers reduces costs
- Capital expenses (CAPEX)
  - Scattered resources are usually bought individually
    - Leading to heterogeneous deployments, which incur higher OPEX
  - Centralizing resources allows buying many identical servers, possibly at a quantity discount
  - Server upgrades can be performed in batches to reduce heterogeneity while still allowing for quantity discounts
    - For example, a company can decide to upgrade 25% of its servers each year

CompSci 401: Cloud Computing

Clusters in the Cloud

Prof. Ítalo Cunha

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

Major drivers to adoption

- Power wall induced move to multicore processors
- Centralizing commodity servers in clusters
    - Better OPEX and CAPEX
    - Less human resources
    - Quantity discounts
    - Less heterogeneity
    - Easier automation

Cloud providers potentialize these advantages

- Cloud providers host multiple tenants
- Build multiple large data centers to handle computing for customers
  - Buy thousands of each component, significant quantity discounts
  - Thorough automation
  - Standardized processes
    - Deploying new servers
    - Replacing parts
    - Handling failures

Cloud providers potentialize these advantages

- Cloud providers host multiple tenants
- Build multiple large data centers to handle computing for customers
  - Buy thousands of each component, significant quantity discounts
  - Thorough automation
  - Standardized processes (e.g., for deploying new servers or replacing parts)
- Isolating tenants is key
  - Performance for a tenant must not depend on other tenants
  - Each tenant's data and code must be kept safe
    - An organization may even want to isolate departments from one another

Facebook’s Fort Worth data center

https://www.facebook.com/FortWorthDataCenter

Inside: corridors, racks, and cabling

CompSci 401: Cloud Computing

Distributed Applications

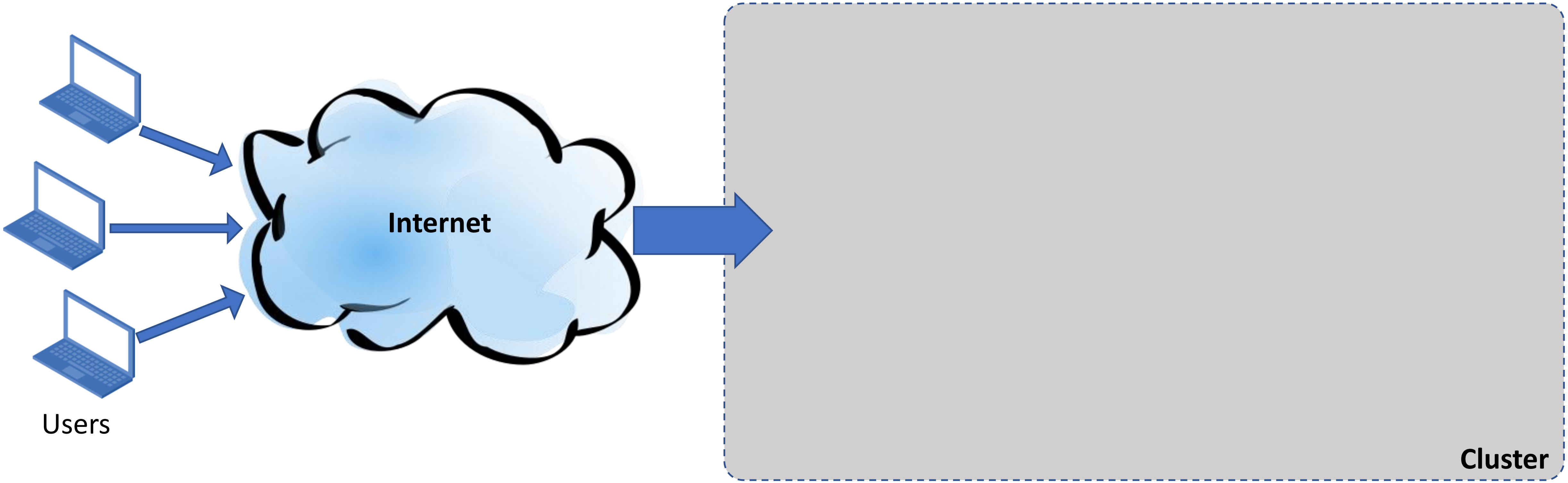
Prof. Ítalo Cunha

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

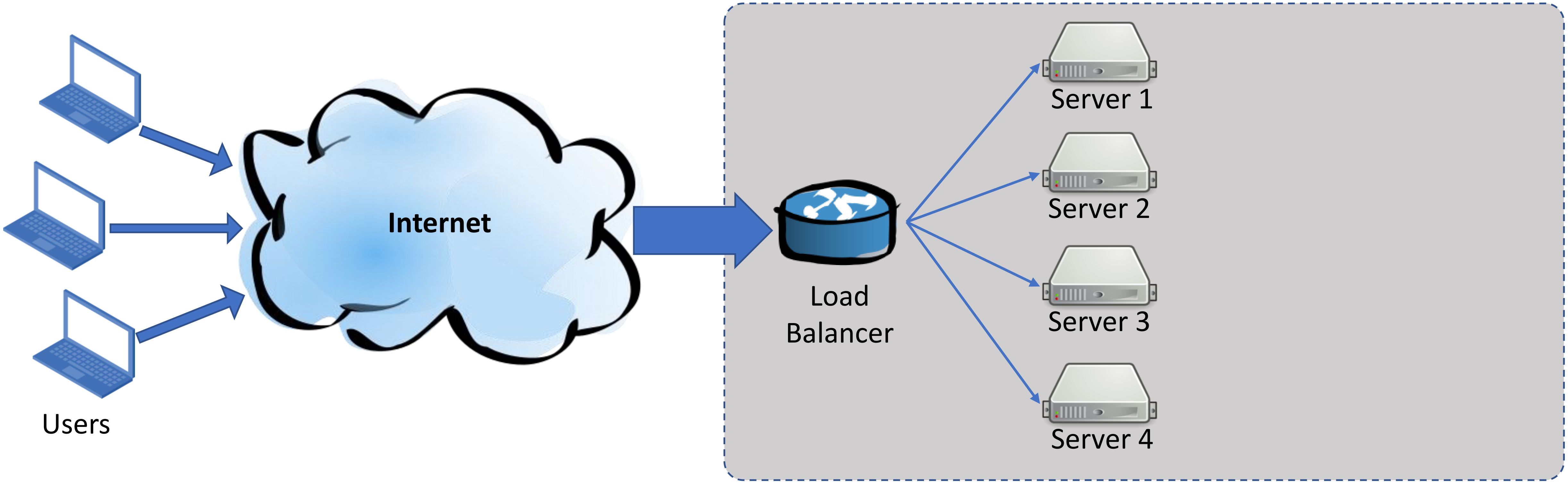
Limitless scaling with distributed applications

- Clusters impose restrictions on applications
- Applications need to fit on commodity servers
- Drive toward distributed solutions

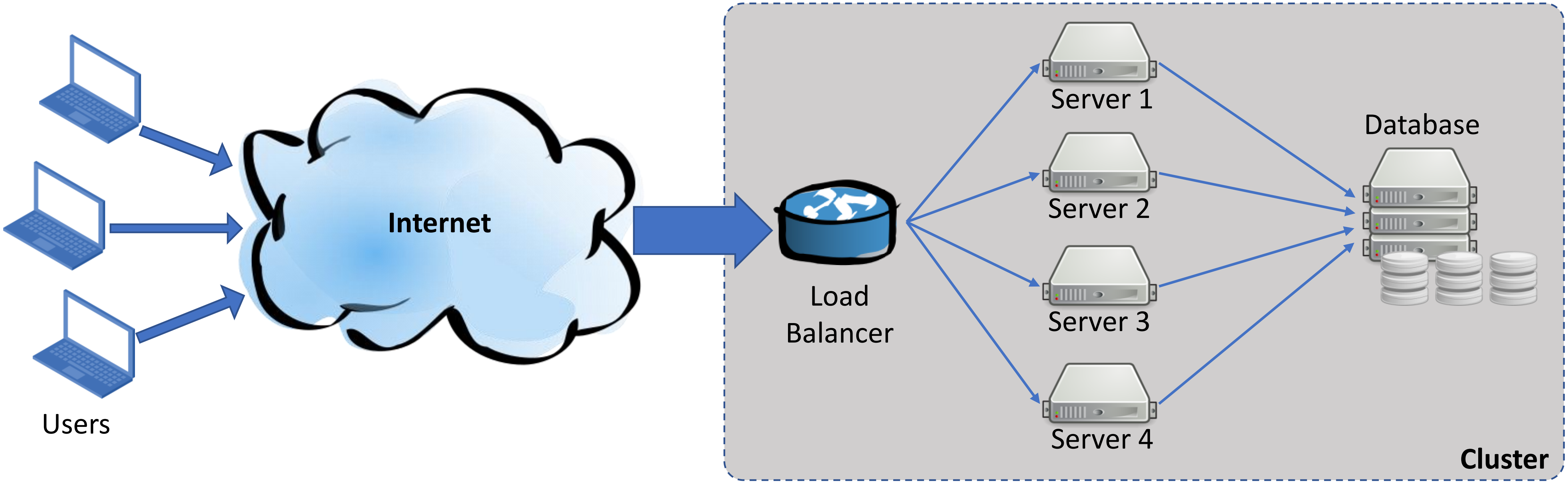
Distributed Web services



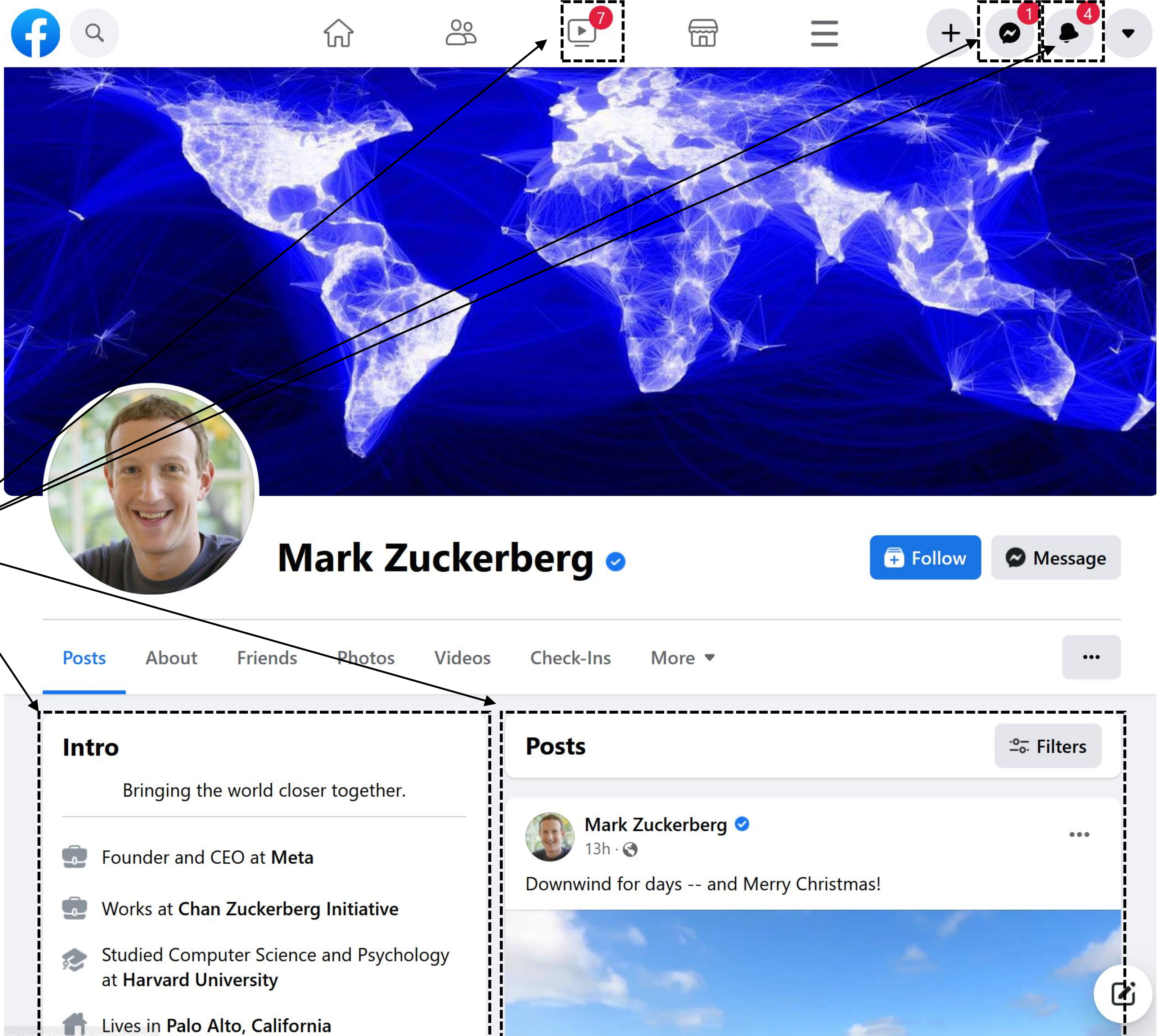
Distributed Web services



Distributed Web services



Each of these components might be delegated to other servers



Distributed Computation

- Large body of research on distributed algorithms
- Several programming paradigms and frameworks
  - MapReduce
  - BigQuery
  - Serverless computing
  - Stream processing