CompSci 401: Cloud Computing

Cloud End-to-end Traffic

Prof. Ítalo Cunha

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

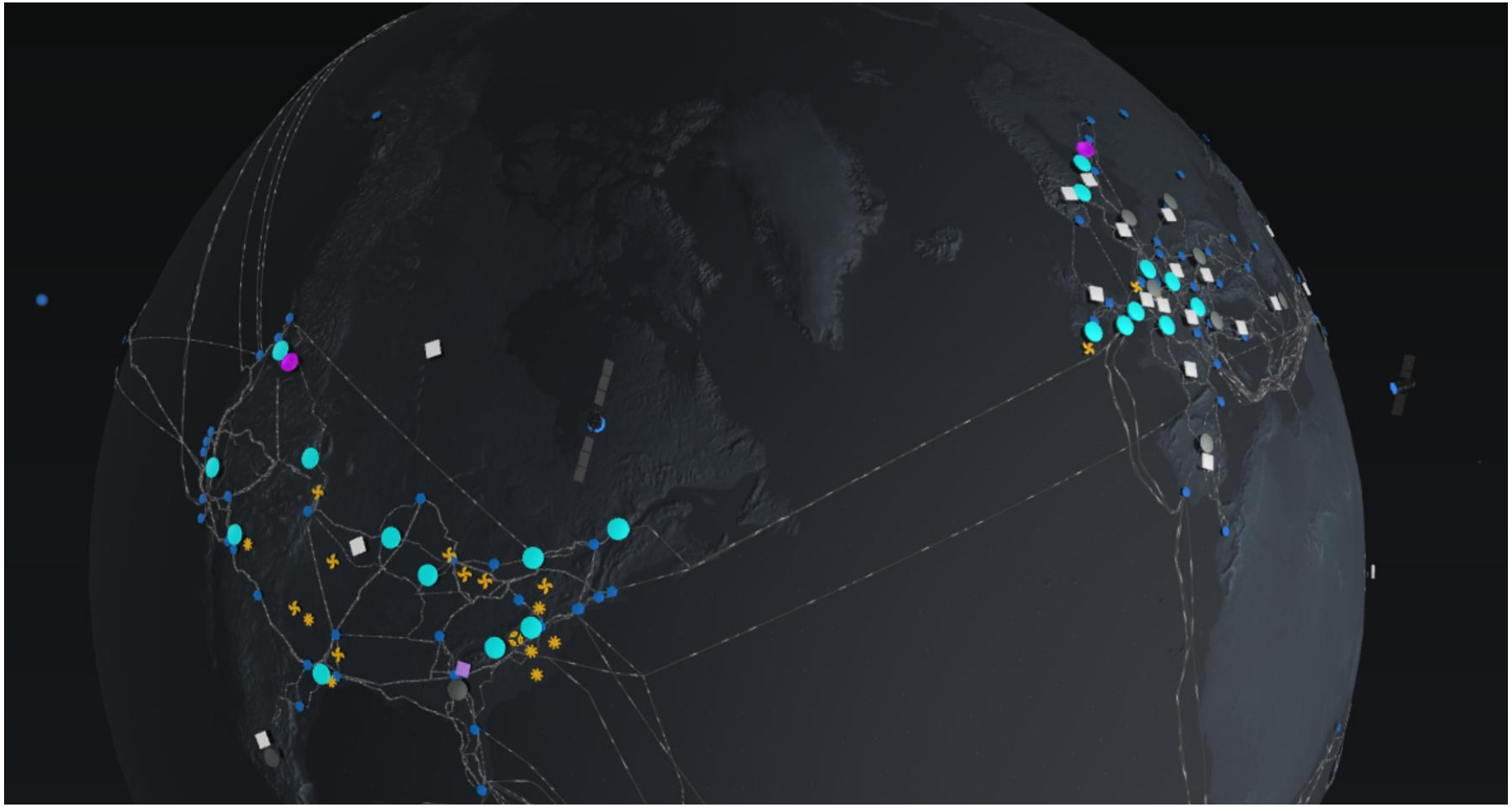
Cloud computing and end-to-end latency

- User traffic must traverse the Internet to reach a datacenter
- Incurs additional latency
  - Many applications do not care about latency
    - Video streaming, most Web applications, basic downloads
  - Some applications do care about latency
    - Gaming, teleconferencing, self-driving vehicles, telesurgery

Distributed, global deployments

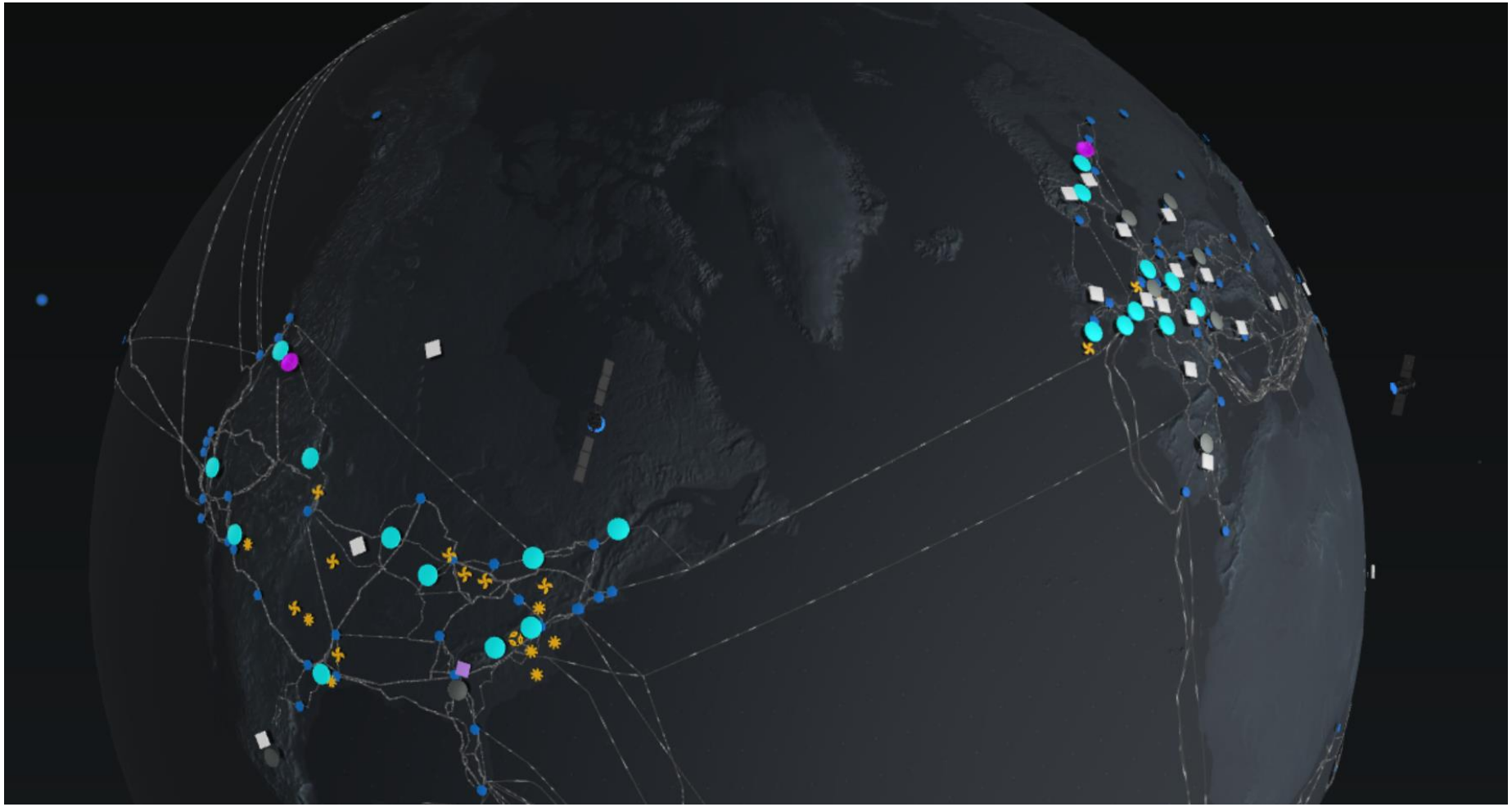
Google

Distributed, global deployments



https://infrastructuremap.microsoft.com/explore

Distributed, global deployments



Fast connectivity:
175K miles of fiber
140 countries

https://infrastructuremap.microsoft.com/explore

Cloud computing improved latency

- End-to-end latency improved significantly
  - 400ms latency between Brazil and the US back in 2000 → unplayable games
  - 120-200ms latency between Brazil and the US in 2020 → some playable games
  - 30-120ms latency within Brazil to São Paulo in 2020 → headshots

Cloud computing improved latency, but…

- End-to-end latency improved significantly
  - 400ms latency between Brazil and the US back in 2000 → unplayable games
  - 120-200ms latency between Brazil and the US in 2020 → some playable games
  - 30-120ms latency within Brazil to São Paulo in 2020 → headshots
- However, 100ms is unacceptable for some applications
  - Telesurgery
  - Collision avoidance for self-driving vehicles
  - Industrial, robot, home automation

Bandwidth constraints

- Datacenters have a lot of bandwidth, but traffic traverses the Internet
  - Incurs load and transit costs on intermediate networks
- But some applications require *a lot* of bandwidth
  - Netflix
  - Youtube

Bandwidth constraints

- Datacenters have a lot of bandwidth, but traffic traverses the Internet
  - Incurs load and transit costs on intermediate networks
- But some applications require *a lot* of bandwidth
  - Netflix
  - Youtube
  - “Net neutrality”

CompSci 401: Cloud Computing

Edge Computing

Prof. Ítalo Cunha

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

Edge Computing

Datacenter

Internet

Local Infrastructure

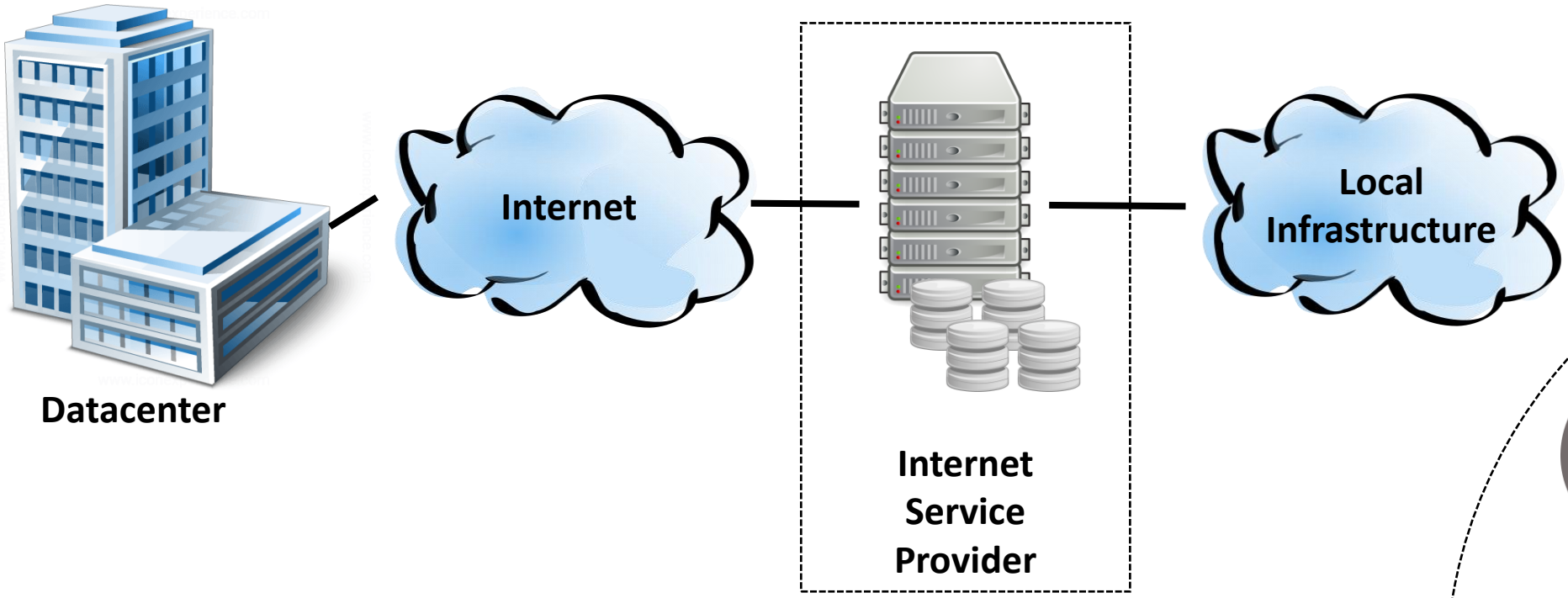
User's Home

User

4G/5G Cell

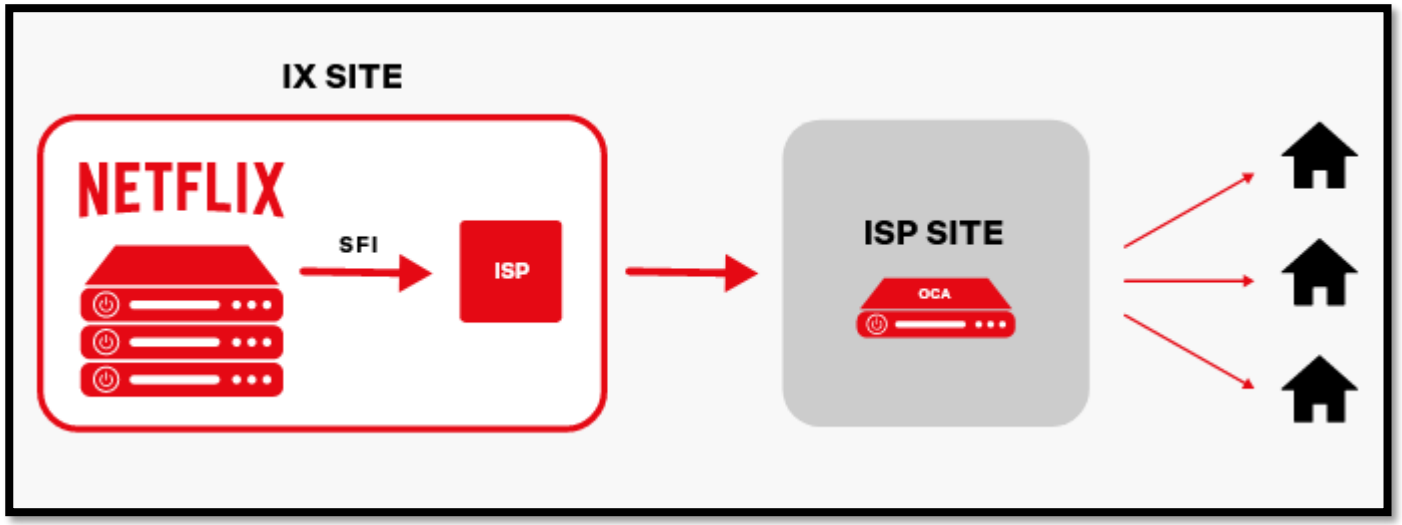
Users

Edge Computing

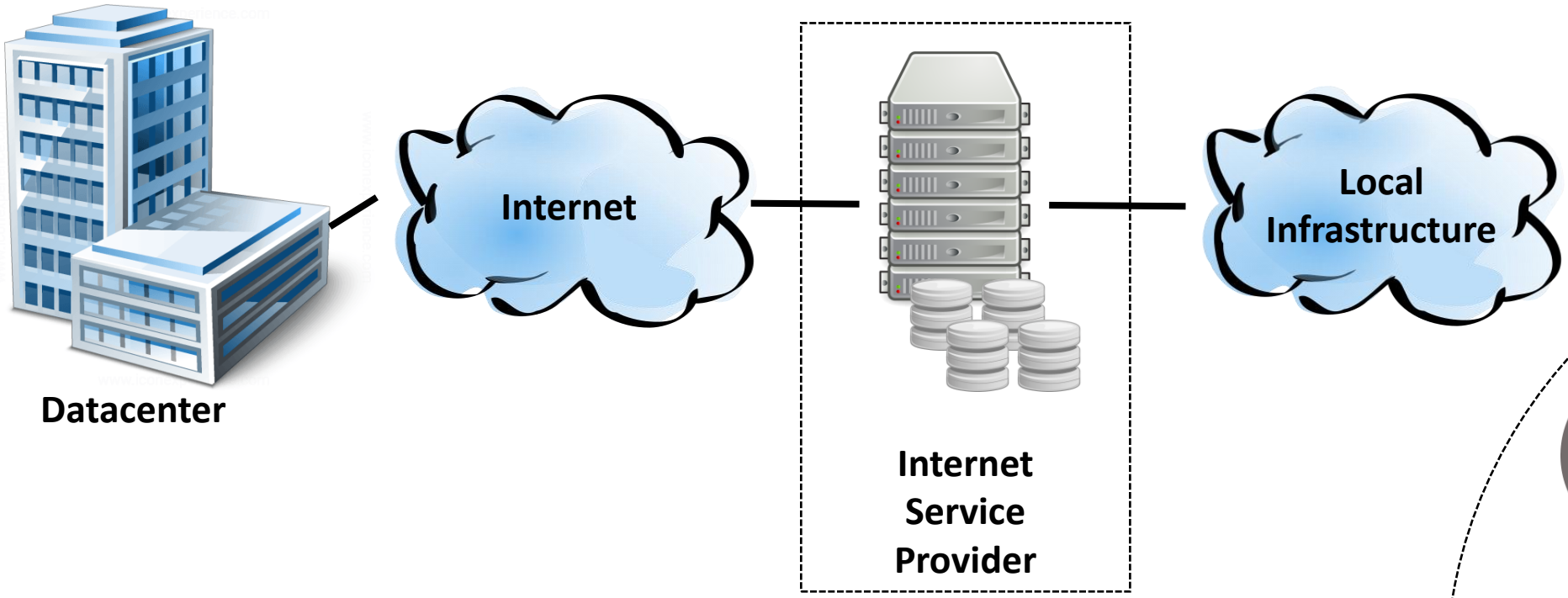


In-ISP caching

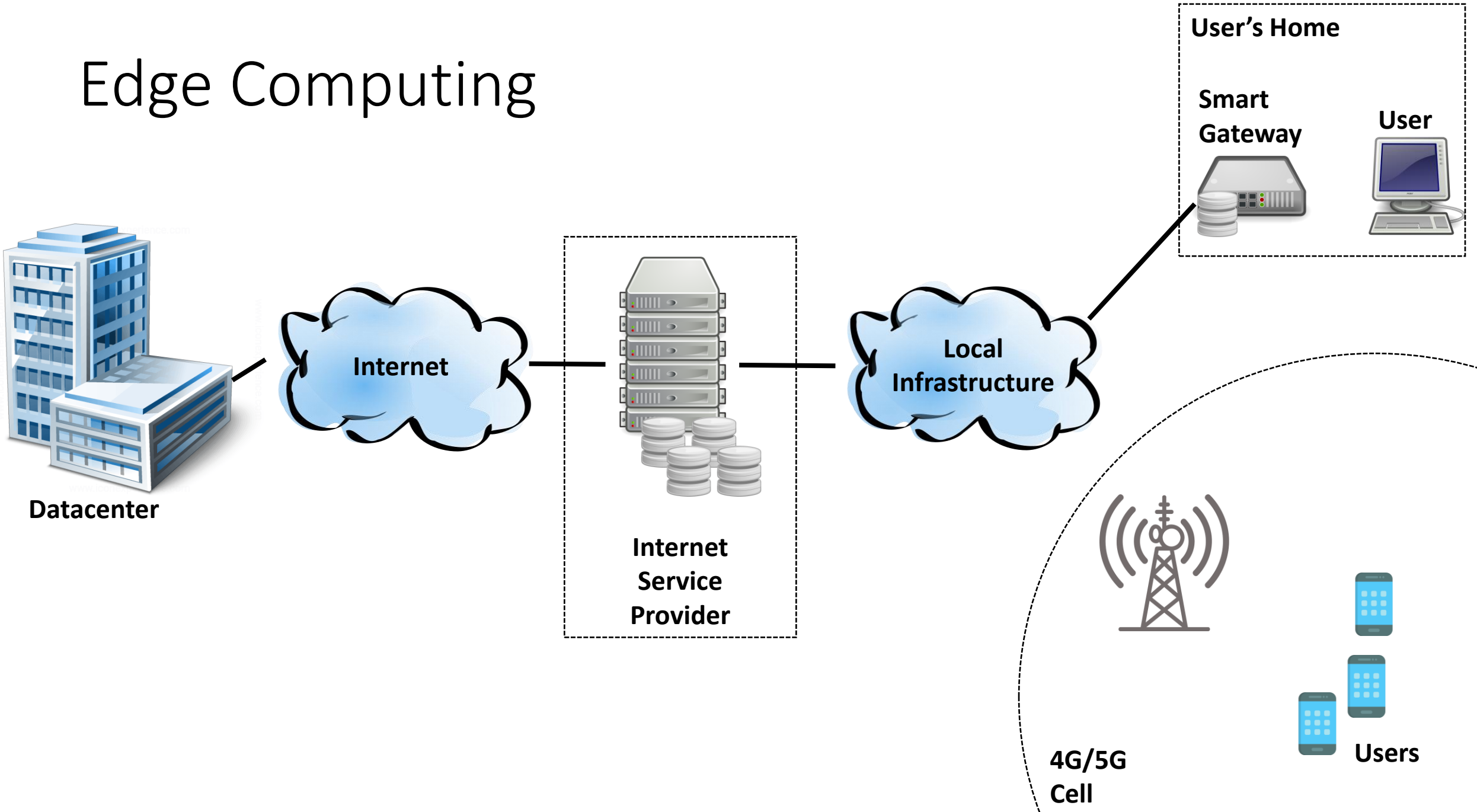
- Reduce latency and localize traffic
  - Facebook FNAs
  - Netflix OpenConnect
  - Akamai Accelerated Network
  - Google Global Cache
  - …



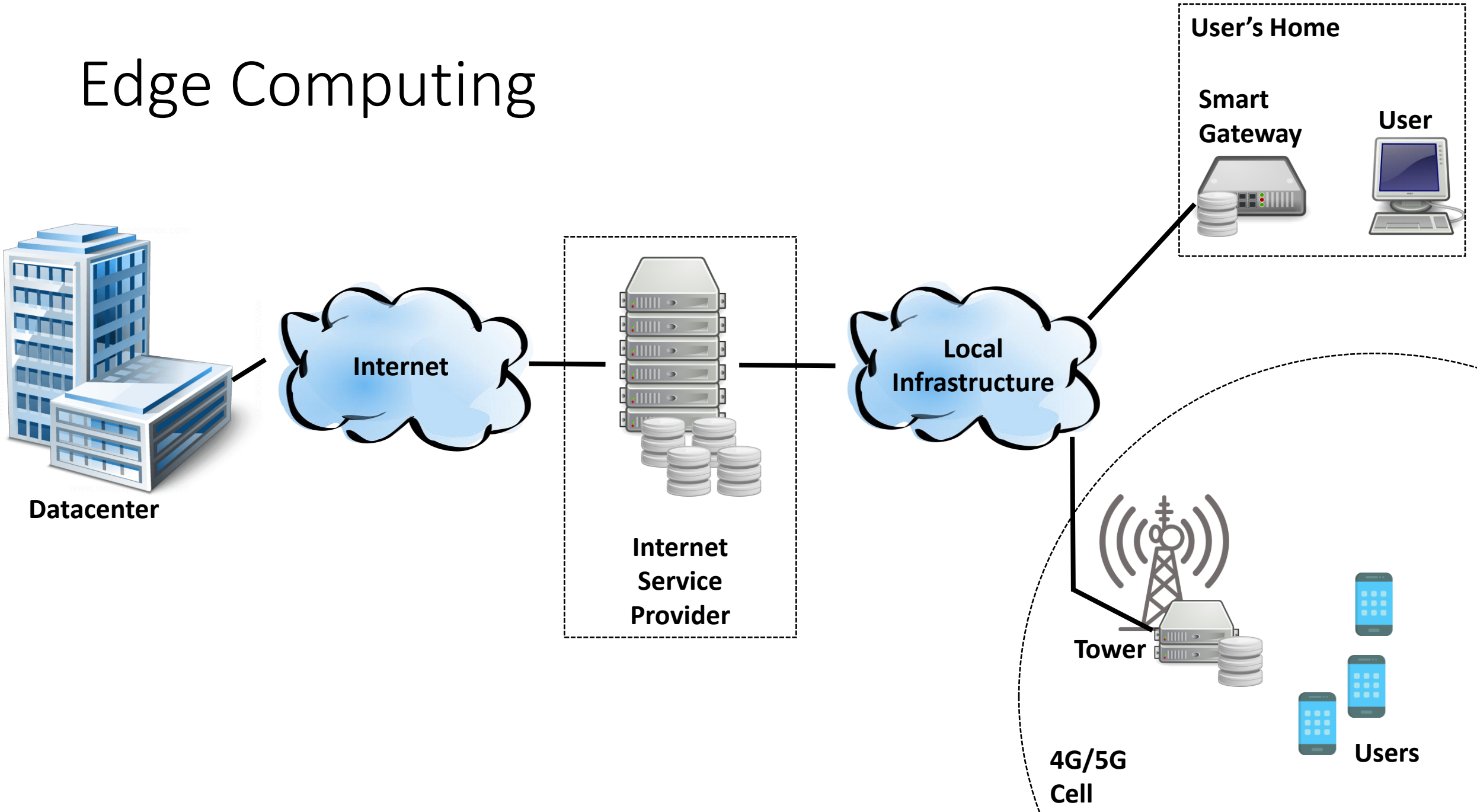
Edge Computing



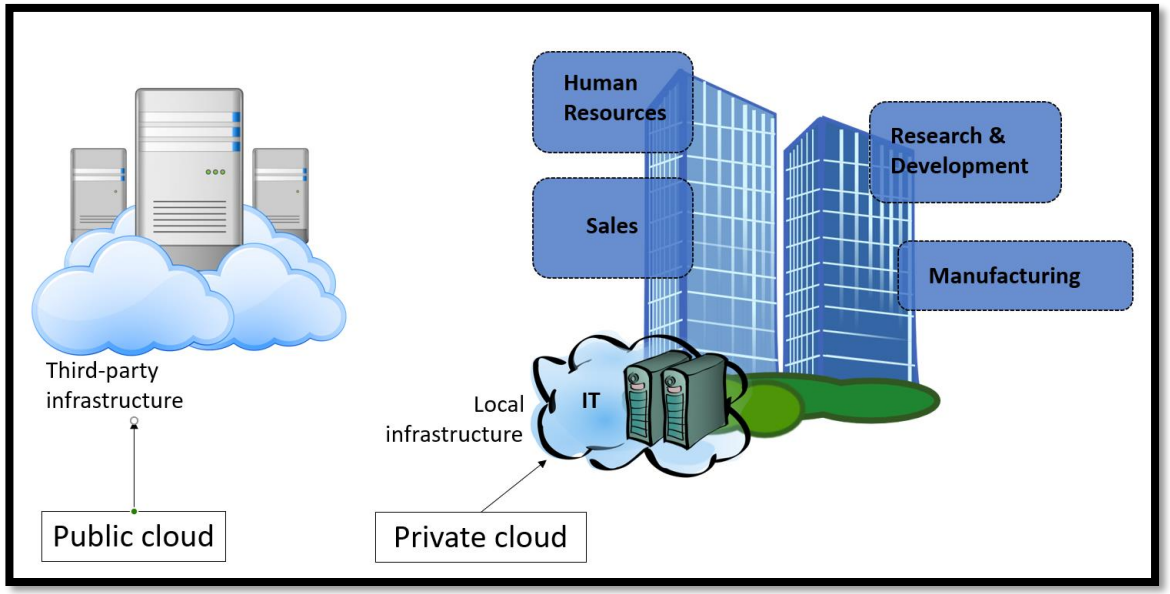
Edge Computing



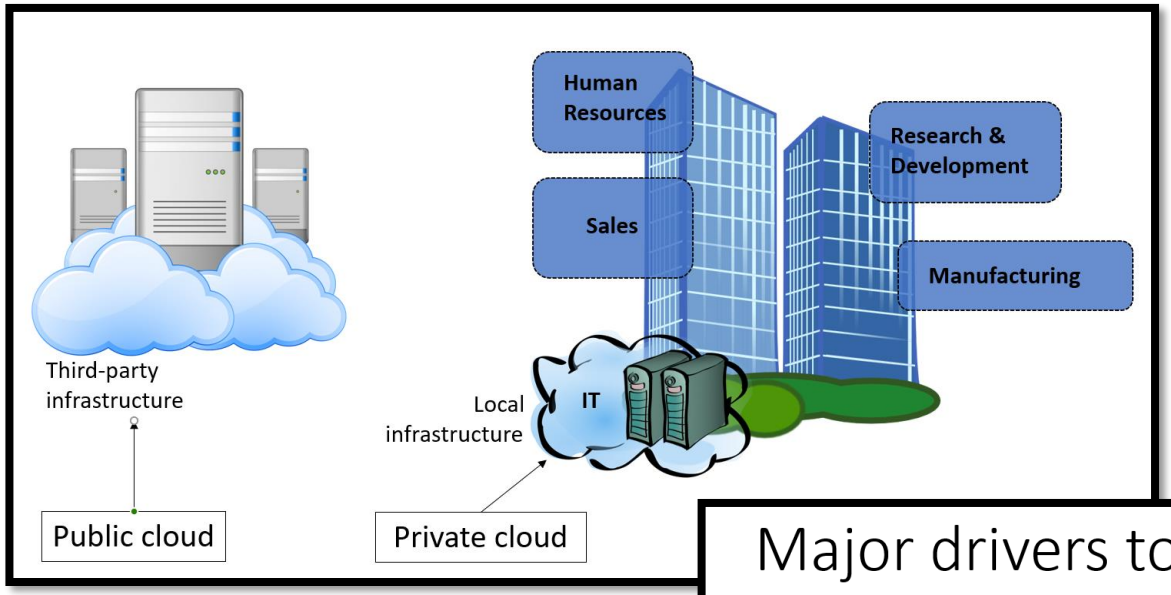
Edge Computing



Centralization



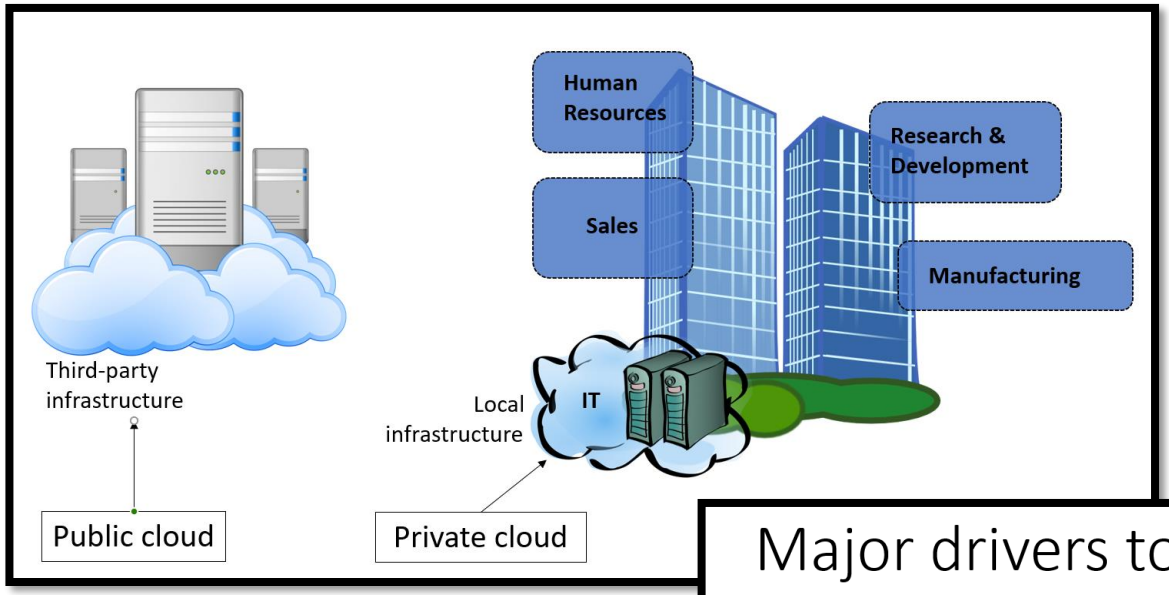
Centralization



Major drivers to adoption

- Centralizing commodity servers in clusters
  - Better OPEX and CAPEX
  - Less human resources
  - Quantity discounts
  - Less heterogeneity
  - Easier automation

Centralization



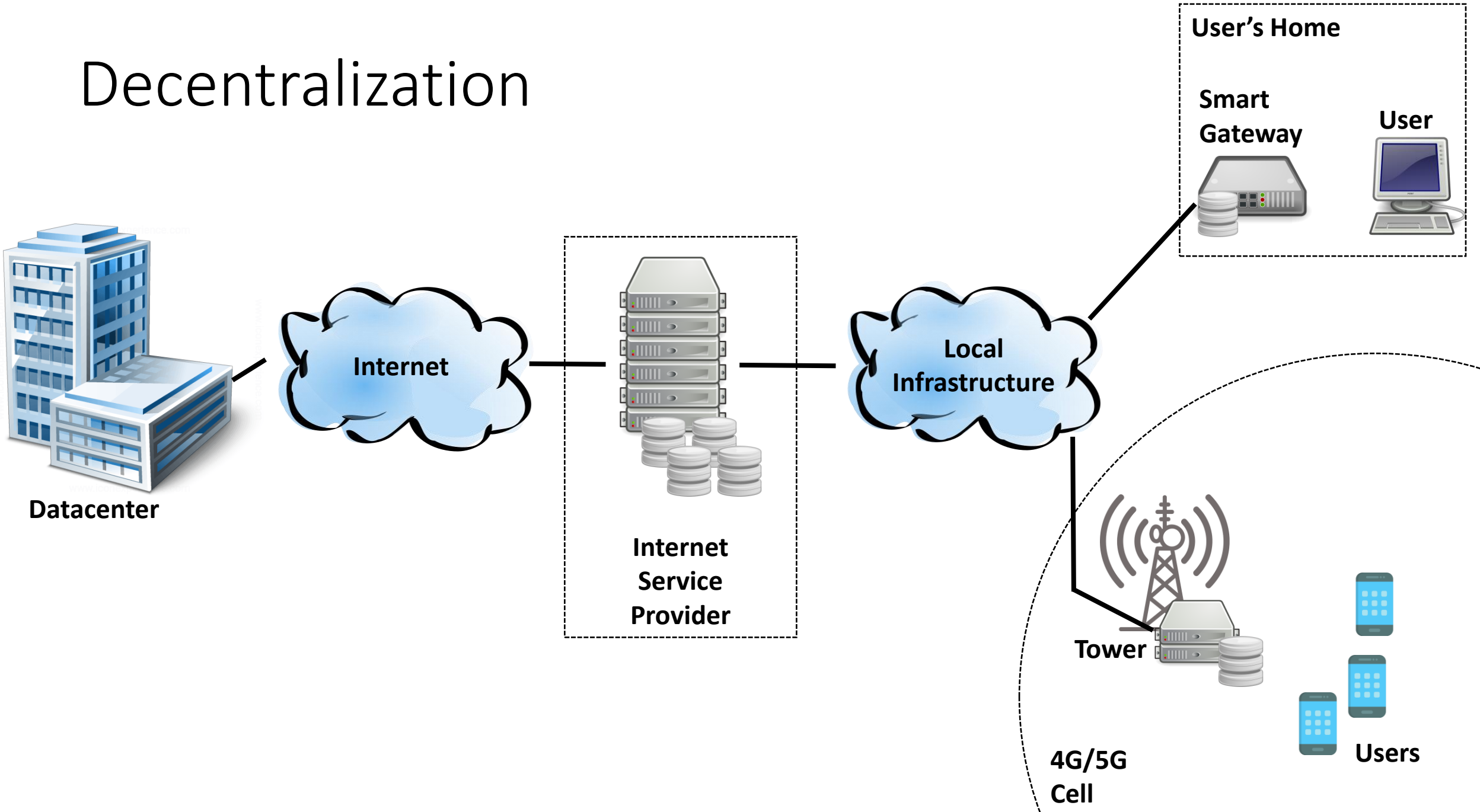
Cloud providers potentialize these advantages

- 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

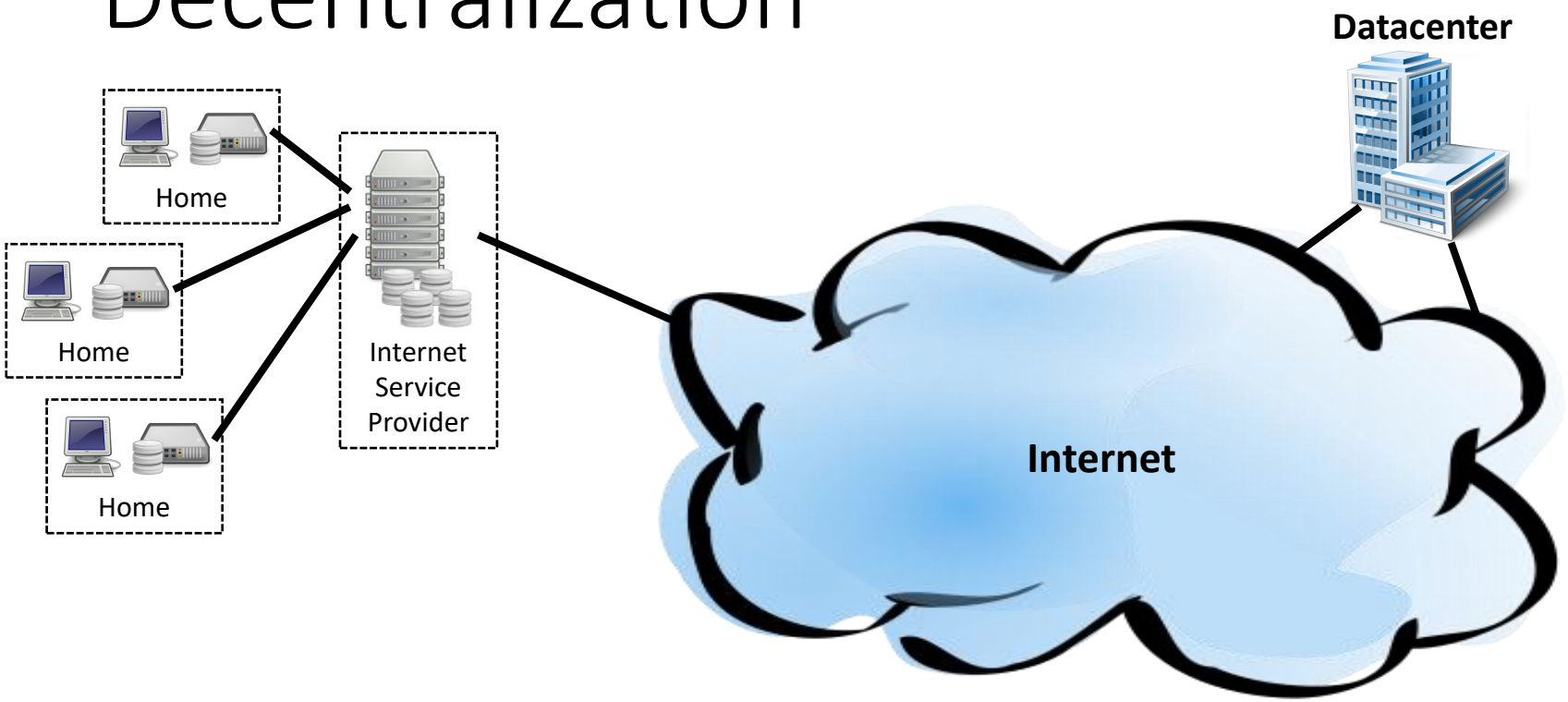
Major drivers to adoption

- Centralizing commodity servers in clusters
  - Better OPEX and CAPEX
  - Less human resources
  - Quantity discounts
  - Less heterogeneity
  - Easier automation

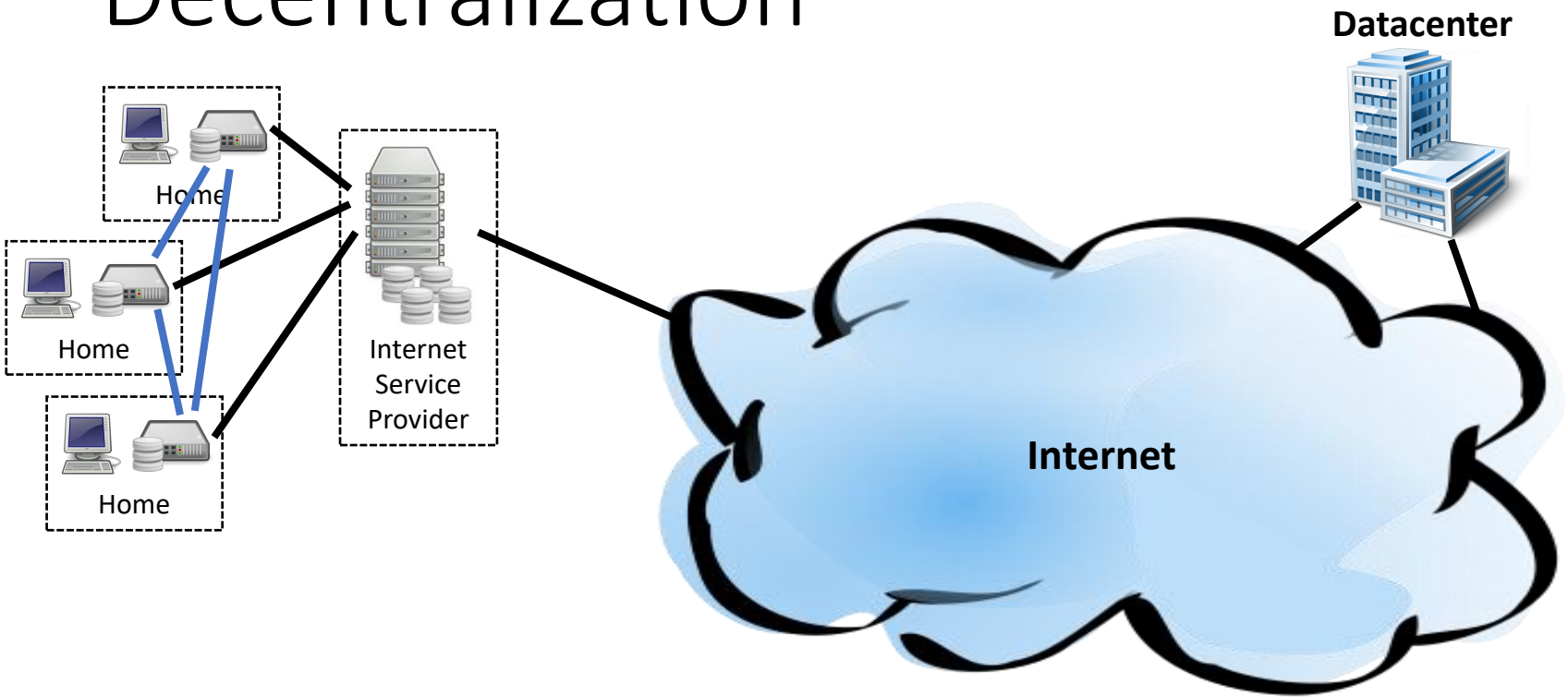
Decentralization



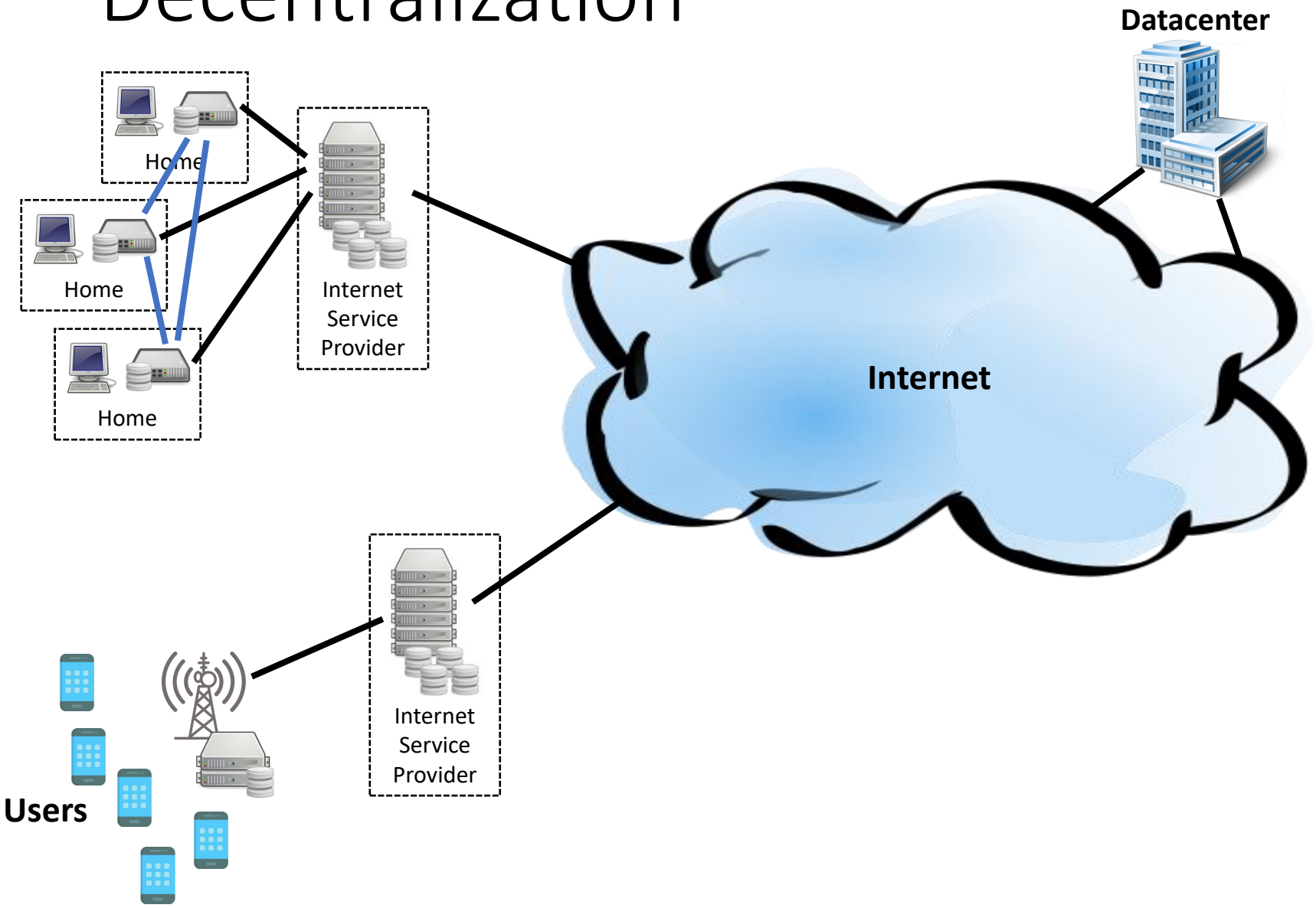
Decentralization



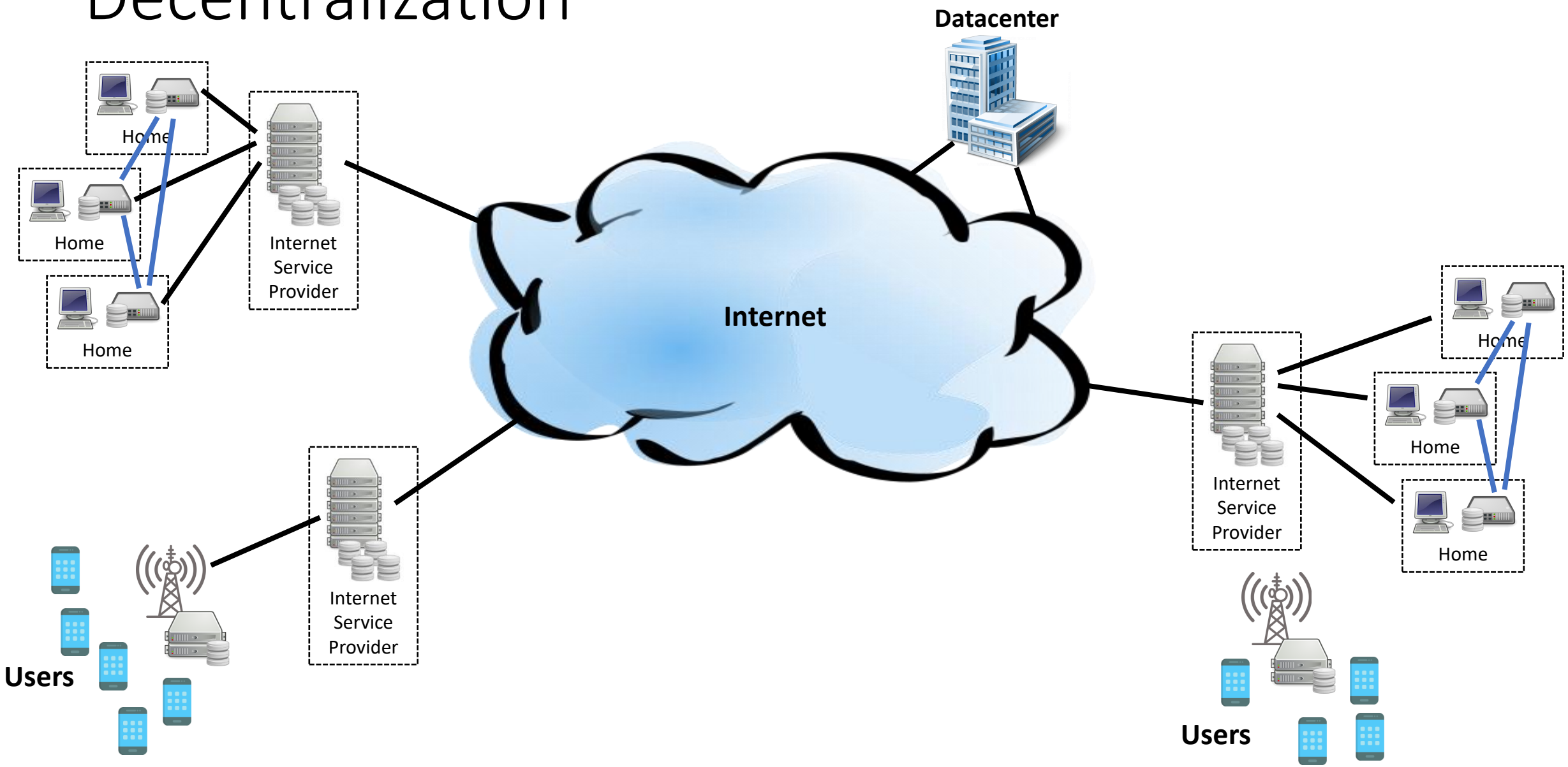
Decentralization



Decentralization



Decentralization



Decentralization