CompSci 401: Cloud Computing

# Racks, Aisles, and Pods

Prof. Ítalo Cunha

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

https://www.switch.com/tahoe-reno/

815MW total power

Up to 670000 m$^2$ of data center space

Rack aisles

Networking racks

Row of server racks

Row of server racks

# Data center building blocks

- Servers and other equipment go in racks
- Racks are organized in rows, with aisles between them
- Rows of racks are grouped into pods
  - Pod may include power distribution units to deliver electricity to the pod
  - Pod may include management facilities
  - Some vendors sell pre-built pods (e.g., the NVIDIA DGX SuperPod) that can be moved to a datacenter or colocation facility
- Pods make up a datacenter

CompSci 401: Cloud Computing

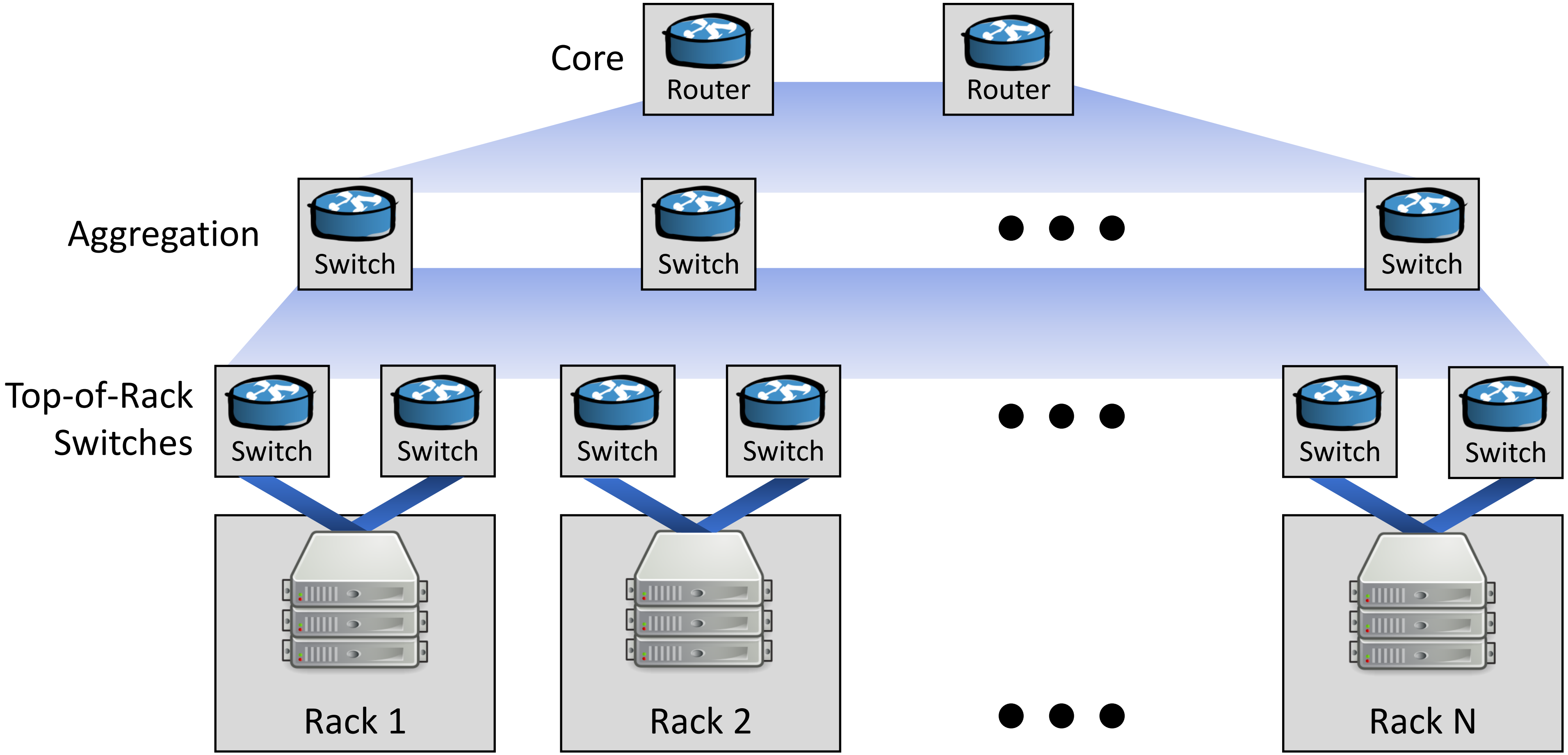
# Datacenter Pods

Prof. Ítalo Cunha

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

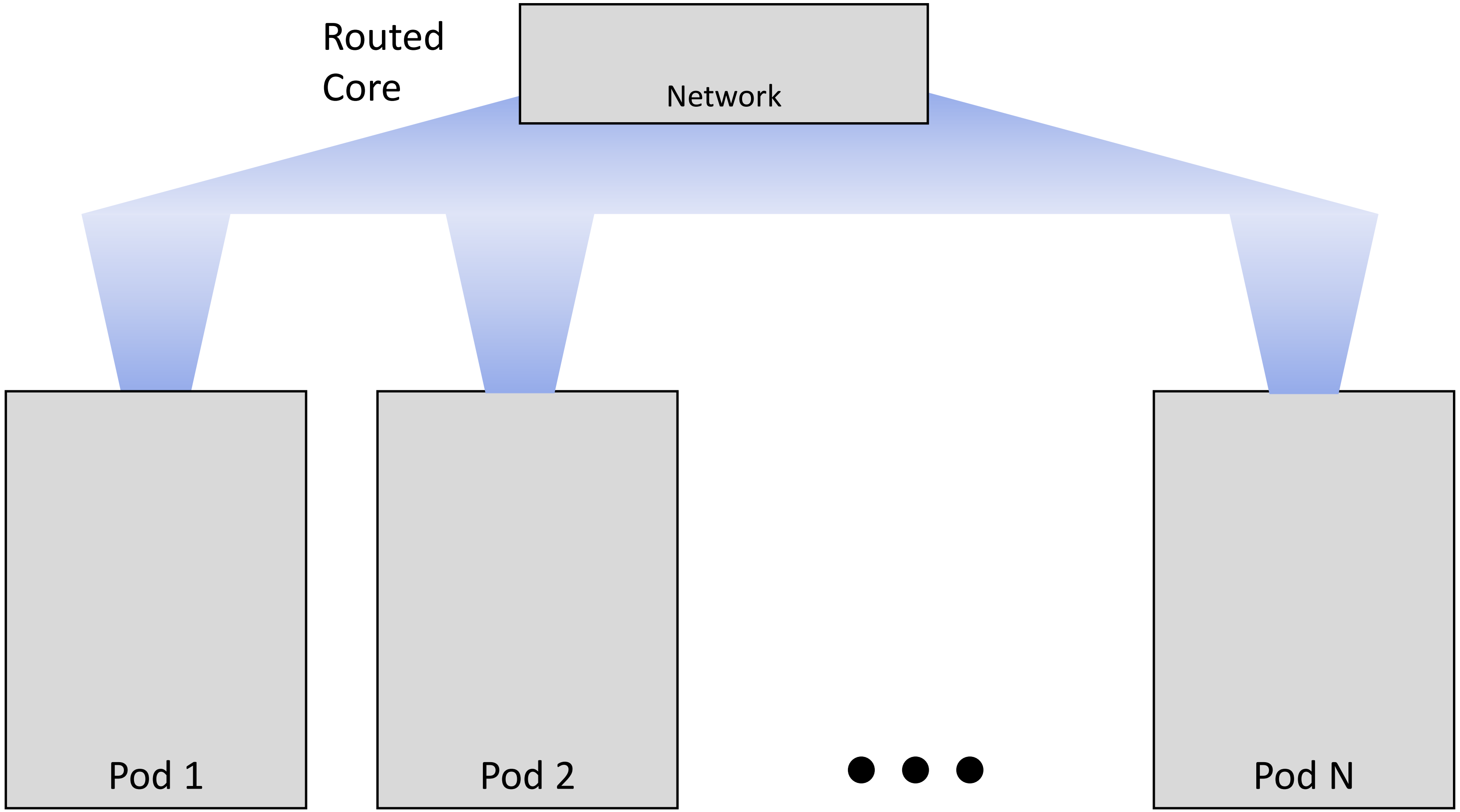
# Traditional approach

- Interconnected racks through ToR and aggregation switches



# Modern approach

- Isolated Pods interconnected through network core



# Pod as a building block

- Built as a unit
    - No “half-built” pods, no partial upgrades
    - Multiple pod “generations”
- Deployed as a unit
- Managed as a unit
- Decommissioned as a unit

- Pods have internal redundancy for resilience
- Failures in a pod do not impact other pods

# Pod as computing resources

- Pod can provide predefined capabilities/interfaces
  - Compute, storage, networking
  - Capabilities may vary across pod versions
    - For example: “memory optimized” vs “storage optimized” vs “compute optimized” pods
- Internals may vary as pod designs advance
  - Lower power consumption, more resources
  - But capabilities/interfaces ensure that applications should still run
- Composition of a pod is an important design decision

# Pod sizes

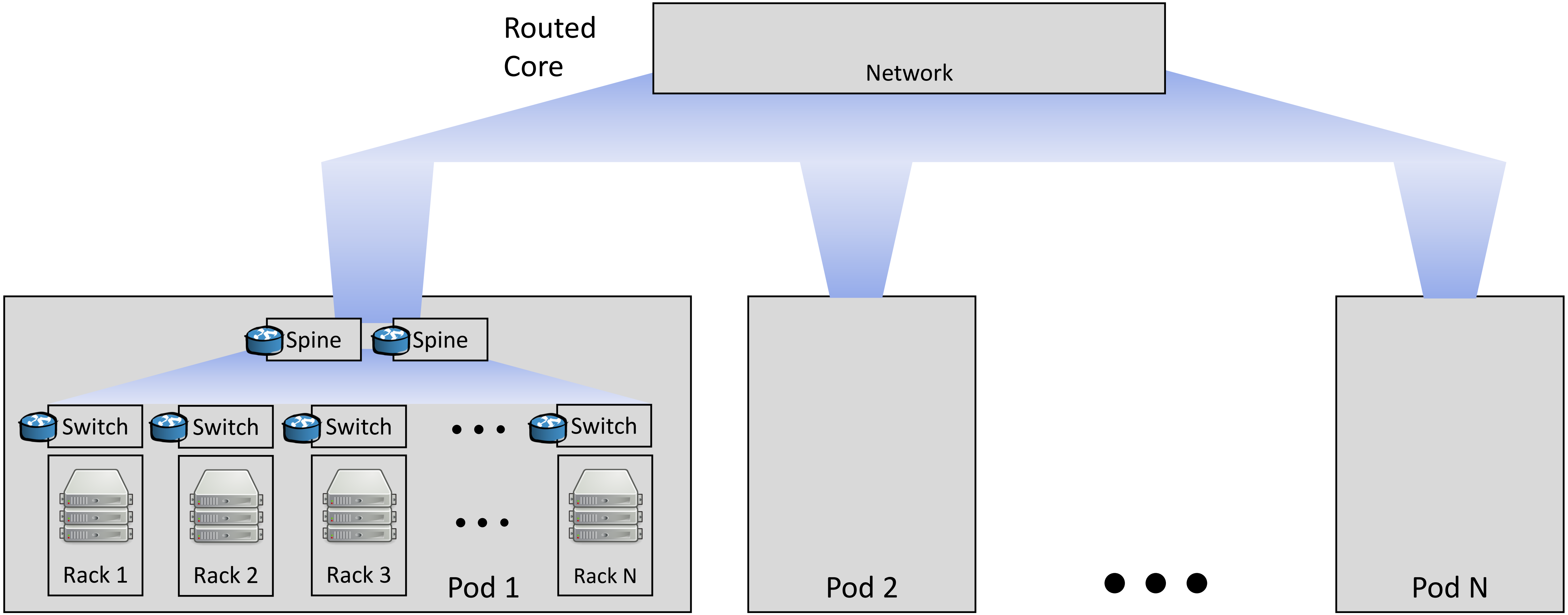
- Considerations
    - Incremental growth
        - Smaller pods make it easier to expand computing capacity
    - Impact of failures
        - Considering 25VMs per rack unit, a 16-rack pod hosts 16 thousand VMs
    - Management
        - Smaller pods are easier to test, manage, and repair
    - Computing capacity
        - Bigger pods support larger applications
        - Smaller pods may require partitioning an application across multiple pods

# Pod sizes

- Considerations
  - Incremental growth
    - Smaller pods make it easier to expand computing capacity
  - Impact of failures
    - Considering 25VMs per rack unit, a 16-rack pod hosts 16 thousand VMs
  - Management
    - Smaller pods are easier to test, manage, and repair
  - Computing capacity
    - Bigger pods support larger applications
    - Smaller pods may require partitioning an application across multiple pods
- Enterprise pod sizes usually between 12-16 racks
  - Pod sizes may vary between generations
  - Larger designs exist, Facebook’s Altoona uses 48 racks per pod

# Pod networking

- Designs vary depending on desired resiliency and bandwidth



CompSci 401: Cloud Computing

# Power and Cooling

Prof. Ítalo Cunha

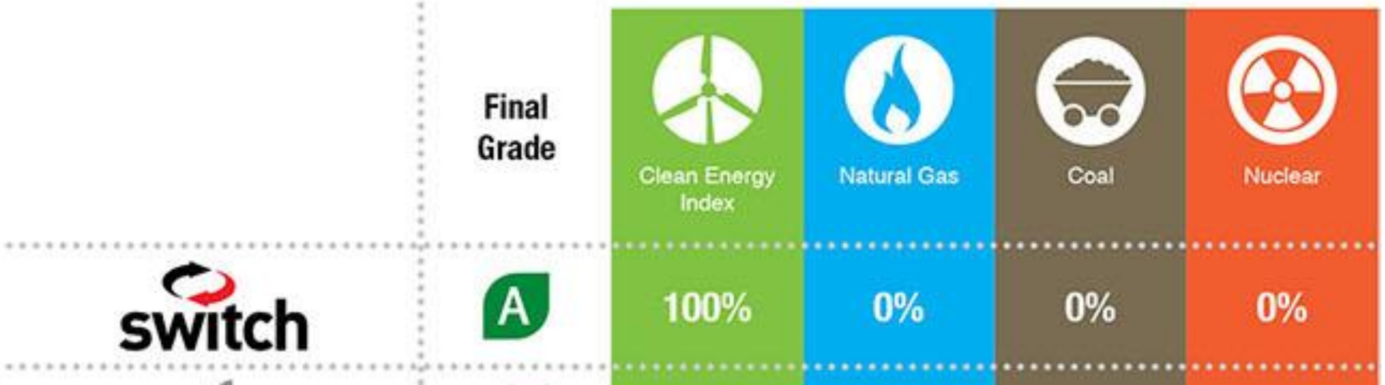
昆山杜克大学
DUKE KUNSHAN UNIVERSITY

# Powering datacenters

- Redundant power sources
  - Independent circuitry, possibly multiple providers
  - Datacenters may be located at locations where power is cheap
  - Not every place has enough power to support a large datacenter
- Batteries and generators
- Renewable energy
  - Google plans to operate 24/7 on renewable sources by 2030

# Powering datacenters

- Redundant power sources
  - Independent circuitry, possibly multiple providers
  - Datacenters may be located at locations where power is cheap
  - Not every place has enough power to support a large datacenter
- Batteries and generators
- Renewable energy
  - Google plans to operate 24/7 on renewable sources by 2030



| | Final Grade | Clean Energy Index | Natural Gas | Coal | Nuclear |
|---|---|---|---|---|---|
| switch | A | 100% | 0% | 0% | 0% |

# Increased power use creates heat

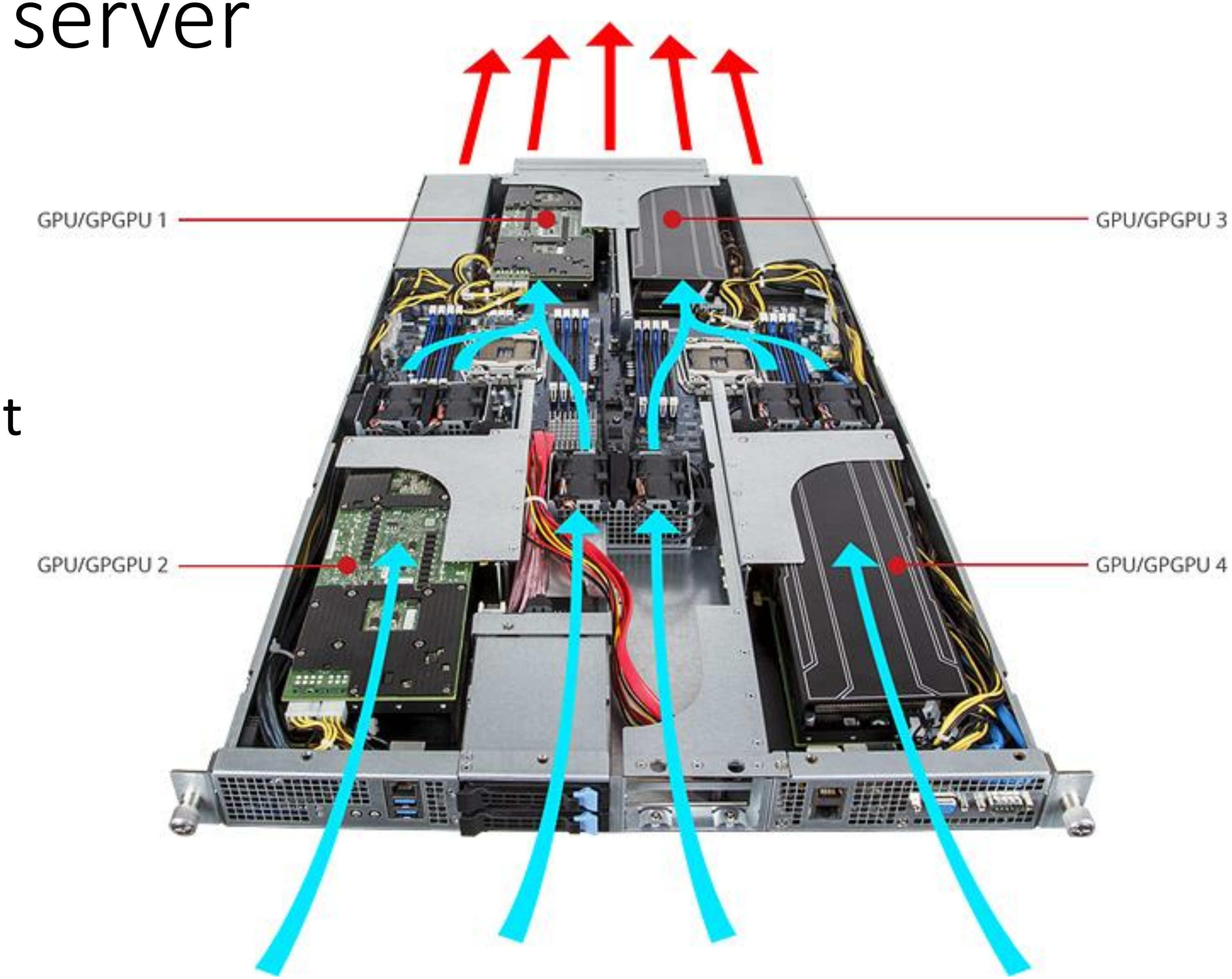
- Components get more efficient, but heat is unavoidable
  - Why your processor has a cooler

# Increased power use creates heat

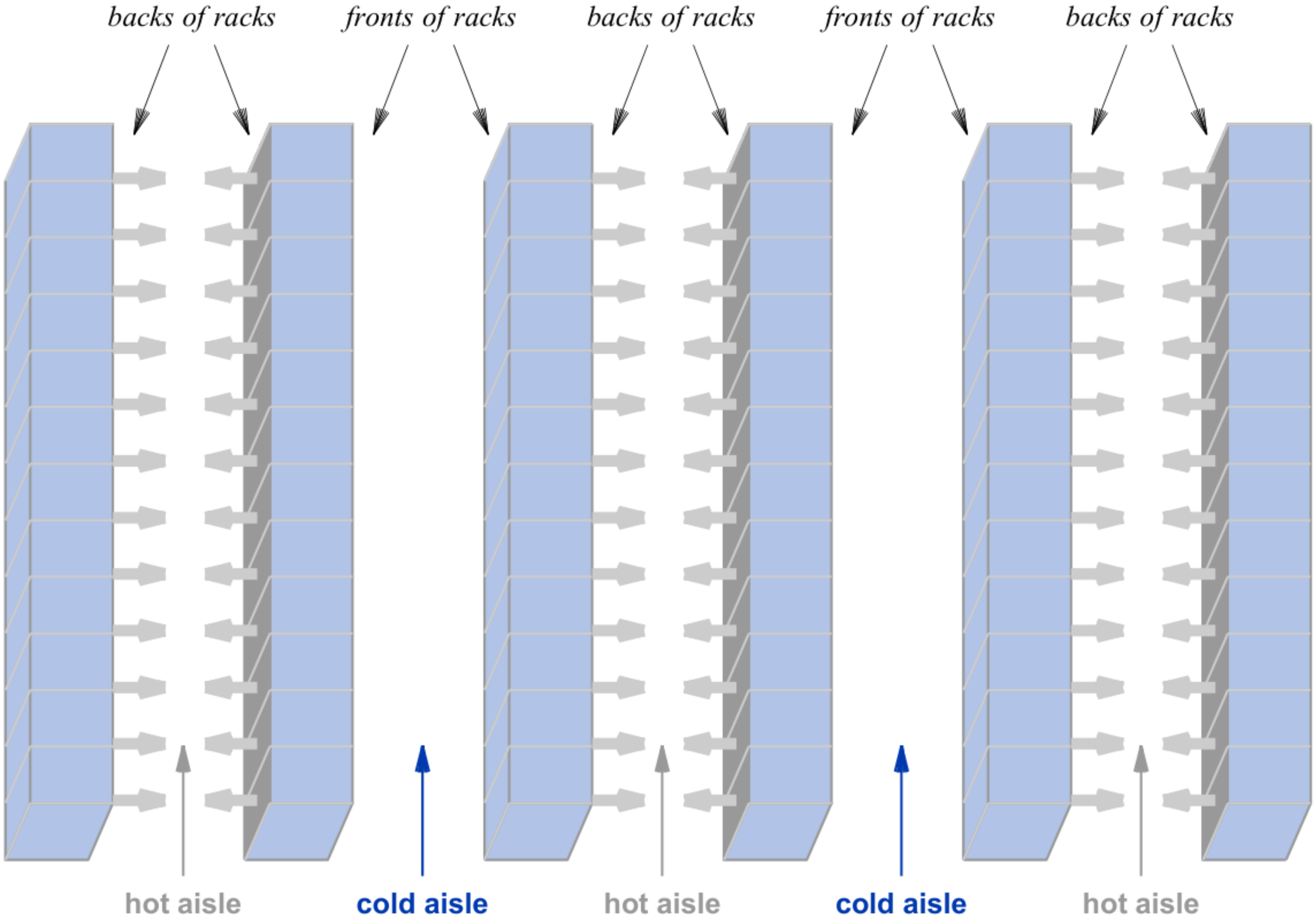
- Components get more efficient,
  but heat is unavoidable
  - Why your processor has a cooler
- Heat management and related technologies are key to datacenter
  - Cooling can get very expensive, so datacenters optimize for it
  - Efficient cooling allows higher-density, more cost-effective designs

# Air flow through server

- Take cold air from the front, release hot air at the back
- Important to prevent hot air from one equipment going into another

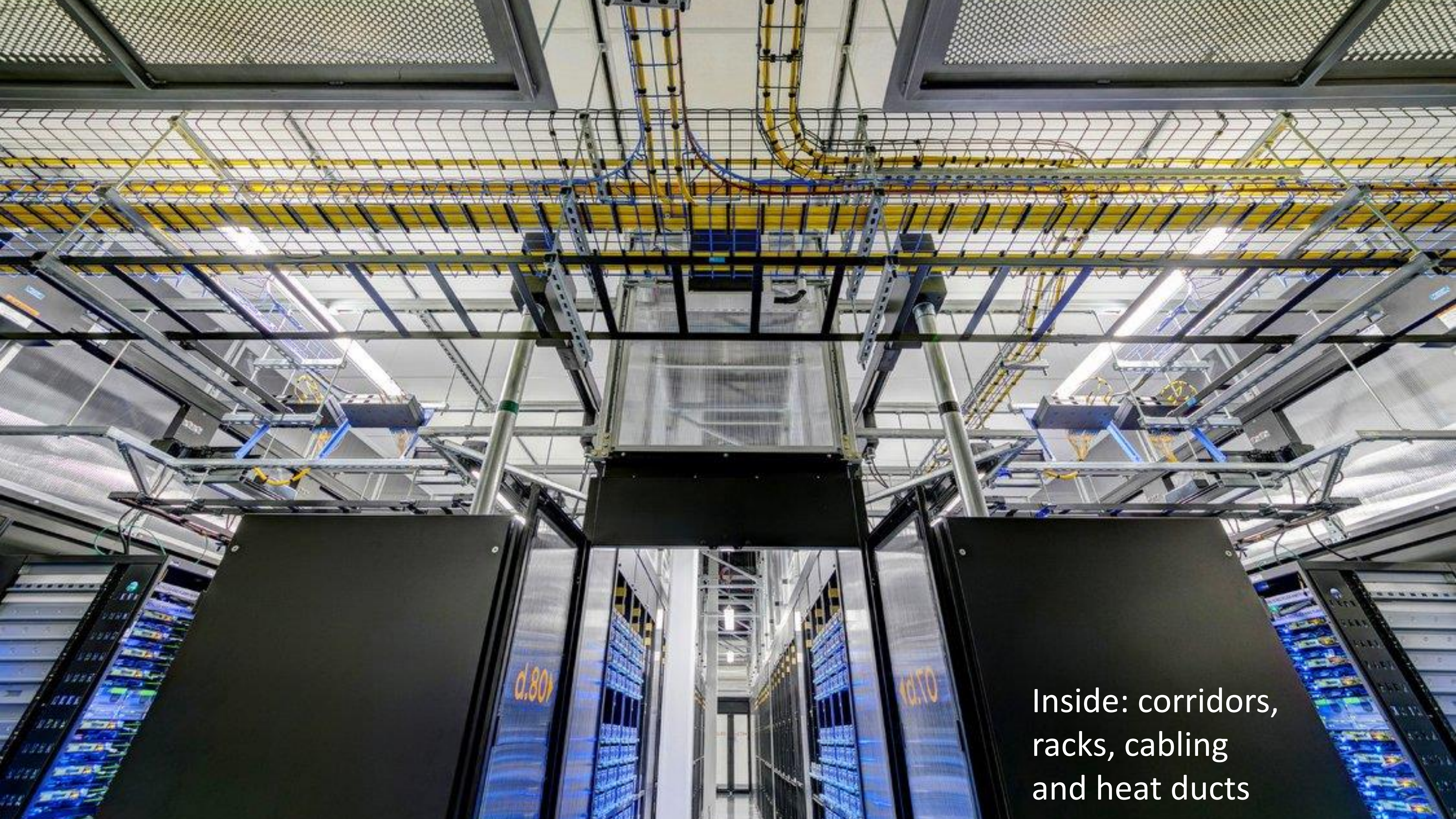


# Heat containment zones



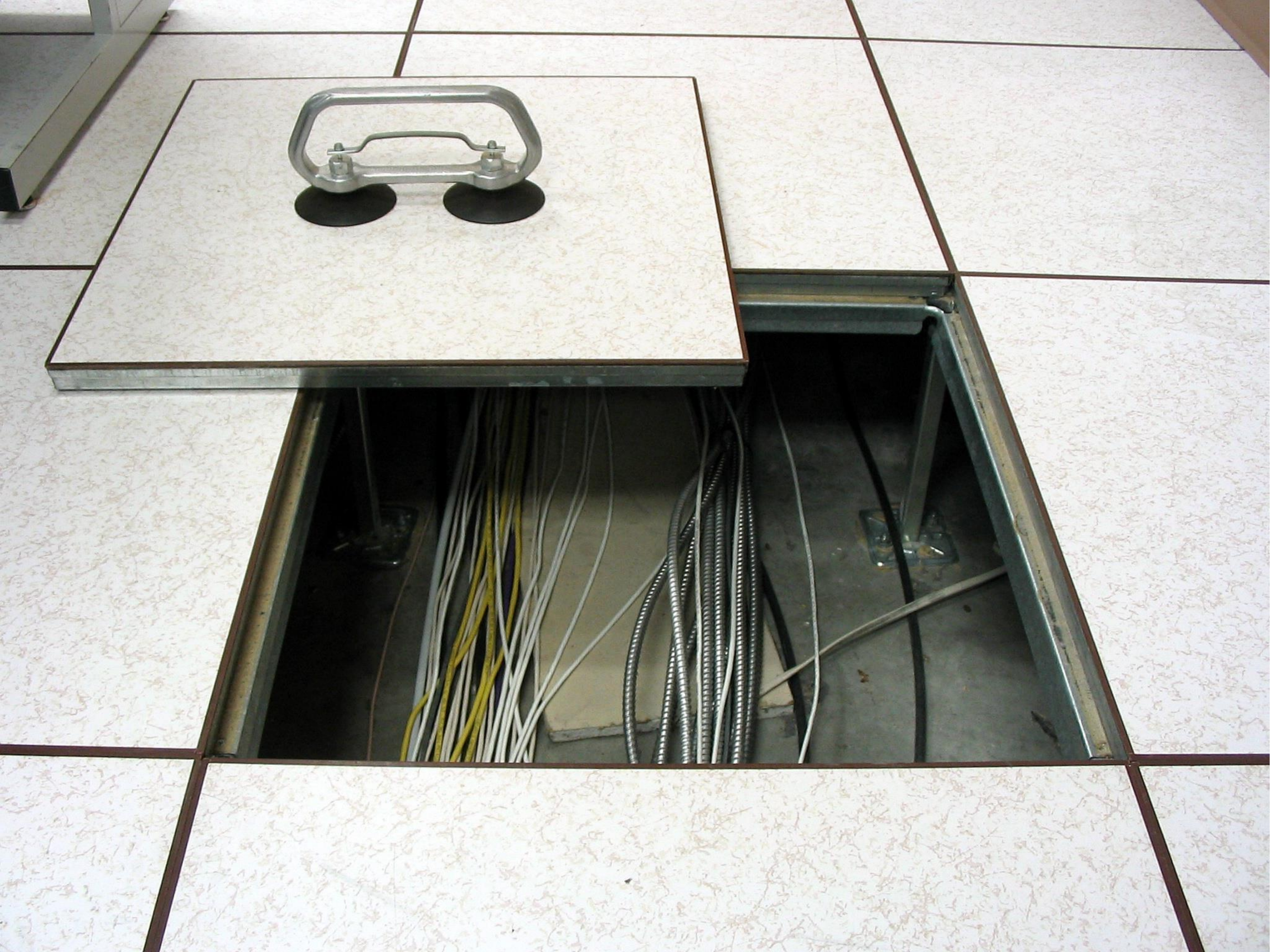
# Dealing with hot spots

- Air flow control
- Leaving empty space between servers
  - Reduce density to help with dissipation



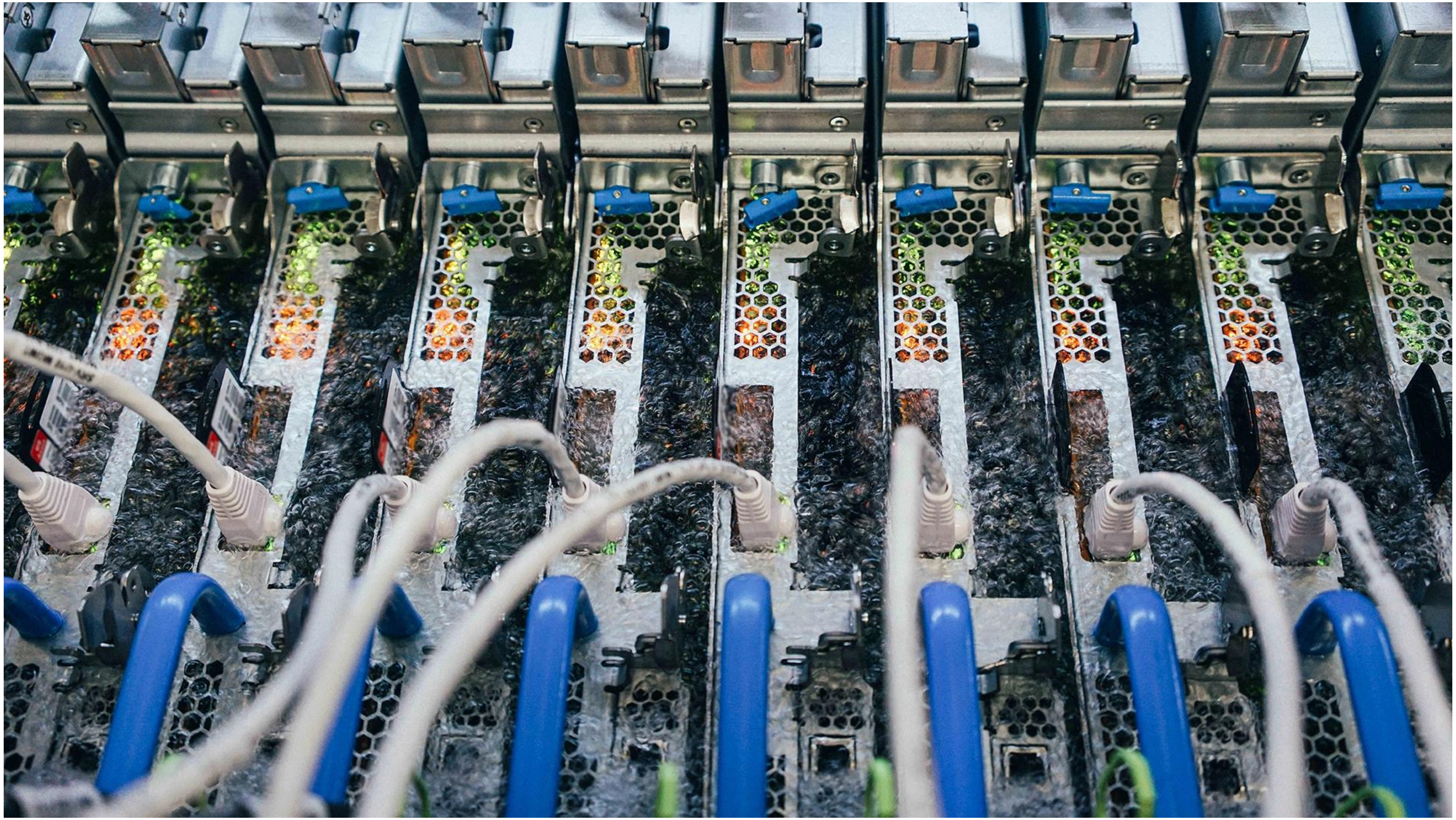
Inside: corridors,
racks, cabling
and heat ducts

Structured racks



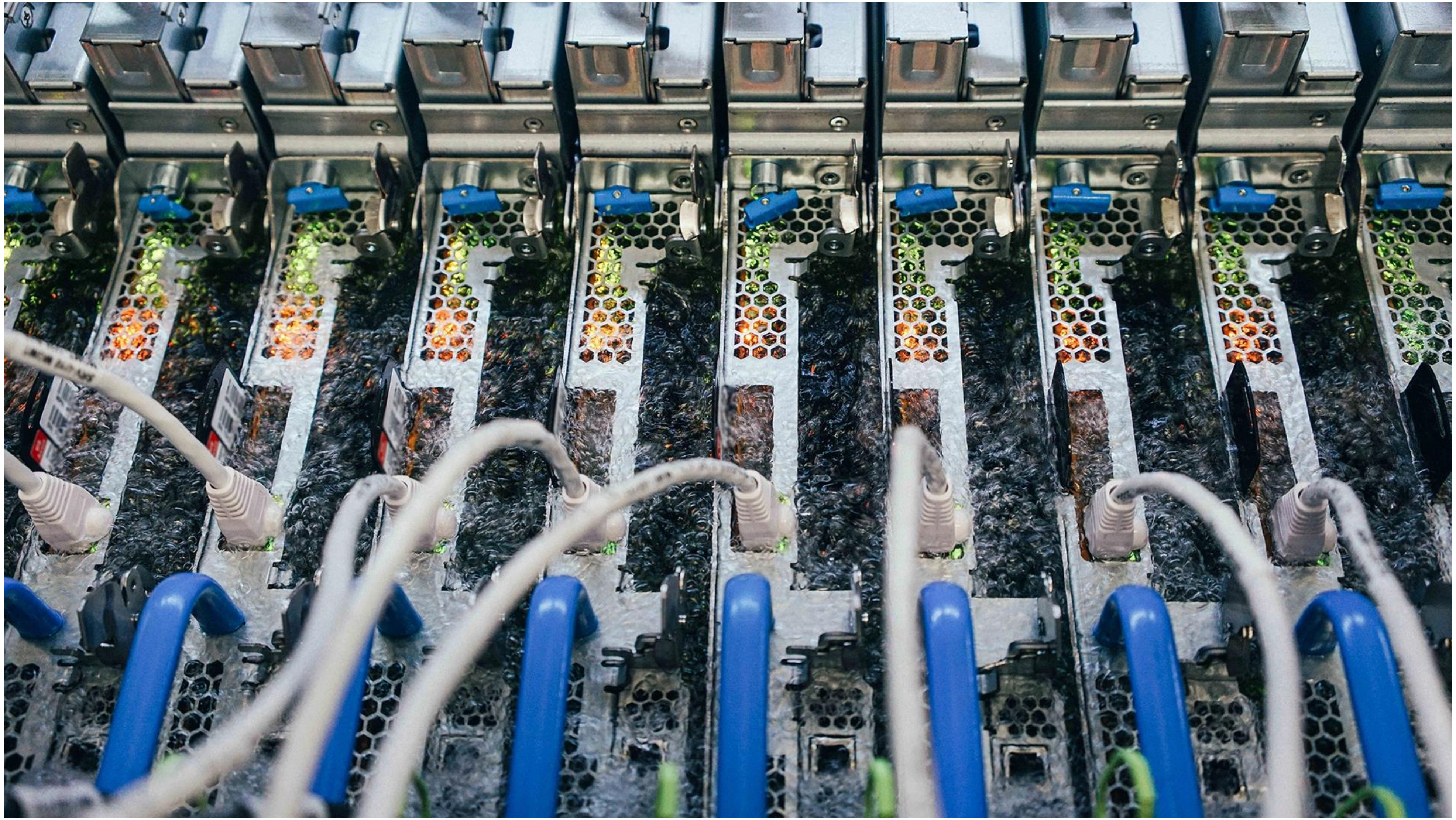
Raised floors to distribute cabling and cool air

# Liquid cooling



Microsoft prototype, servers submersed in non-conductive liquid boiling at 55$^{o}$C

# Liquid cooling



Microsoft prototype, servers submersed in non-conductive liquid boiling at 55$^{o}$C

CompSci 401: Cloud Computing

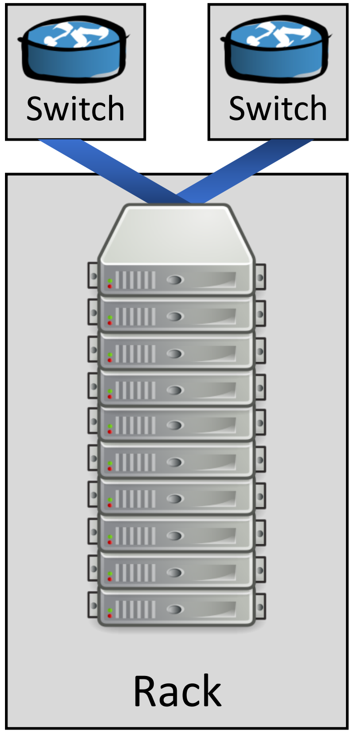
# Datacenter Traffic

Prof. Ítalo Cunha

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

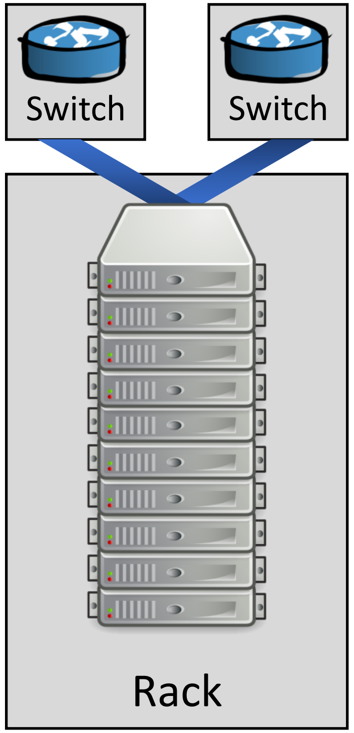
# Rack and server networking

- Racks have one or two top-of-rack (ToR) switches
- Servers can have interfaces with multiple ports
  - Two ports to connect to two ToR switches
  - Typically, 10Gbps between server and ToR
  - ToR switches have multiple uplinks, typically 40Gbps each



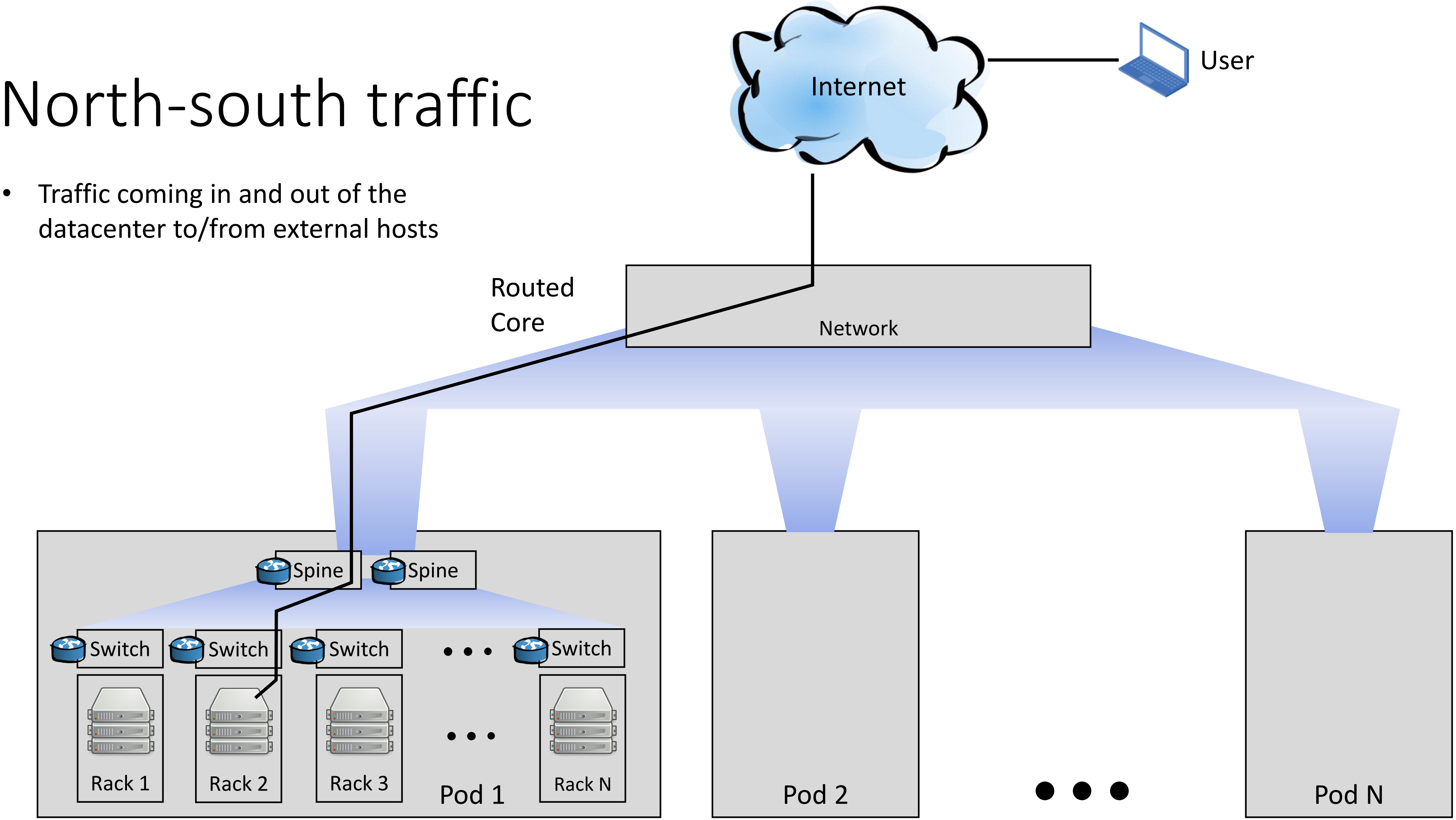
# Rack and server networking

- Racks have one or two top-of-rack (ToR) switches
- Servers can have interfaces with multiple ports
  - Two ports to connect to two ToR switches
  - Typically, 10Gbps between server and ToR
  - ToR switches have multiple uplinks, typically 40Gbps each
- Servers offload processing to network cards
  - Higher network utilization and more CPU for computation
  - IP checksum computation and validation
  - TCP segmentation
  - Encryption and decryption
  - And much more on programmable cards



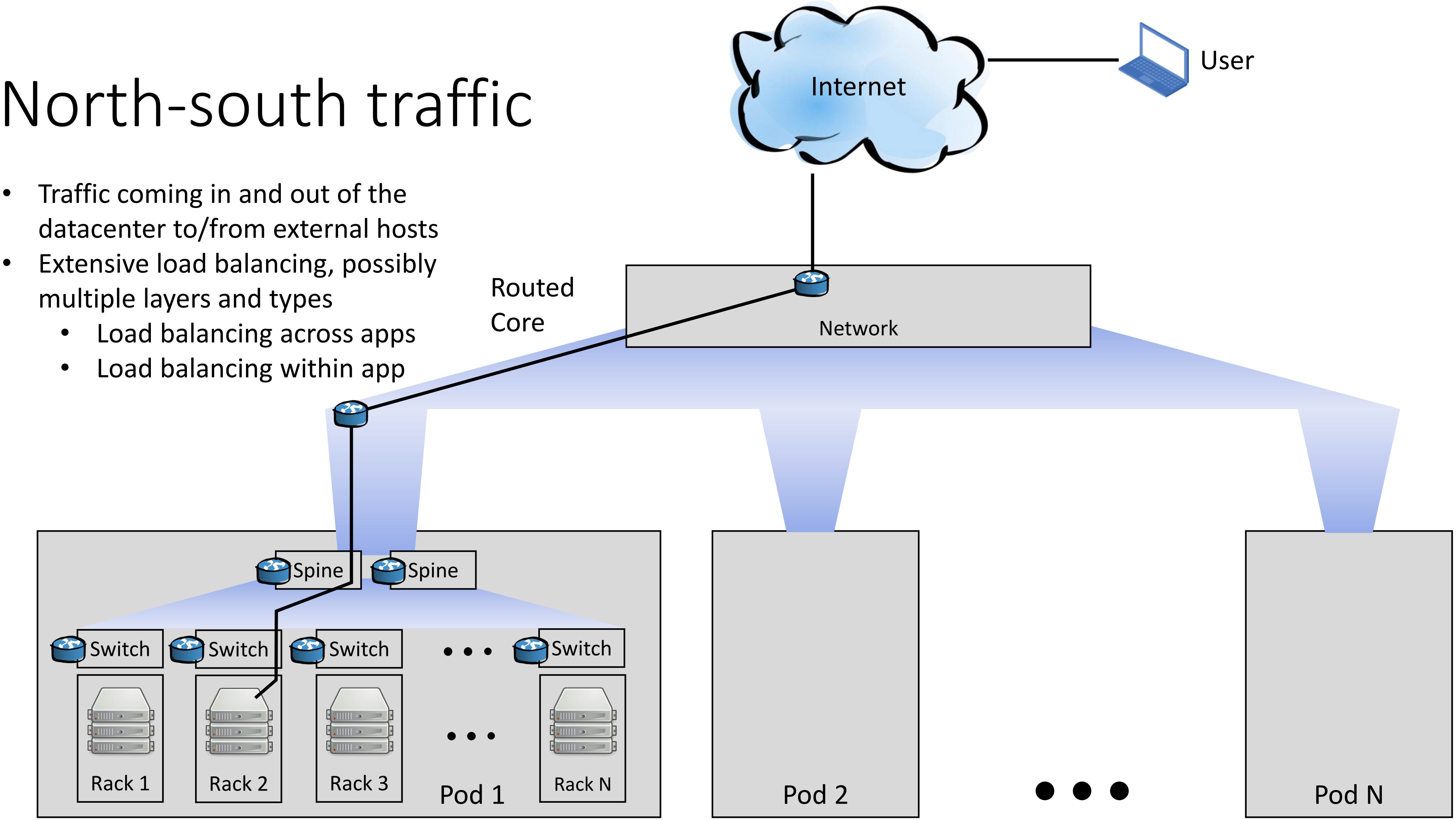
# North-south traffic

- Traffic coming in and out of the datacenter to/from external hosts



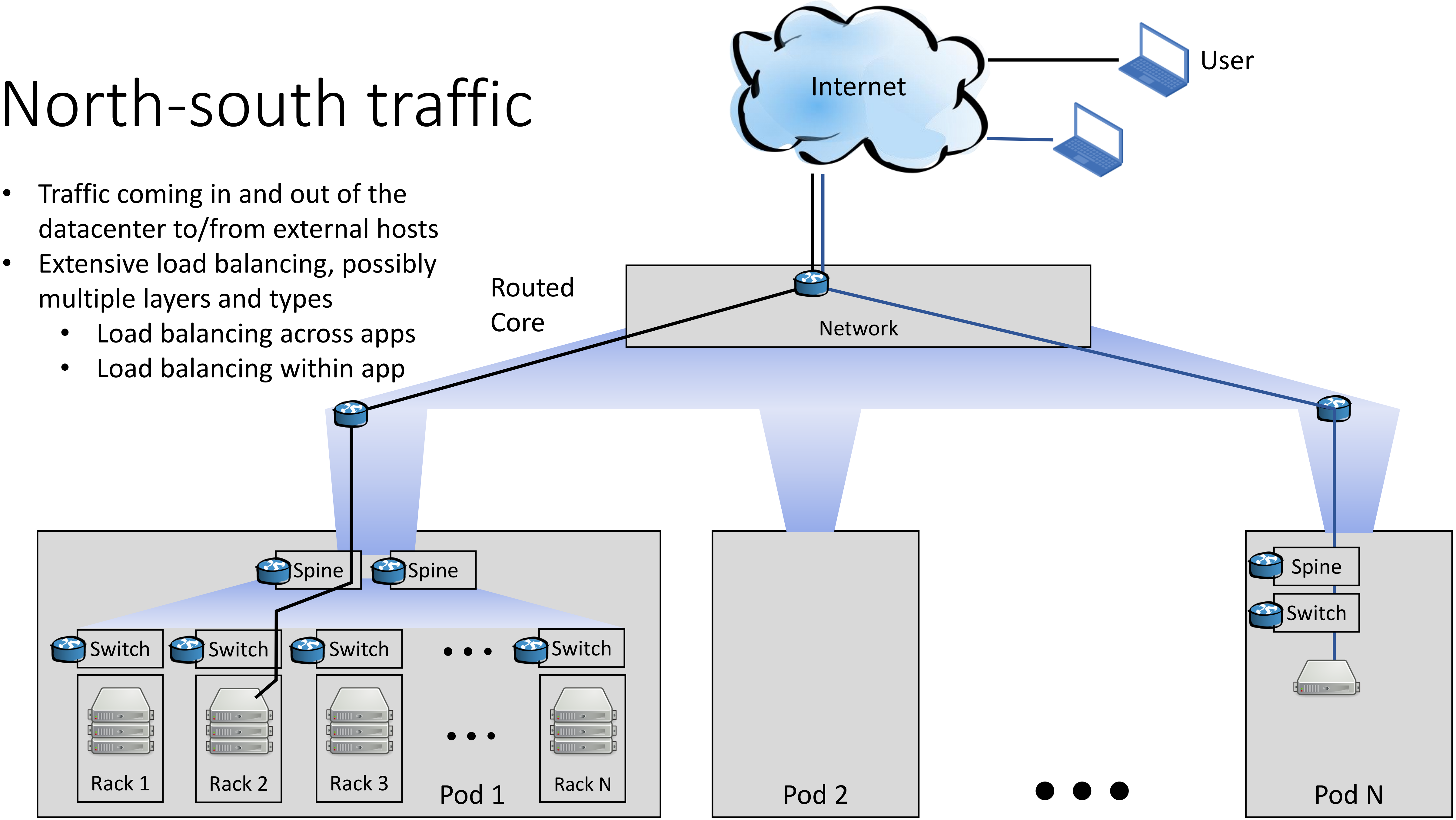
# North-south traffic

- Traffic coming in and out of the datacenter to/from external hosts
- Extensive load balancing, possibly multiple layers and types
  - Load balancing across apps
  - Load balancing within app



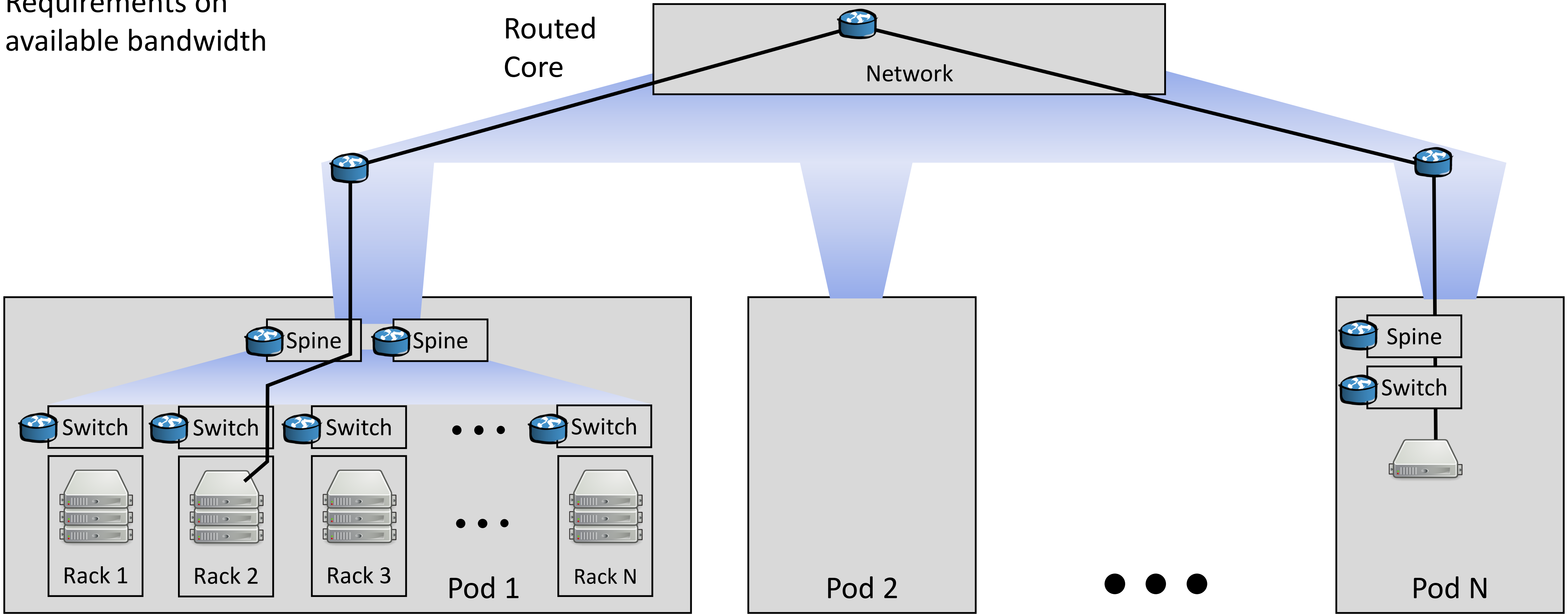
# North-south traffic

- Traffic coming in and out of the datacenter to/from external hosts
- Extensive load balancing, possibly multiple layers and types
    - Load balancing across apps
    - Load balancing within app



# East-west traffic

- One application may be spread across multiple pods
- Applications may communicate within the datacenter
- Requirements on available bandwidth



CompSci 401: Cloud Computing

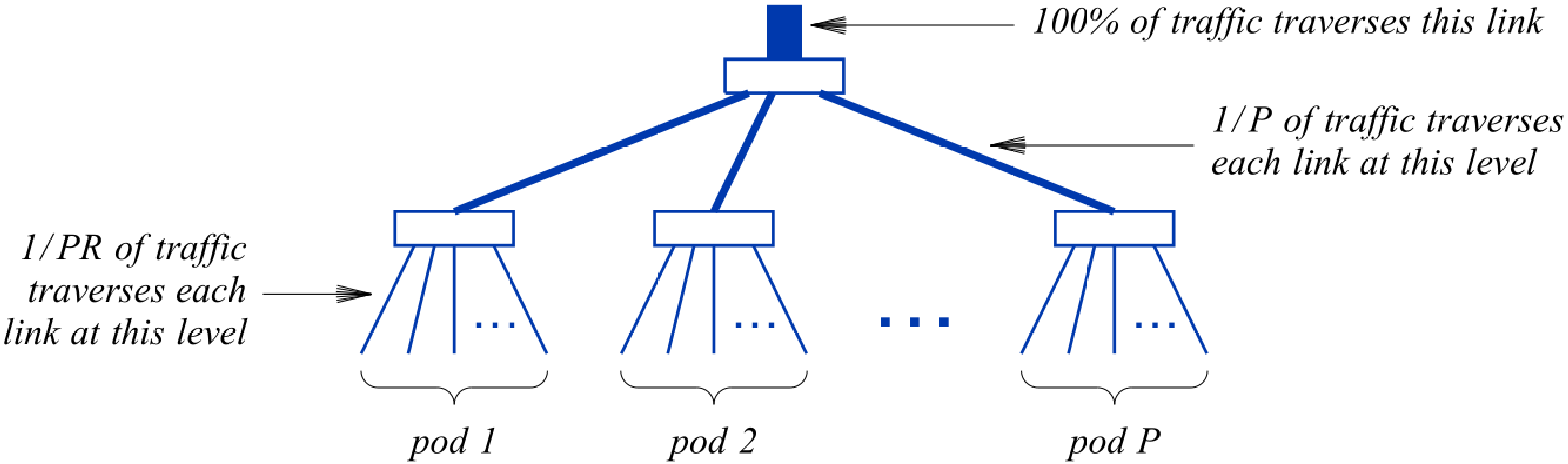
# Datacenter Networking

Prof. Ítalo Cunha

昆山杜克大学
DUKE KUNSHAN
UNIVERSITY

# Network topology and link capacities

- Traffic aggregates on the higher layers of the hierarchy
- Fat trees



100% of traffic traverses this link

1/P of traffic traverses each link at this level

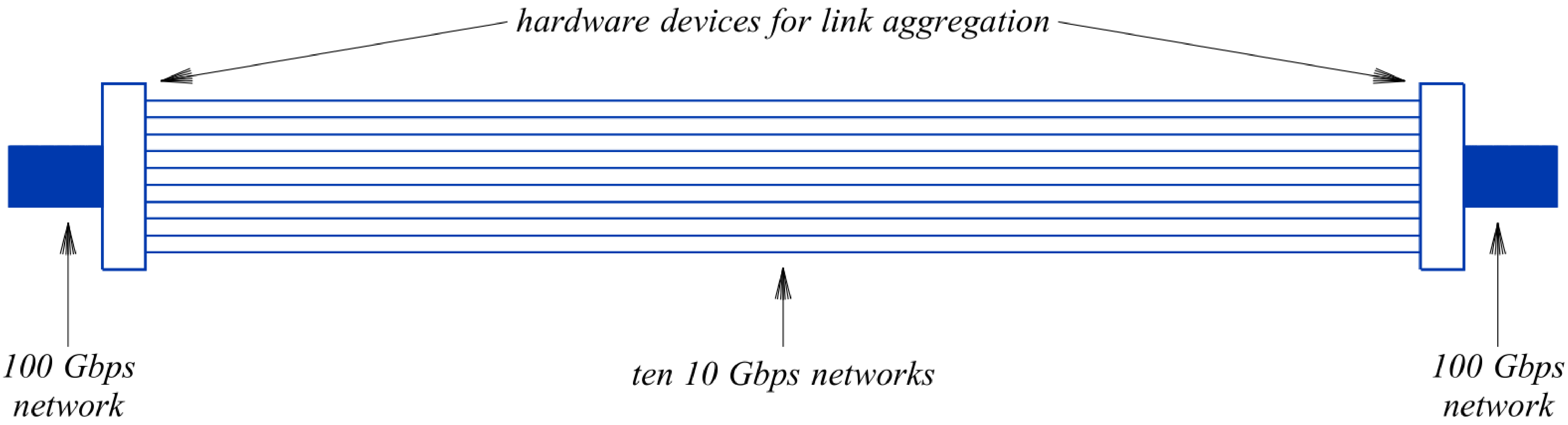
1/PR of traffic traverses each link at this level

# High link capacities and link aggregation

- Providers face constraints when building a datacenter network
- Link capacities available
  - Ethernet links have 1, 10, 40, 100, and 400Gbps
  - Need to design considering how to combine available capacities
- Cost and reliability of high-capacity hardware
  - Higher-capacity hardware is more expensive
  - Higher-capacity hardware may be less reliable or less tested than lower-capacity hardware that have been around for longer

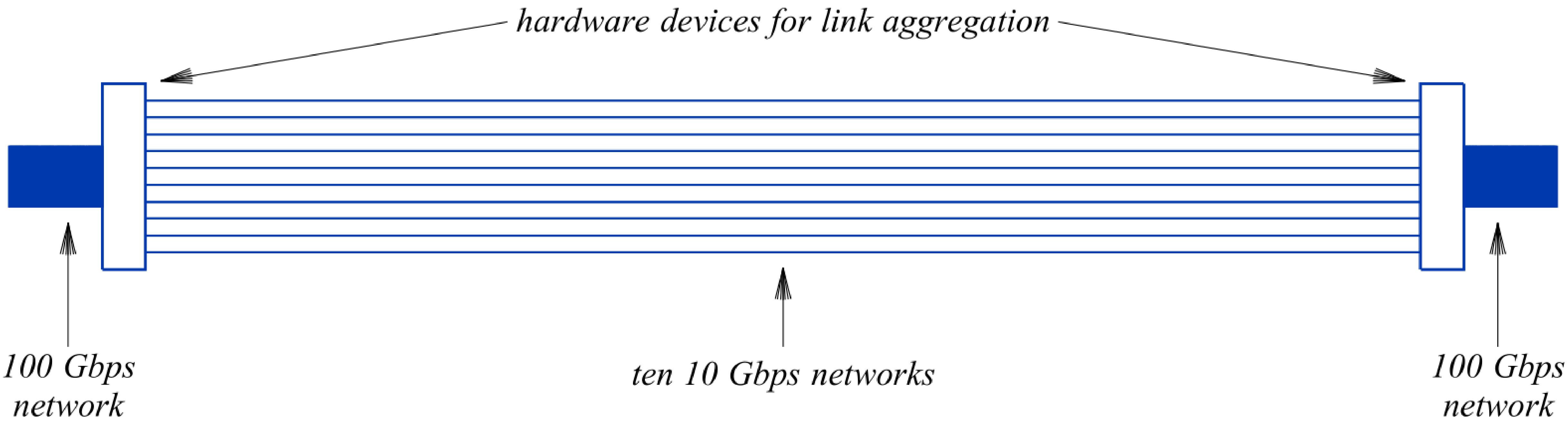
# High link capacities and link aggregation

- Link aggregation, or bonding, join multiple links into a single link



# High link capacities and link aggregation

- Link aggregation, or bonding, join multiple links into a single link
  - Requires additional hardware, cabling
  - Hardware may constrain number of links and link capacities that can be used

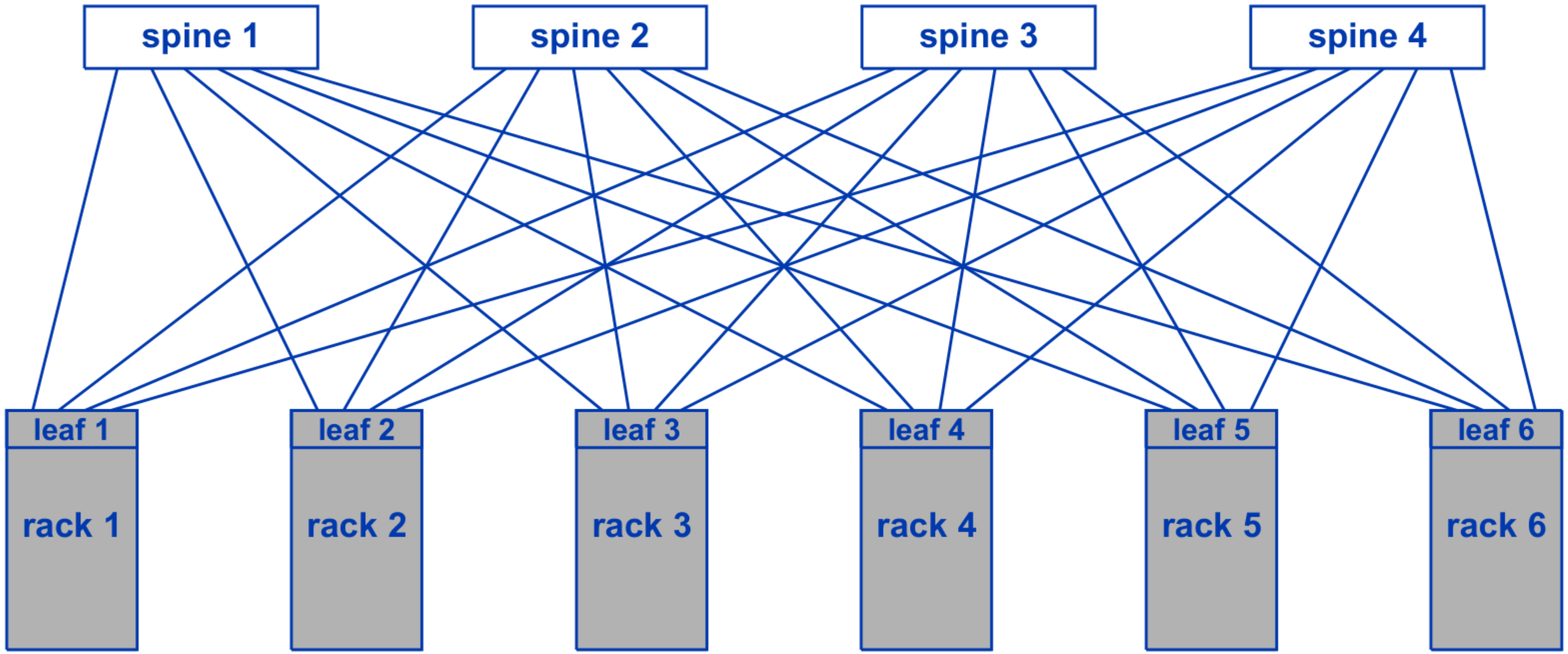


# Leaf-spine topology for east-west traffic

- How to support large volumes of east-west traffic without a hierarchical design?

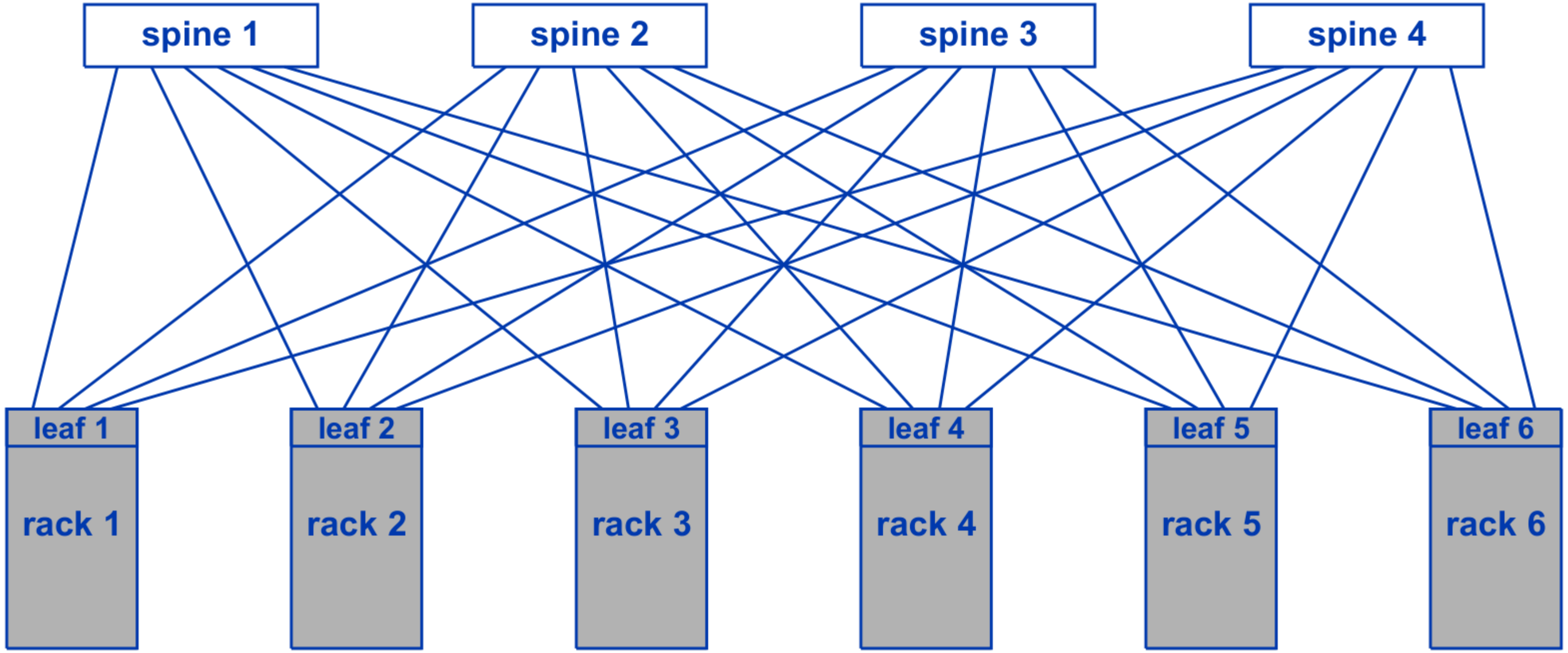
# Leaf-spine topology for east-west traffic

- How to support large volumes of east-west traffic without a hierarchical design?



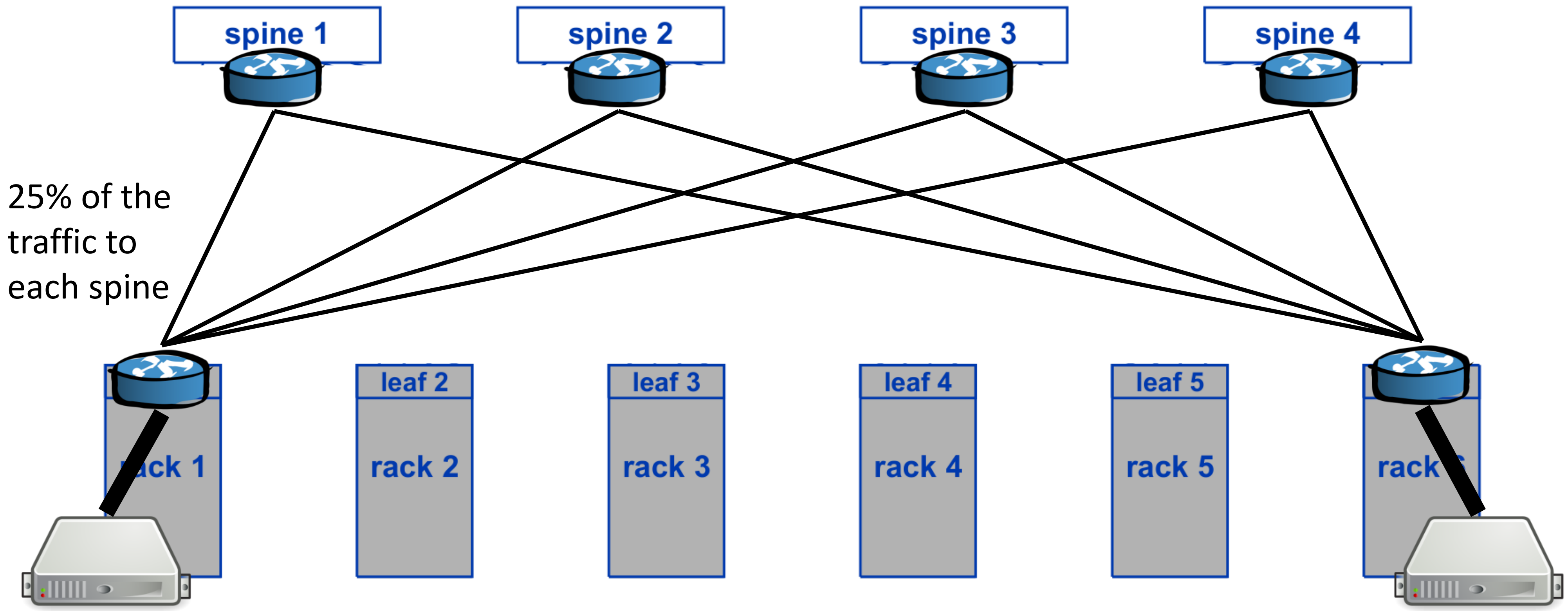
# Leaf-spine topology for east-west traffic

- High capacity for east-west traffic
  - Parallelism provides additional bandwidth



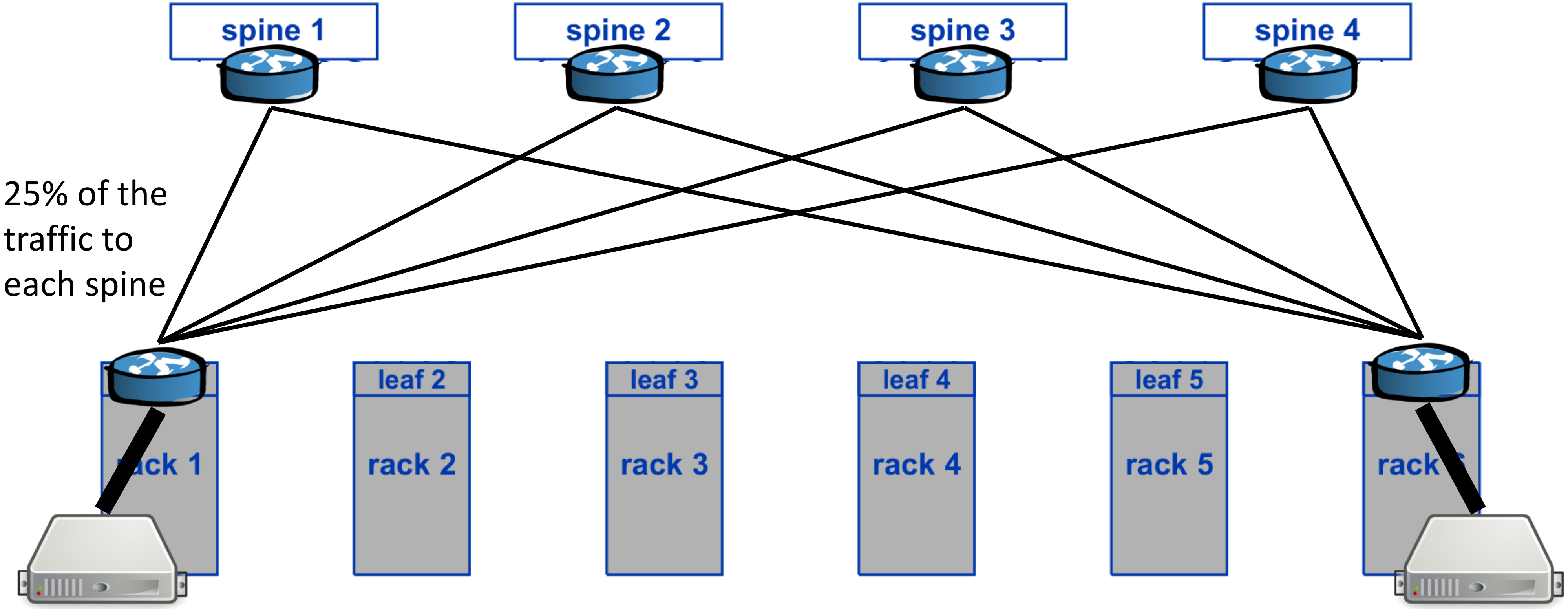
# Leaf-spine topology for east-west traffic

- High capacity for east-west traffic
  - Parallelism provides additional bandwidth
  - Switches employ Equal Cost Multipath (ECMP) to balance traffic



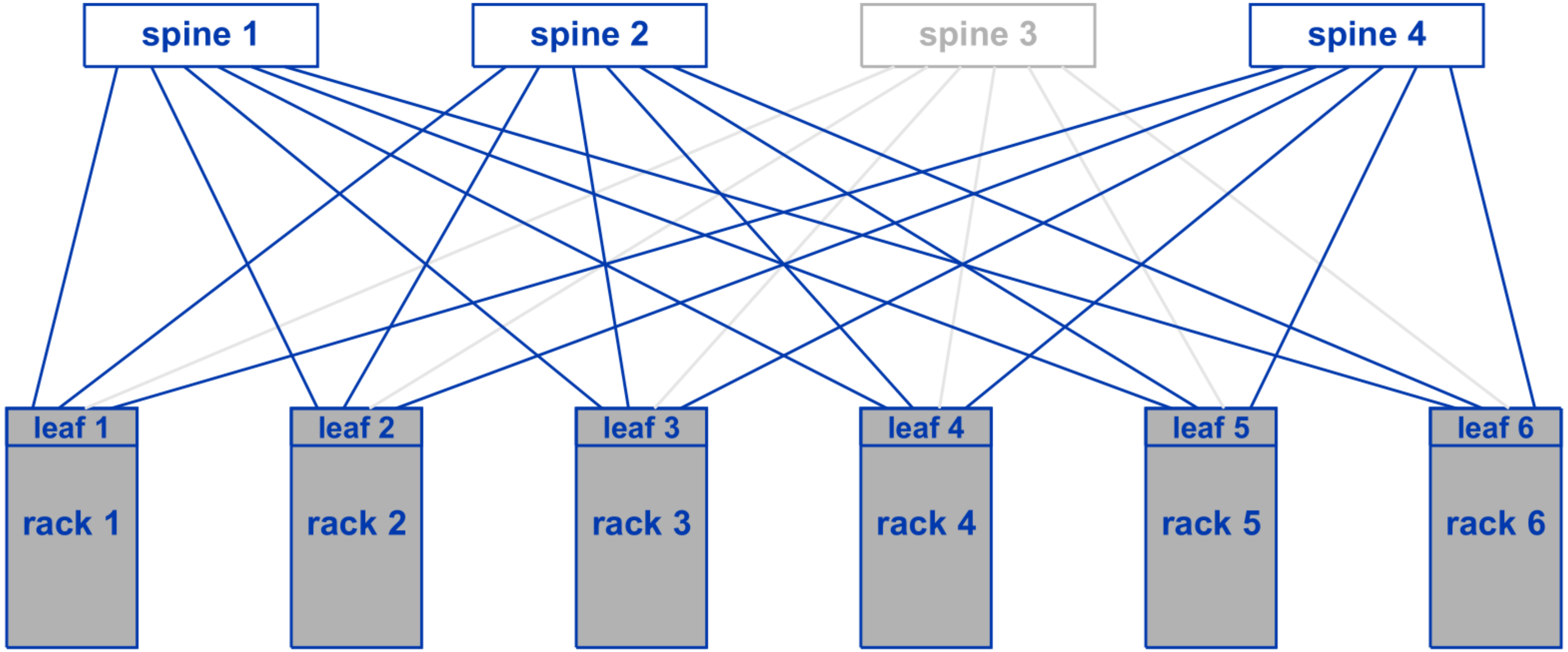
# Leaf-spine topology for east-west traffic

- High capacity for east-west traffic
  - Parallelism provides additional bandwidth
  - Switches employ Equal Cost Multipath (ECMP) to balance traffic



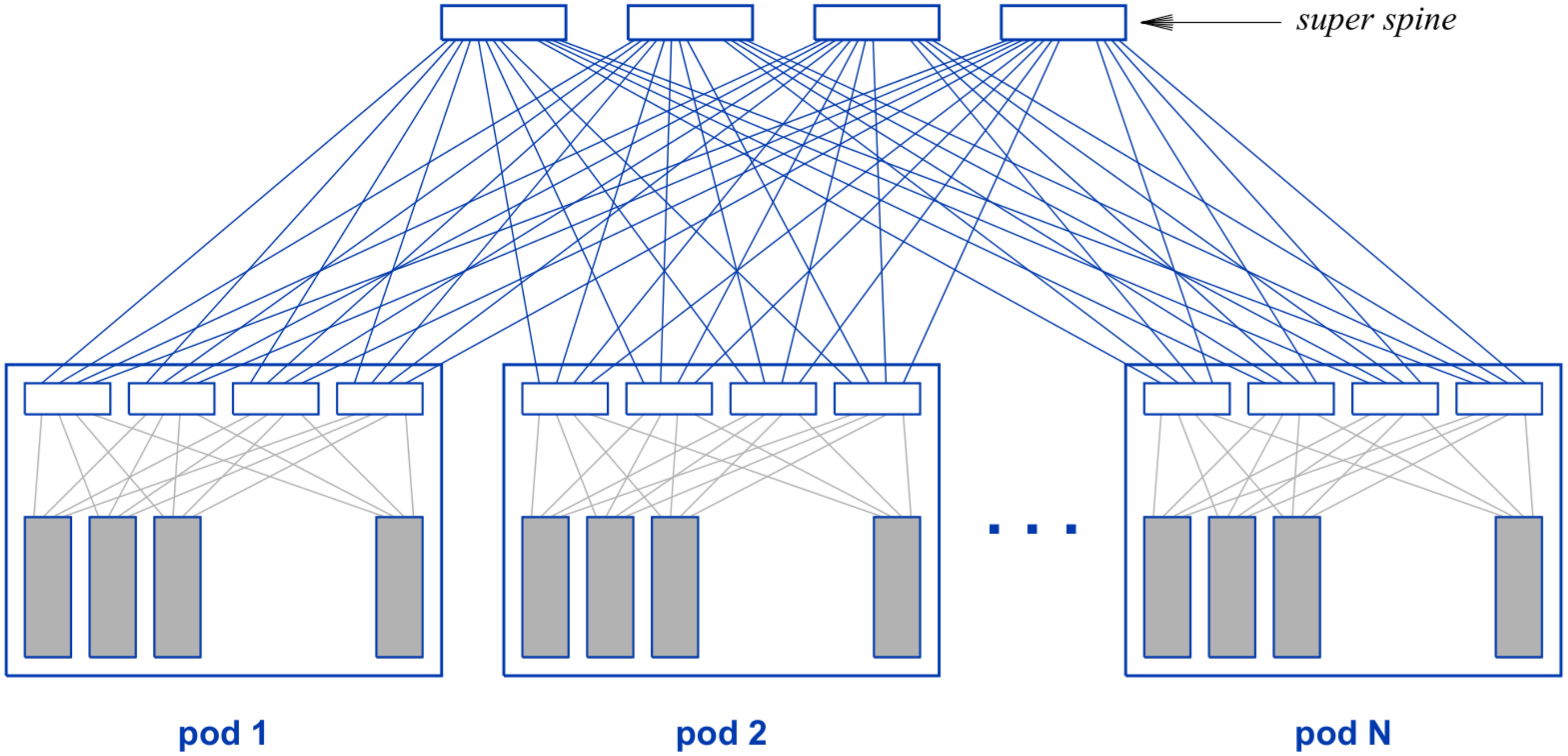
# Leaf-spine topology for east-west traffic

- Resiliency to failures
  - Bandwidth decreases, but alternate paths exist and communication continues
  - Routing and ECMP need to be updated to avoid the failed paths



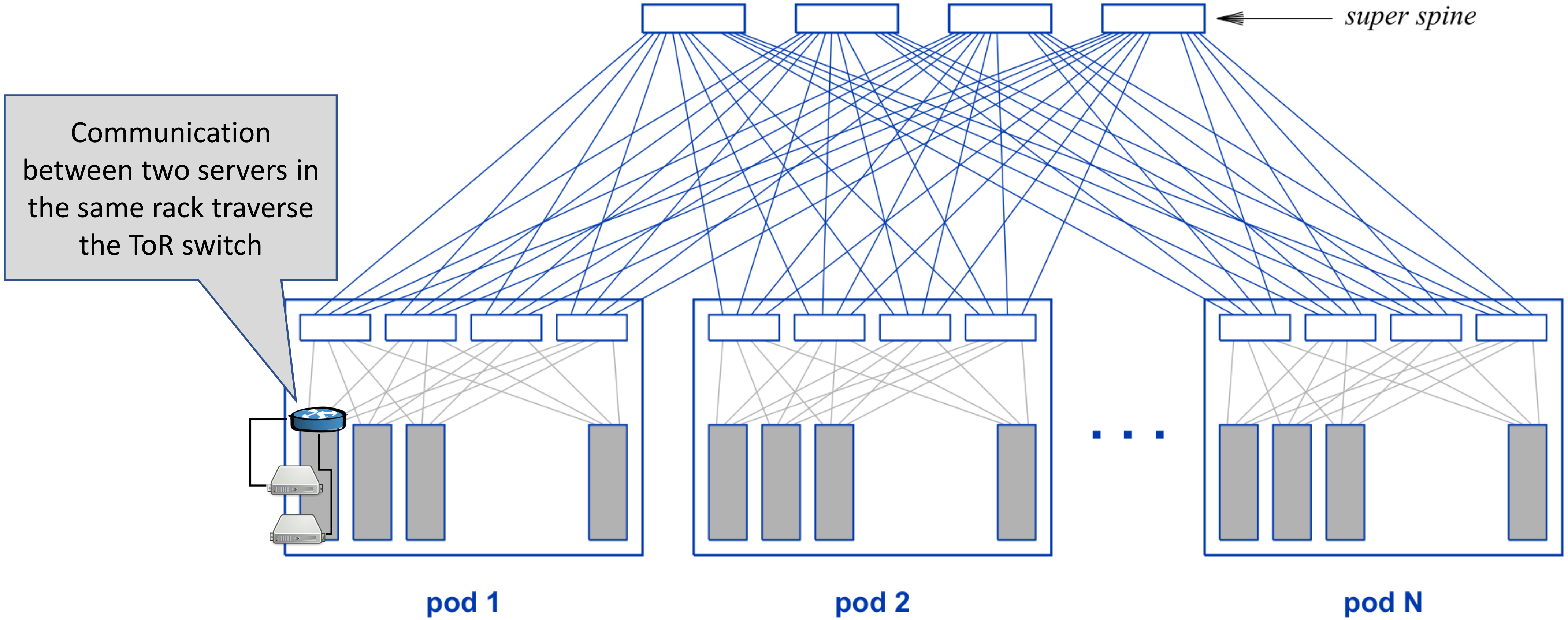
# Multi-layer topologies

- Switches have a limited number of ports
- Large datacenters need multiple layers of spines



# Multi-layer topologies

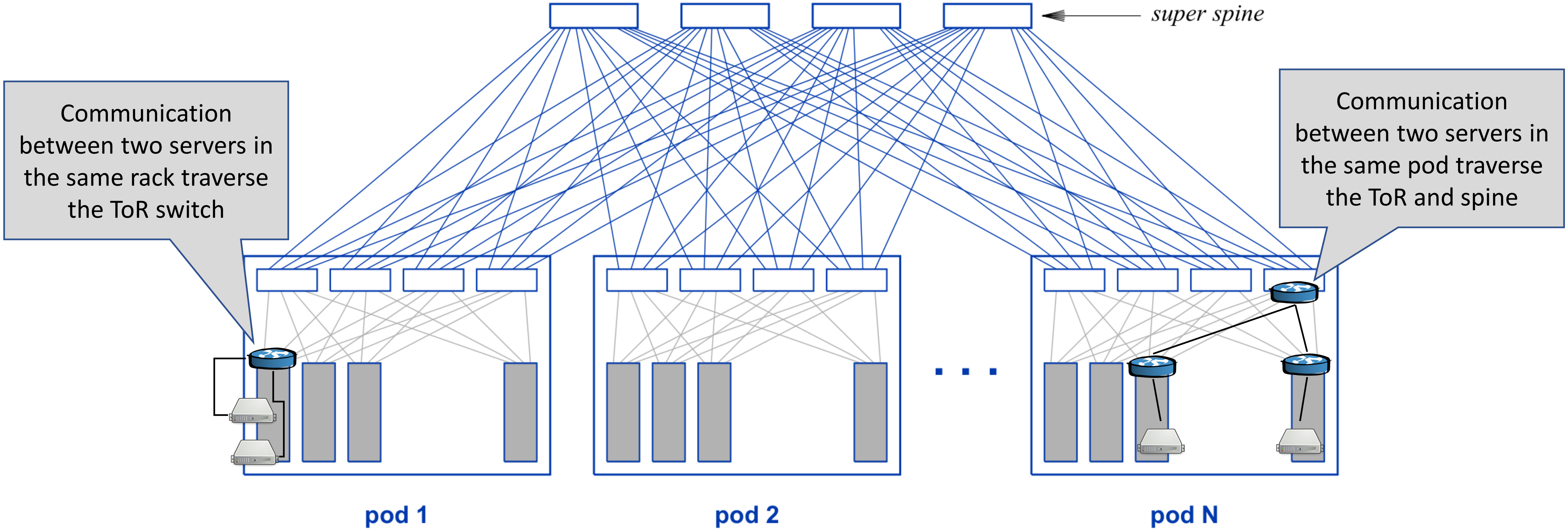
- Short paths for local communication



Communication between two servers in the same rack traverse the ToR switch

# Multi-layer topologies

- Short paths for local communication

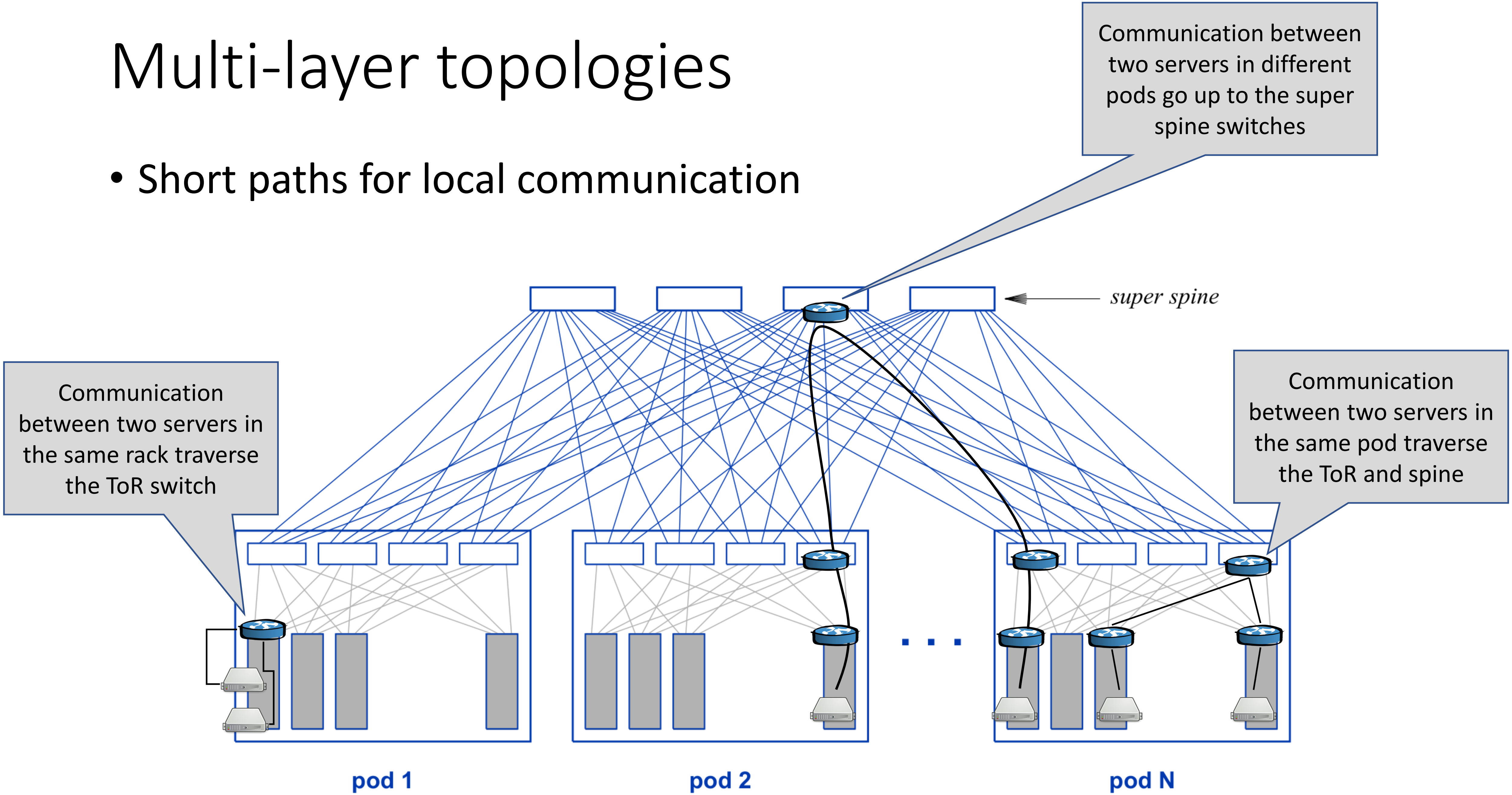


Communication between two servers in the same rack traverse the ToR switch

Communication between two servers in the same pod traverse the ToR and spine

# Multi-layer topologies

- Short paths for local communication



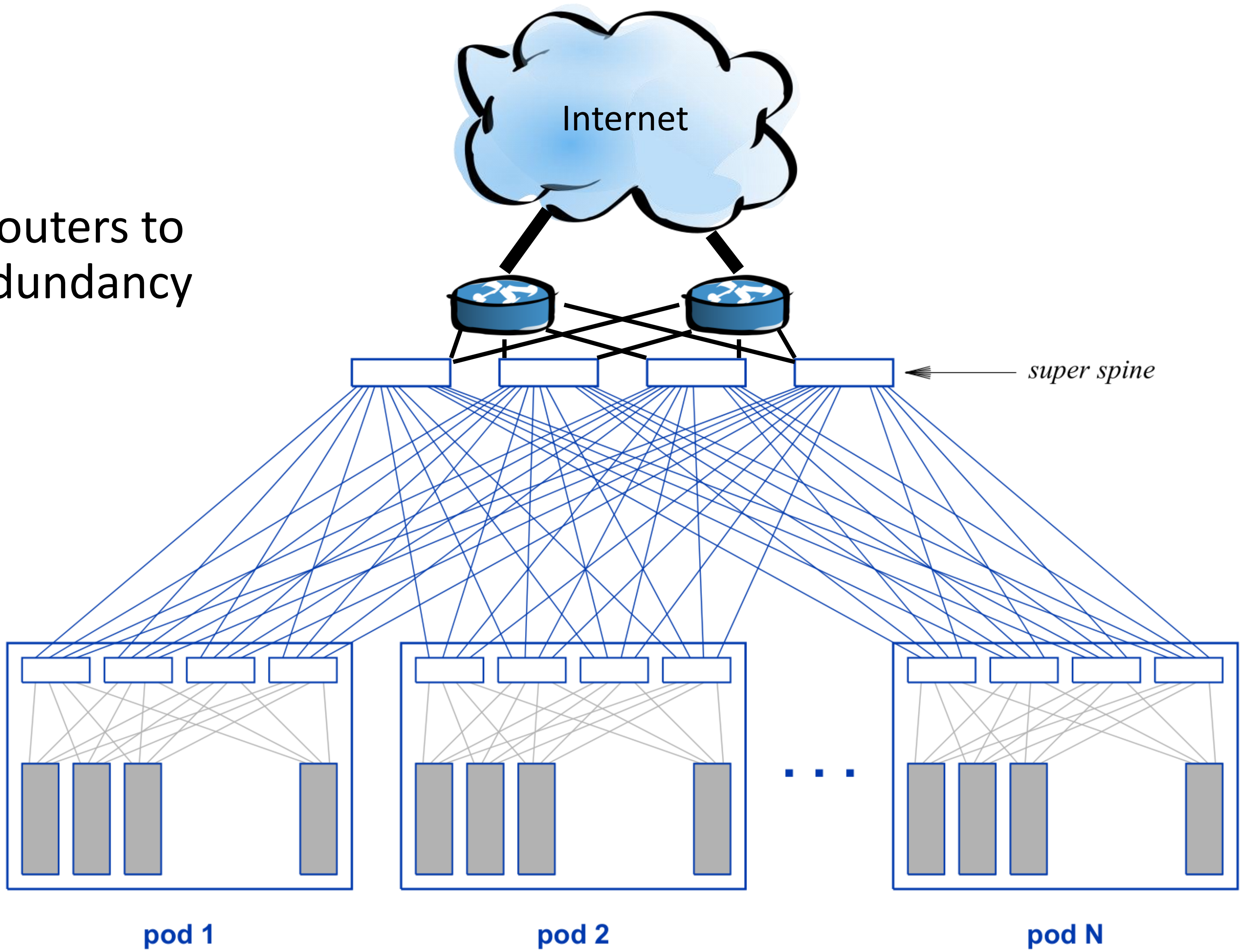
Communication between two servers in different pods go up to the super spine switches

Communication between two servers in the same rack traverse the ToR switch

Communication between two servers in the same pod traverse the ToR and spine

# Connecting to the Internet

- Multiple routers to ensure redundancy



CompSci 401: Cloud Computing

# Storage in Datacenters

Prof. Ítalo Cunha

昆山杜克大学
DUKE KUNSHAN UNIVERSITY

# Evolution of storage in datacenters

- Classical approach
  - Servers with their own spinning disks (HDDs)
  - Data replicated on multiple servers
  - External disks accessed through specialized hardware (SCSI, Fibre Channel)
- Modern approach
  - Servers with minimal amounts of storage
  - Data stored remotely
    - Data accesses go through the network, no specialized hardware
  - Solid state drives (SSDs)
  - Replication only when necessary

# Block storage

- Storage in datacenters is virtualized
- Each VM has a virtual disk, stored on over-the-network storage
- Each disk access goes over the network
  - Storage needs to be placed strategically (e.g., within the rack or pod)
  - Reduce network latency and bandwidth use