PADERBORN UNIVERSITY

Advanced Networked Systems SS24

Data Center Networks

Prof. Lin Wang, Ph.D.
Computer Networks Group
Paderborn University
https://en.cs.uni-paderborn.de/cn

Learning objectives

How to **interconnect millions of servers** in a data center?

How to achieve **flexible management** in a data center network?

Cloud computing

Elastic resources

- Expand and contract resources
- Pay-per-use, infrastructure on demand

Multi-tenancy

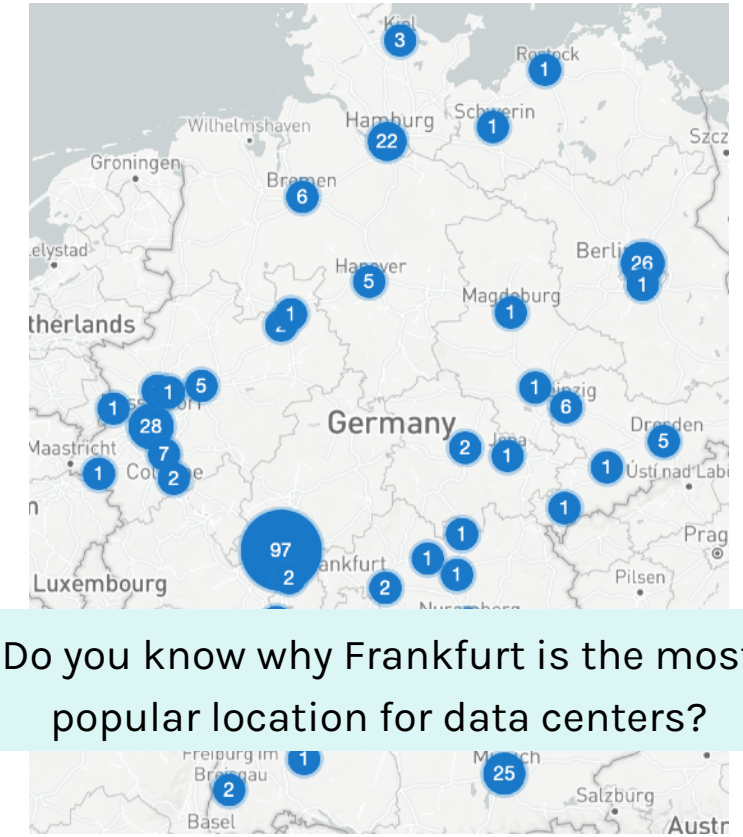
- Multiple independent users, resource isolation
- Amortize the cost of the shared infrastructure

Flexible service management

- Resilience: isolate failures of server and storage
- Workload migration: move work to other locations

What is behind cloud computing?

Large-scale data centers

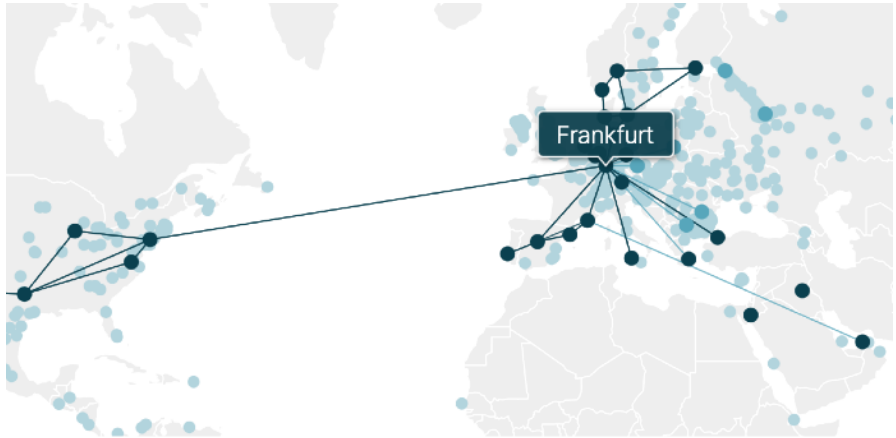


Do you know why Frankfurt is the most popular location for data centers?

https://www.datacentermap.com/germany/

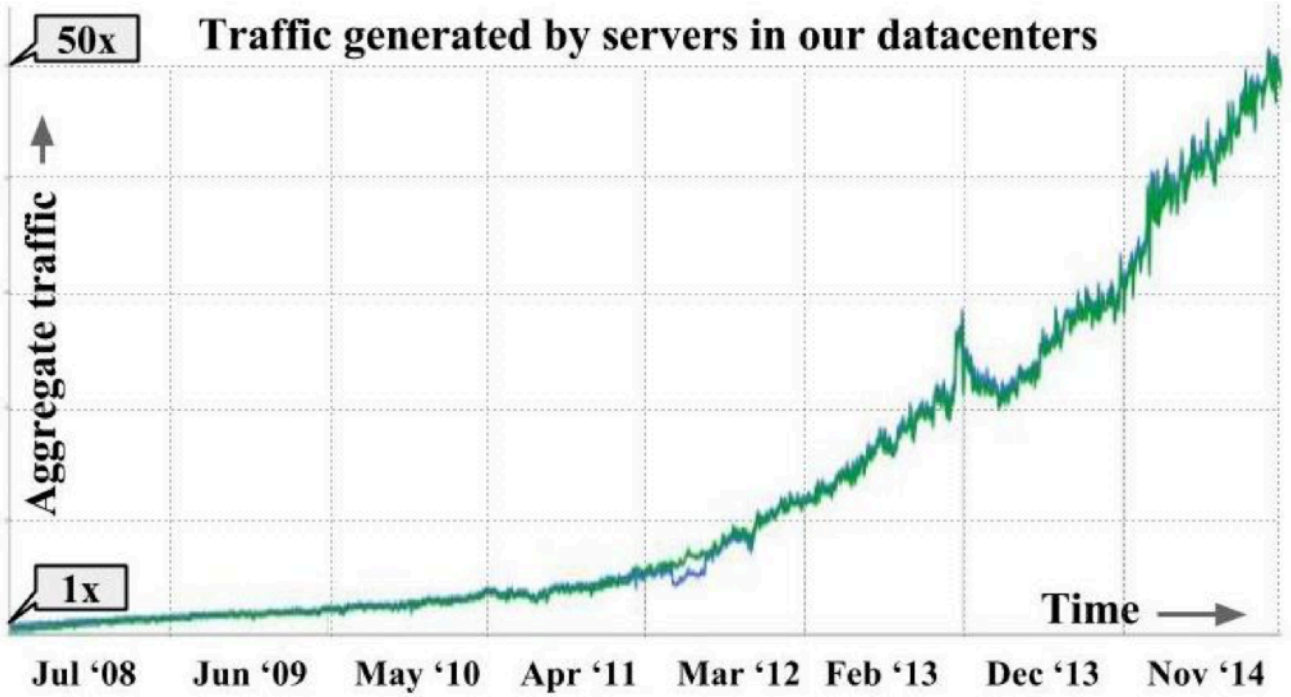
Internet exchange points

Facilitate efficient interconnection between ISP networks



https://www.datacentermap.com/ixp/g/frankfurt/

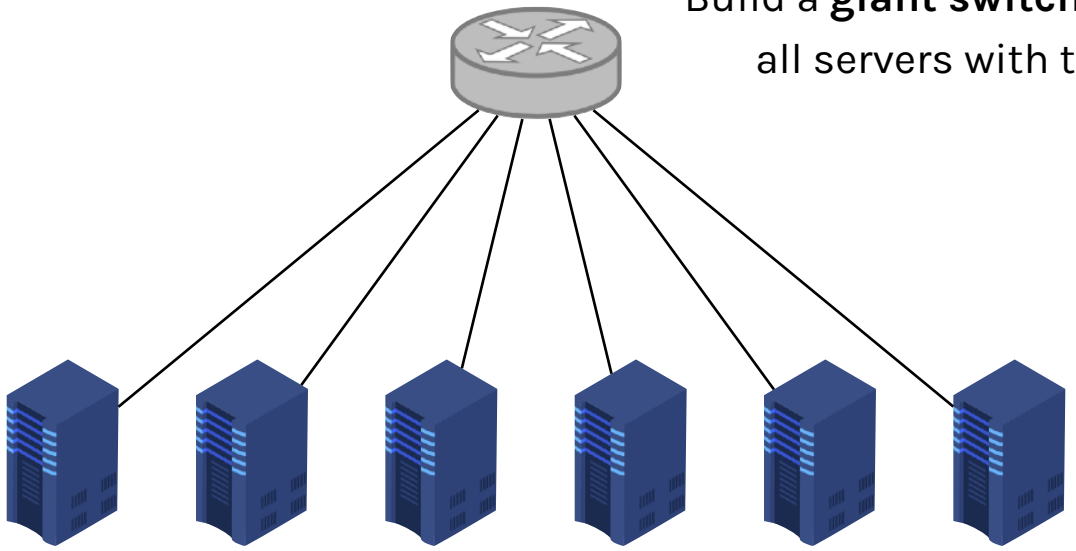
Aggregated server traffic in Google's data centers



How to interconnect the servers with sufficient bandwidth?

How to interconnect many servers?

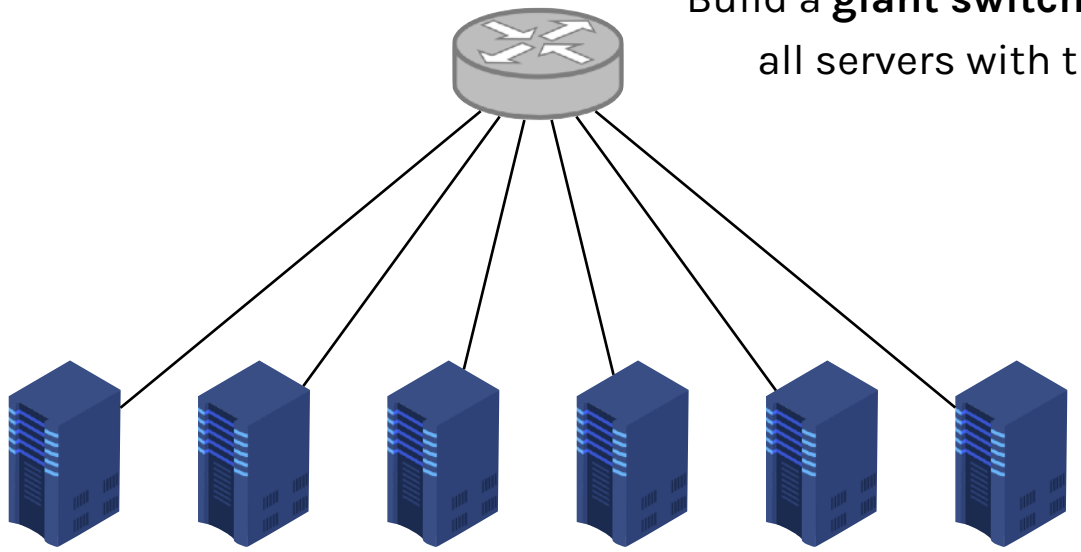
Build a **giant switch** and connect all servers with the switch



What problems can you think of with such a design?

How to interconnect many servers?

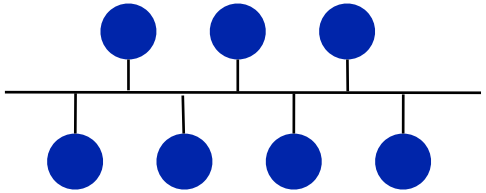
Build a **giant switch** and connect all servers with the switch



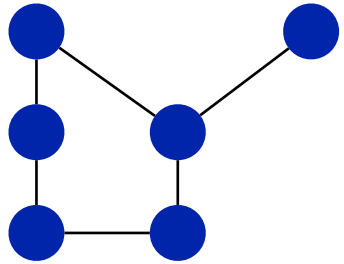
Limited port density, monetary cost, broadcast storms, isolation…

A dedicated network for the data center

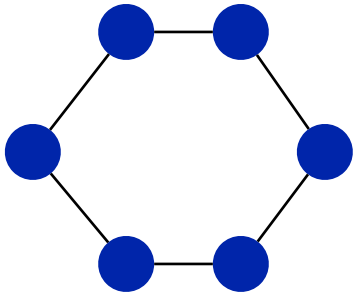
Line



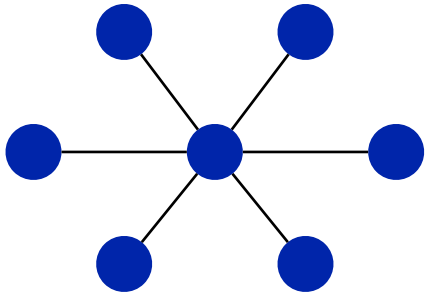
Bus



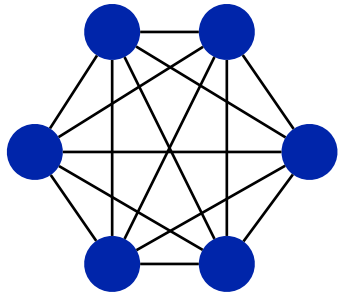
Mesh



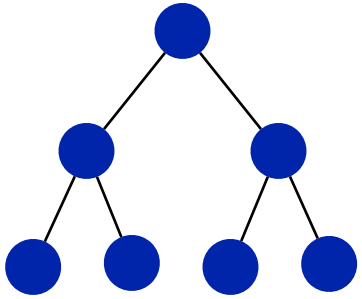
Ring



Star



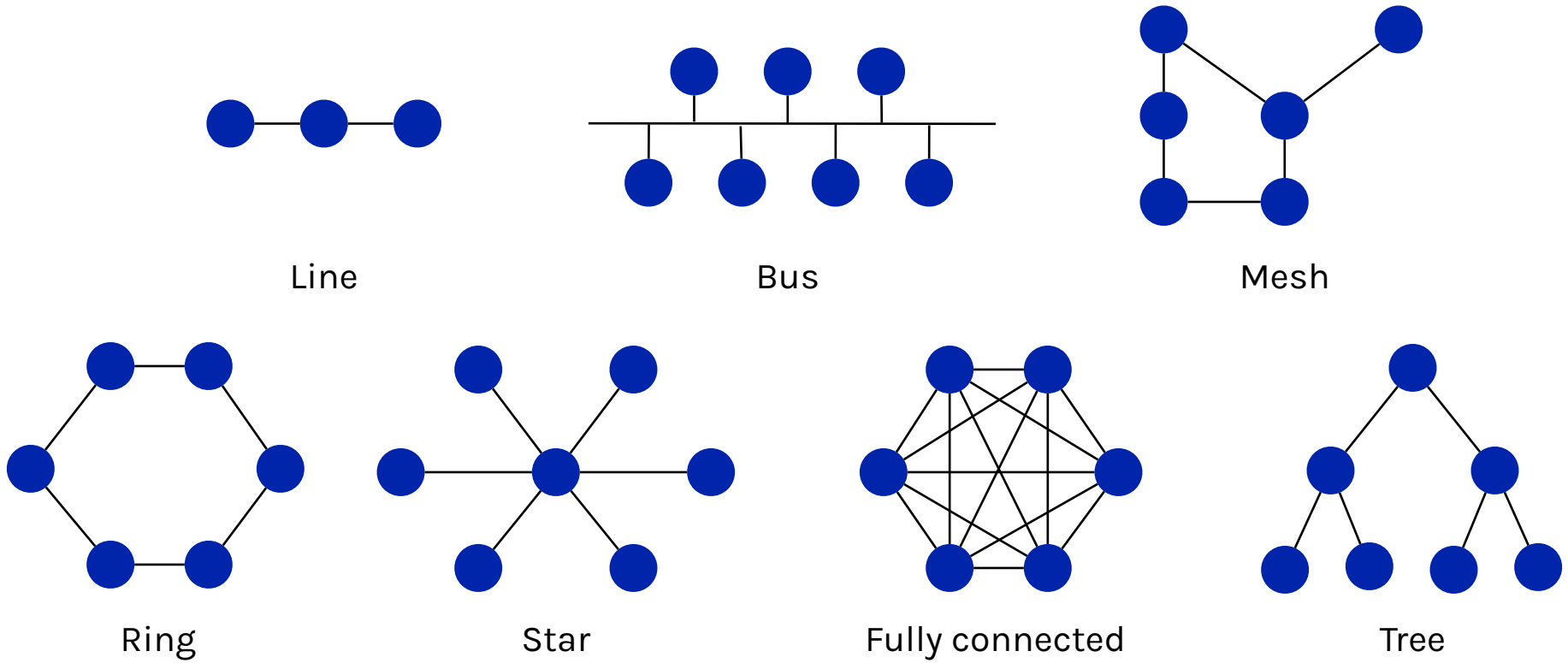
Fully connected



Tree

Which ones are more suitable for a data center?

A dedicated network for the data center



Tradeoff between connectivity and complexity

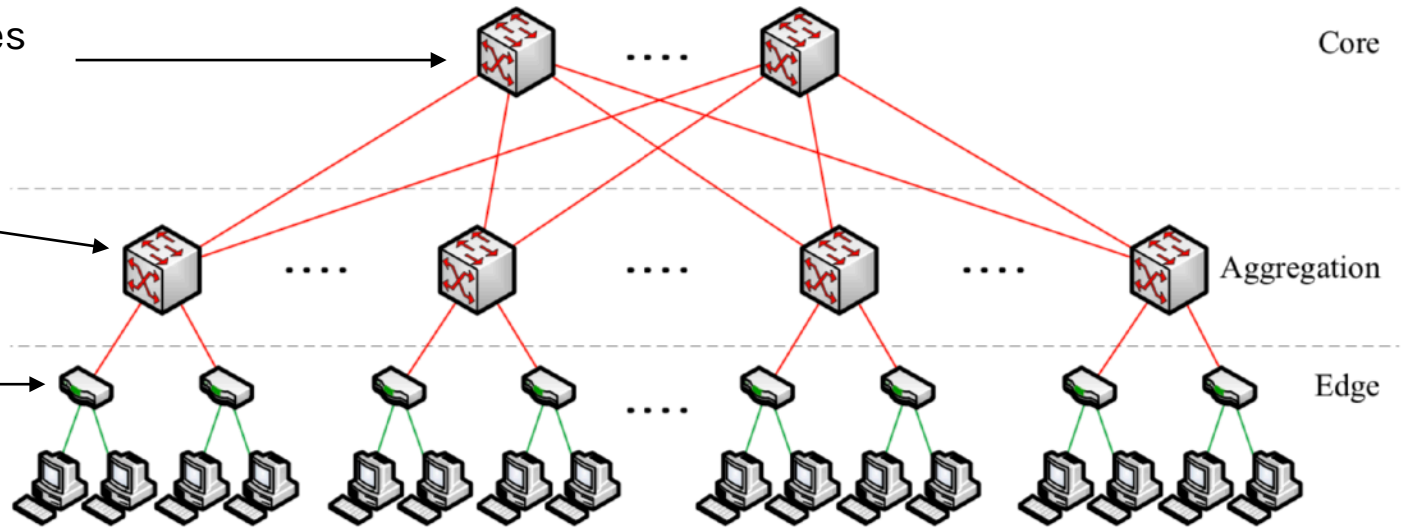
A tree-based data center network

A 3-tier tree architecture

Core switches (10Gbps)

Aggregation switches (10Gbps)

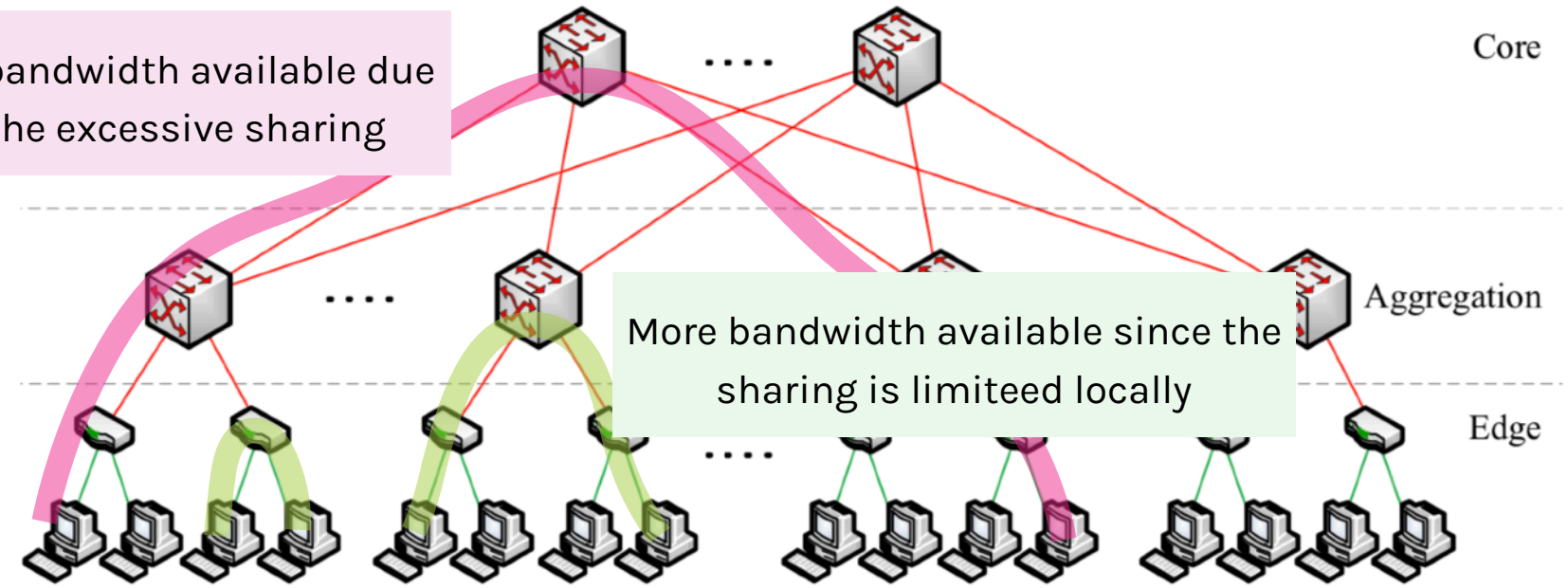
Top-of-rack (ToR) switches (1Gbps)



What if ToR switches go for 10 Gbps or beyond? Nowadays we are talking about 400 GbE links.

Bottleneck in tree-based networks

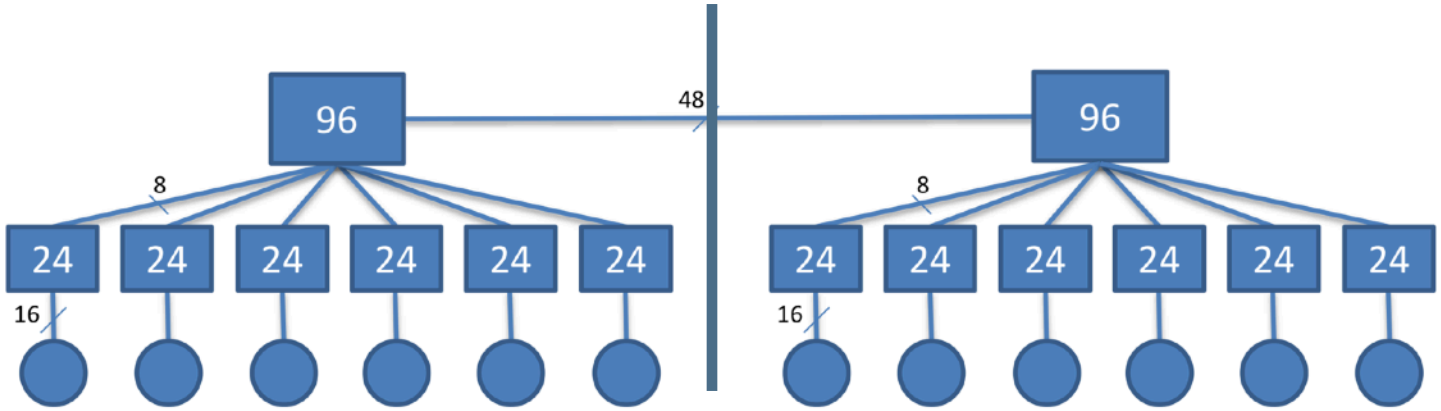
Less bandwidth available due to the excessive sharing



More bandwidth available since the sharing is limiteed locally

How to quantitatively measure the connectivity?

Network performance metrics



Bisection width	The minimum number of links cut to divide the network into two halves
Bisection bandwidth	The minimum bandwidth of the links that divide the network into two halves
Full bisection bandwidth	Nodes in one half can communicate simultaneously with nodes in the other half, at their full uplink capacity

Oversubscription ratio

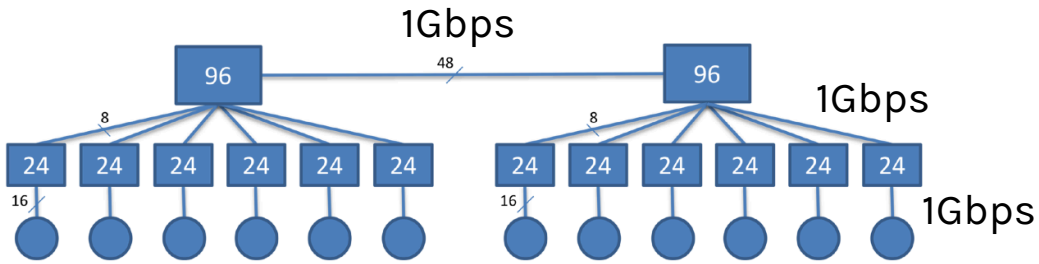
Definition

- Ratio of worst-case required aggregate bandwidth to the total uplink bandwidth of a network device
- Ability of hosts to fully utilize its uplink capabilitie

Examples

- 1:1 → All hosts can use full uplink capacity
- 5:1 → Only 20% of host bandwidth may be available

Typical data center oversubscription ratio is 2.5:1 to 8:1



What is the oversubscription ratio in the above topology?

Oversubscription ratio

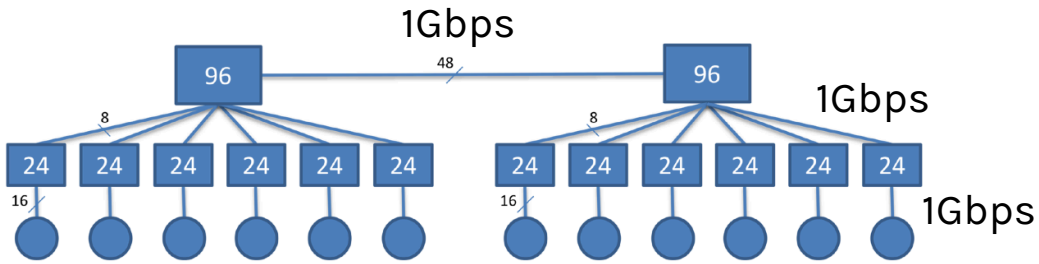
Definition

- Ratio of worst-case required aggregate bandwidth to the total uplink bandwidth of a network device
- Ability of hosts to fully utilize its uplink capabilitie

Examples

- 1:1 → All hosts can use full uplink capacity
- 5:1 → Only 20% of host bandwidth may be available

Typical data center oversubscription ratio is 2.5:1 to 8:1



Oversubscription ratio at the aggregation layer: 16 x 6 / 48 = 2:1

Oversubscription ratio at the core layer: 8 x 6 / 48 = 1:1

Factors behind data center network designs

Commoditization in the data center

- Inexpensive, commodity servers and storage devices
- Highly specialized network with proprietary devices

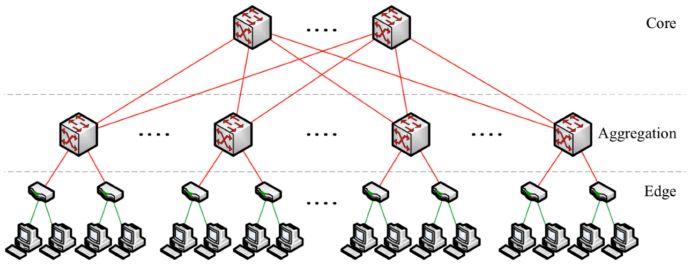
Data center is not a “small Internet”

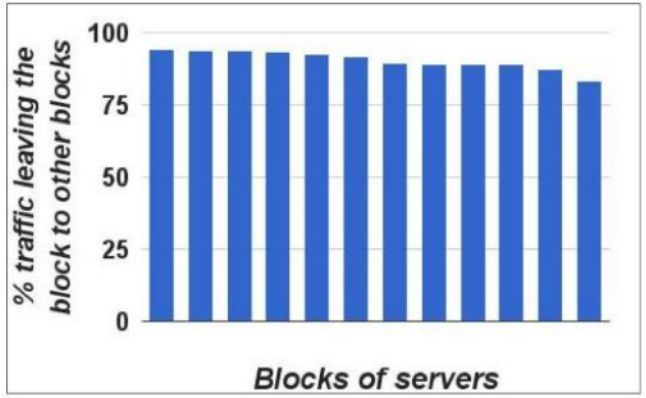
- One admin domain, not adversarial, limited policy routing, etc…

Bandwidth is often the bottleneck

- Data-intensive workloads (big data, graph processing, machine learning)
- Low traffic locality

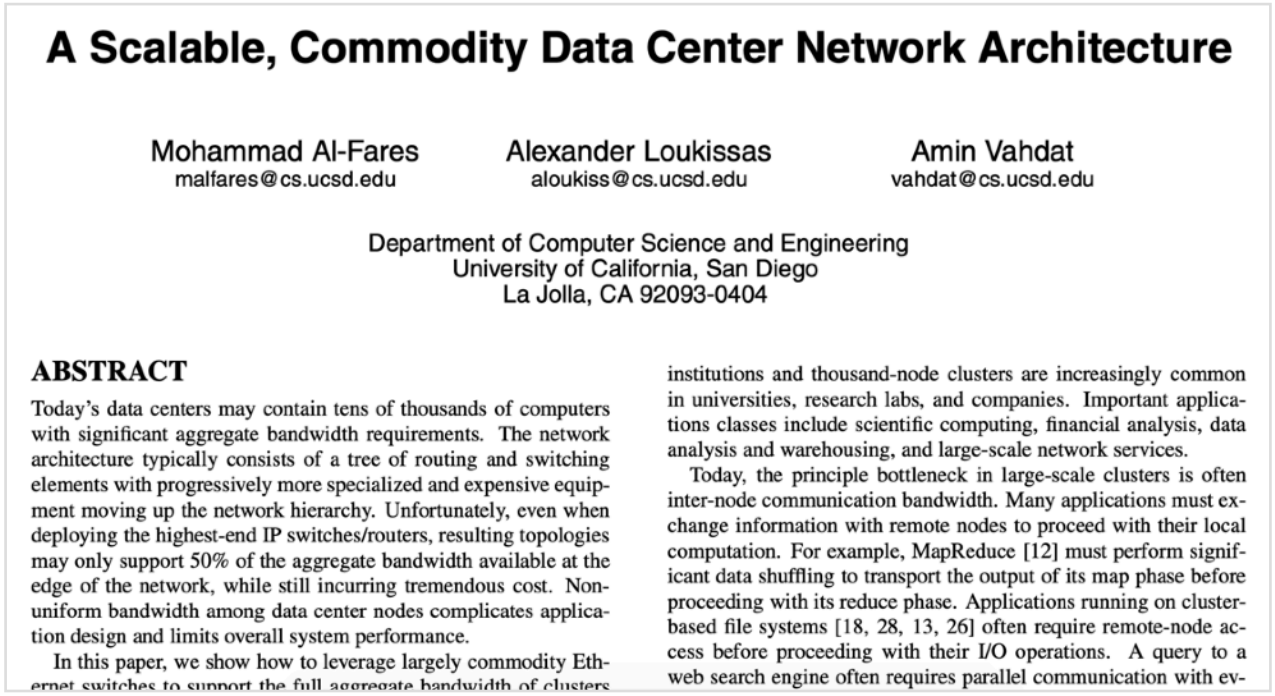
Large-fanout proprietary switch





Low traffic locality

Fat-tree



A Scalable, Commodity Data Center Network Architecture

Mohammad Al-Fares
malfares@cs.ucsd.edu

Alexander Loukissas
aloukiss@cs.ucsd.edu

Amin Vahdat
vahdat@cs.ucsd.edu

Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093-0404

ABSTRACT

Today's data centers may contain tens of thousands of computers with significant aggregate bandwidth requirements. The network architecture typically consists of a tree of routing and switching elements with progressively more specialized and expensive equipment moving up the network hierarchy. Unfortunately, even when deploying the highest-end IP switches/routers, resulting topologies may only support 50% of the aggregate bandwidth available at the edge of the network, while still incurring tremendous cost. Non-uniform bandwidth among data center nodes complicates application design and limits overall system performance.

In this paper, we show how to leverage largely commodity Ethernet switches to support the full aggregate bandwidth of clusters

institutions and thousand-node clusters are increasingly common in universities, research labs, and companies. Important applications classes include scientific computing, financial analysis, data analysis and warehousing, and large-scale network services.

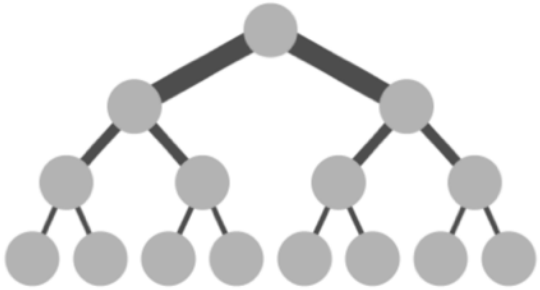
Today, the principle bottleneck in large-scale clusters is often inter-node communication bandwidth. Many applications must exchange information with remote nodes to proceed with their local computation. For example, MapReduce [12] must perform significant data shuffling to transport the output of its map phase before proceeding with its reduce phase. Applications running on cluster-based file systems [18, 28, 13, 26] often require remote-node access before proceeding with their I/O operations. A query to a web search engine often requires parallel communication with ev-

A special instance of a Clos network, instead of the traditional fat-tree;
but generally referred to as fat-tree by researchers

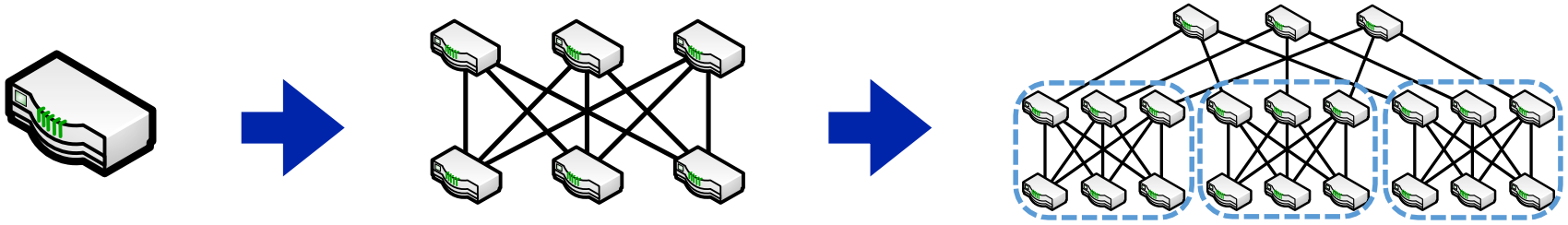
Fat-tree topology

A special instance of the Clos topology

- Clos networks are originally designed for telephone switches
- Emulate a single huge switch with many smaller switches
- Invented in 1938 by Edson Erwin and formalized by Charles Clos in 1953
- Fat-tree was proposed by Charles Leiserson in 1985, which means a different topology (shown in the right side)



An original fat-tree, not to be confused with the data center fat-tree topology



Fat-tree: design goals

Scalable interconnection bandwidth

- Full bisection bandwidth between all pairs of hosts (oversubscription ratio?)

Economies-of-scale

- Price/port is constant with the number of hosts, leverage commodity merchant silicon

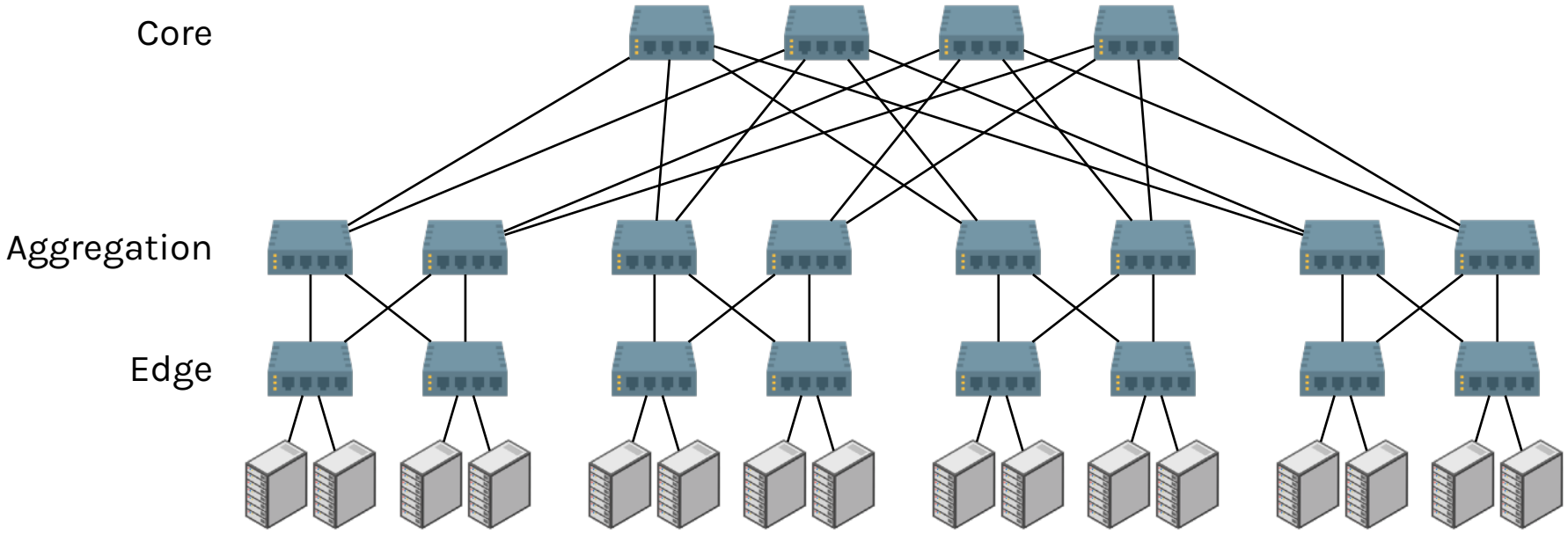
Compatibility

- Support Ethernet and IP without host modifications

Easy management

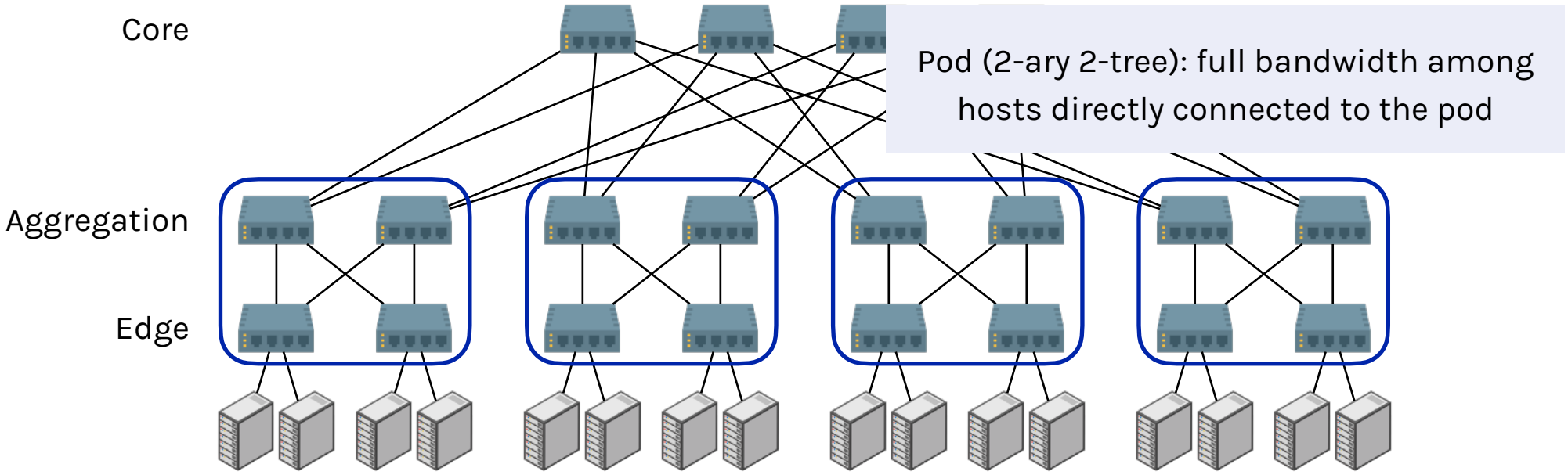
- Modular design, avoid manual management

Fat-tree example



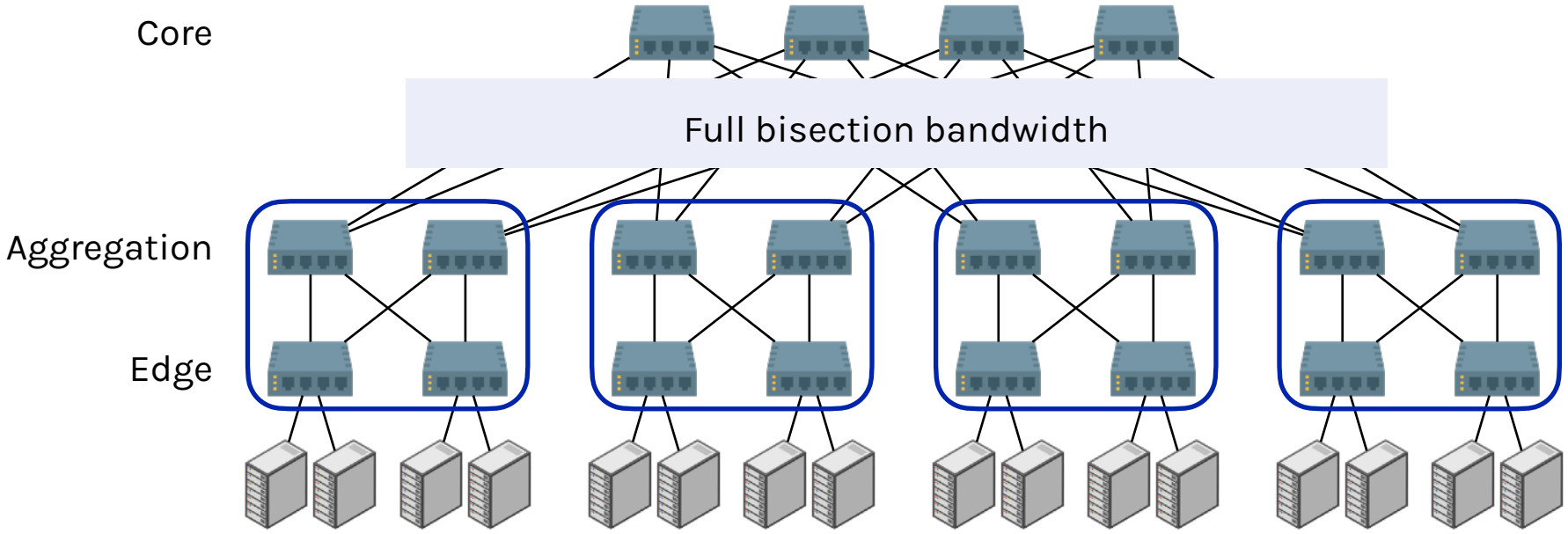
A fat-tree network built from 4-port **identical** switches

Fat-tree pod



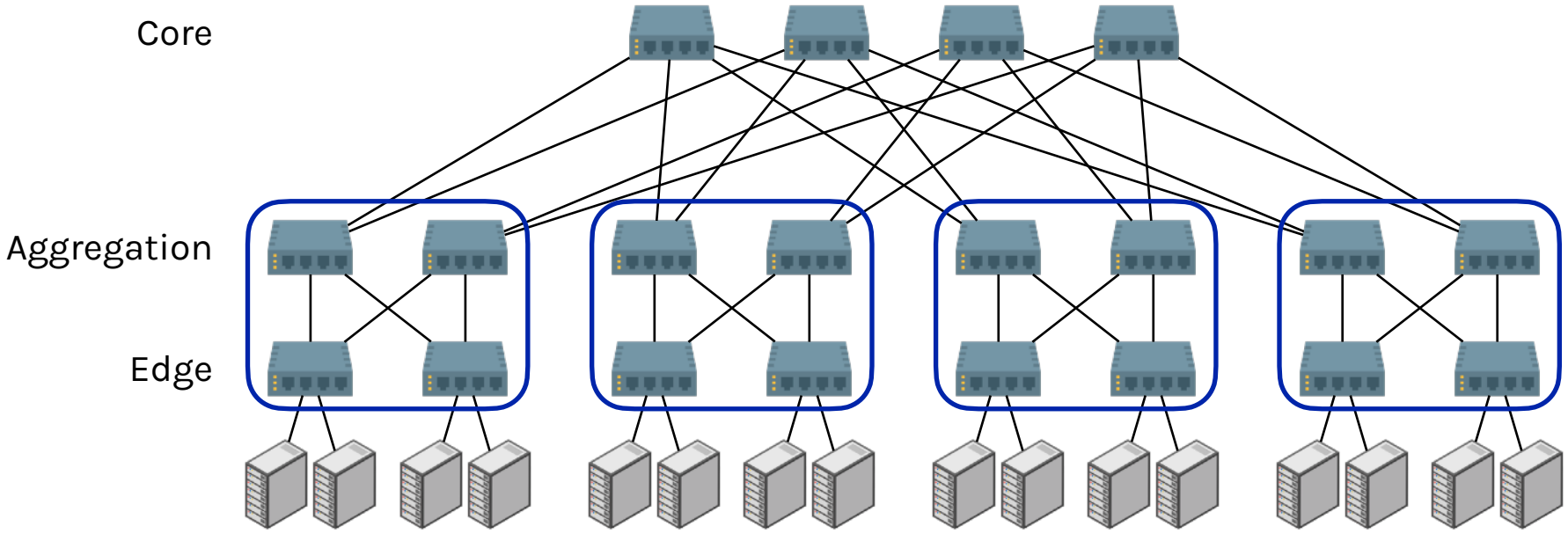
A fat-tree network built from 4-port **identical** switches

Fat-tree bisection bandwidth



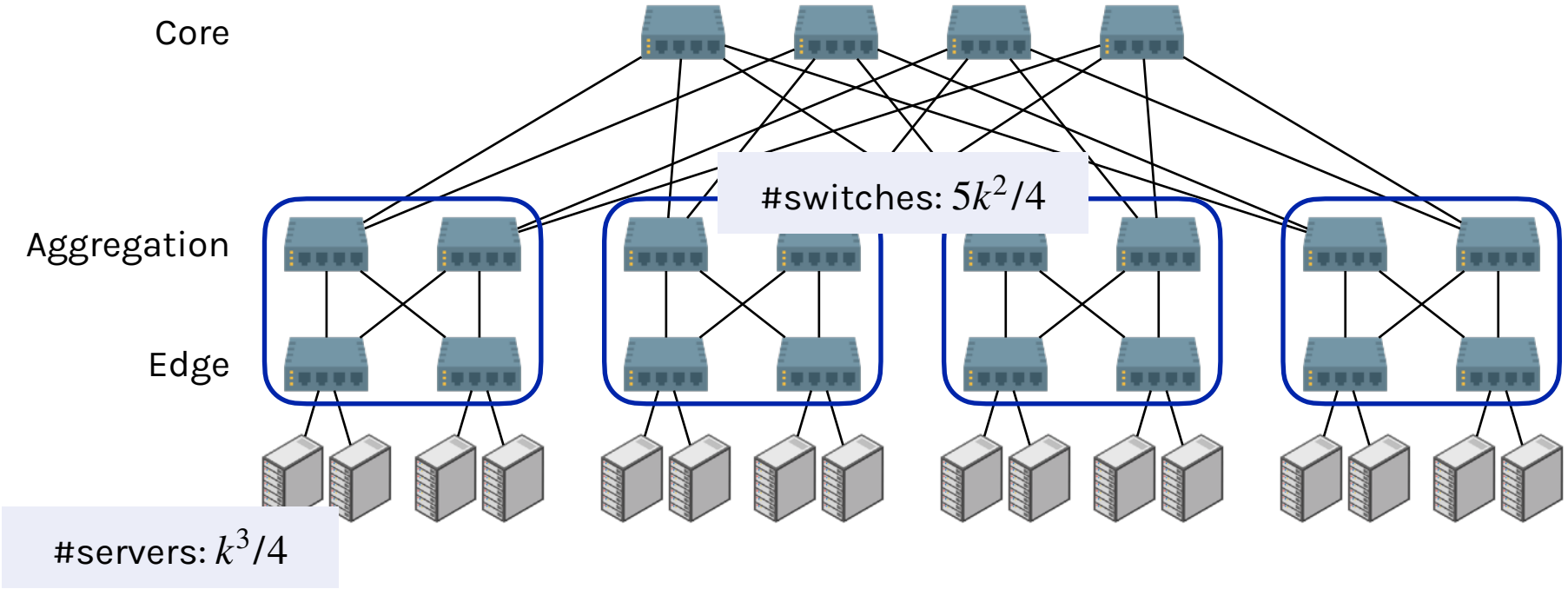
A fat-tree network built from 4-port **identical** switches

Fat-tree scalability



Suppose we use ***k*-port switches**, **how many servers** can we interconnect with fat-tree, and **how many switches** are needed?

Fat-tree scalability



Fat-tree can scale to any link capacity at the edge: 40 Gbps, 100 Gbps, 400 Gbps …

Why this has not been done before?

Needs to be backward compatible with IP/Ethernet

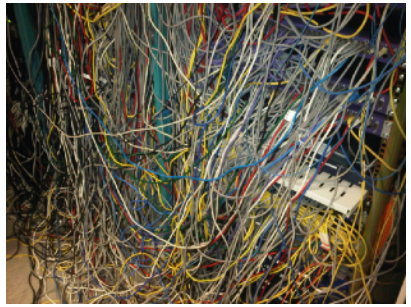
- Existing routing and forwarding protocols do not work for fat-tree
- Scalability challenges with millions of end points

Management

- Thousands of individual elements that must be programmed individually

Cabling explosion at each level of fat-tree

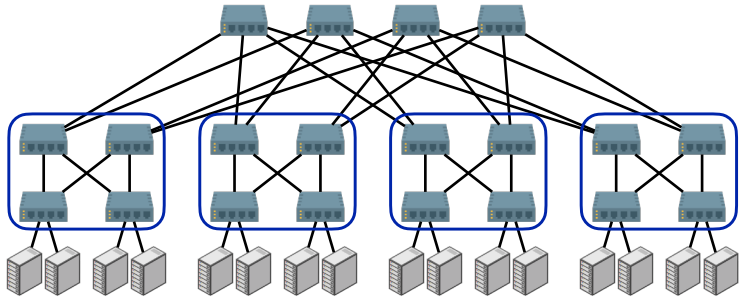
- Tens of thousands of cables running across the data center



Challenges with fat-tree

Backward compatible with IP/Ethernet

- Routing algorithms (such as OSPF) will naively choose a single shortest path to use between subnets
- Leads to bottleneck quickly
- $(k/2)^2$ shortest paths available, should use them all equally

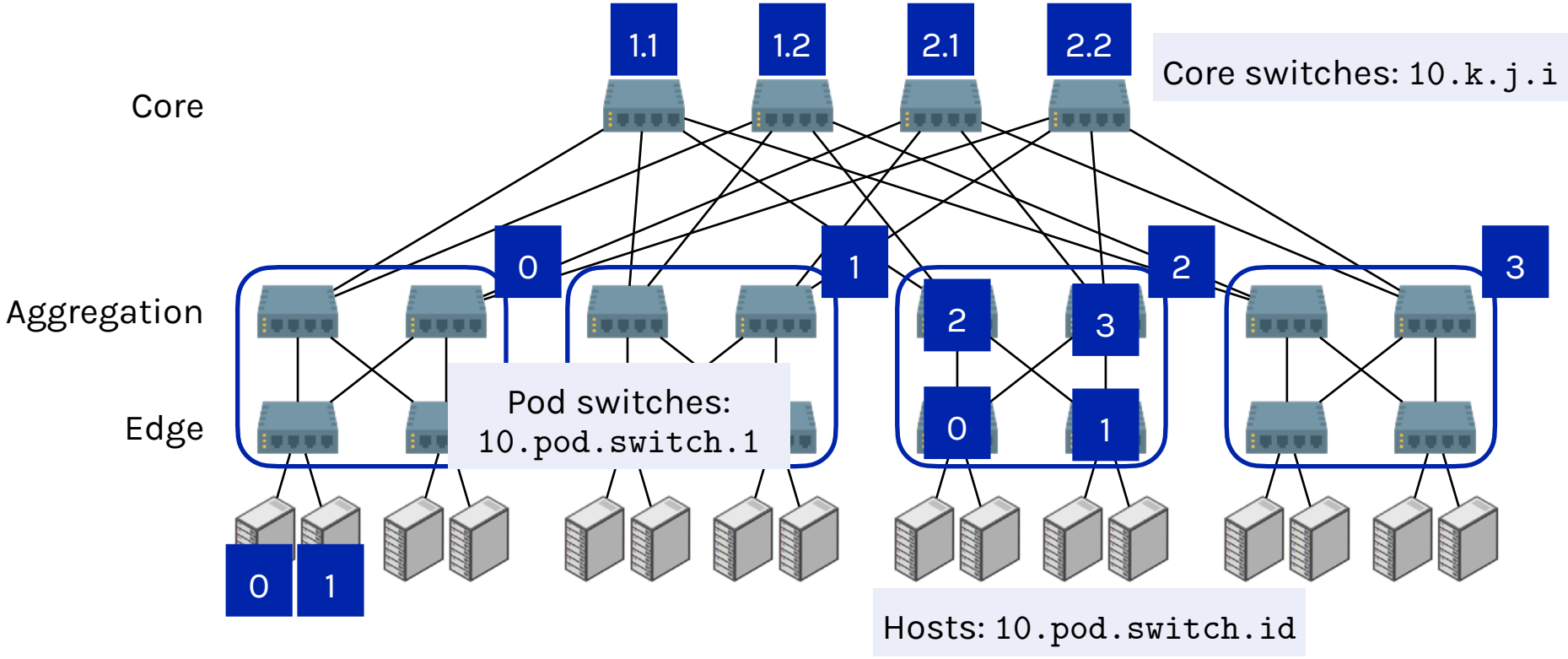


Complex wiring due to lack of high-speed ports

Hints: take advantage of the **regularity of the fat-tree structure** to simplify protocol design and improve performance

Addressing in fat-tree

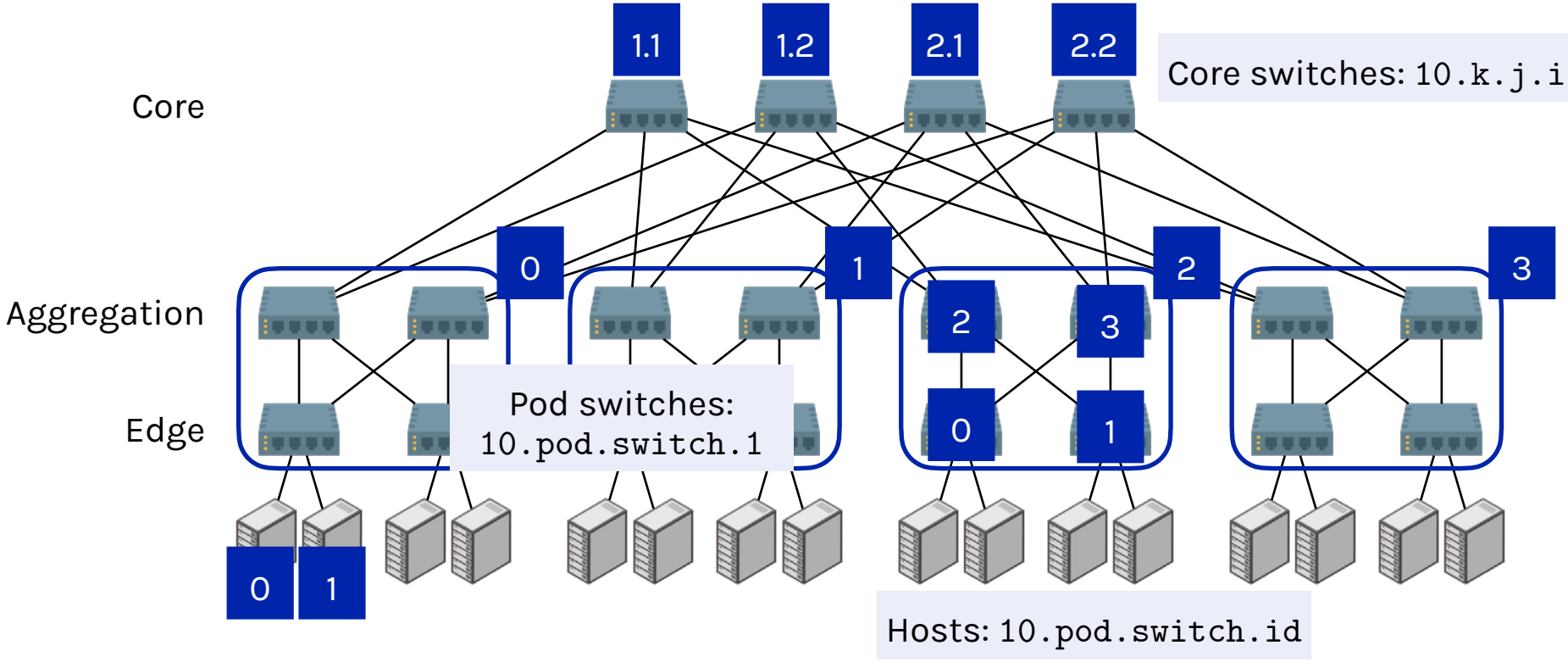
Use 10.0.0.0/8 private address block



What factors could limit the max. size of a fat-tree network?

Addressing in fat-tree

Use 10.0.0.0/8 private address block



$k < 256$

Forwarding on fat-tree

Two-level lookup table

- Prefixes used to forward intra-pod traffic
- Suffixes used to forward inter-pod traffic

Host IP: `10.pod.switch.id`

Prefix	Output port
10.2.0.0/24	0
10.2.1.0/24	1
0.0.0.0/0	

→

Suffix	Output port
0.0.0.2/8	2
0.0.0.3/8	3

Hosts in different pods are forwarded based on the host ID

TCAM

10.2.0.X
10.2.1.X
X.X.X.2
X.X.X.3

→ Encoder →

RAM

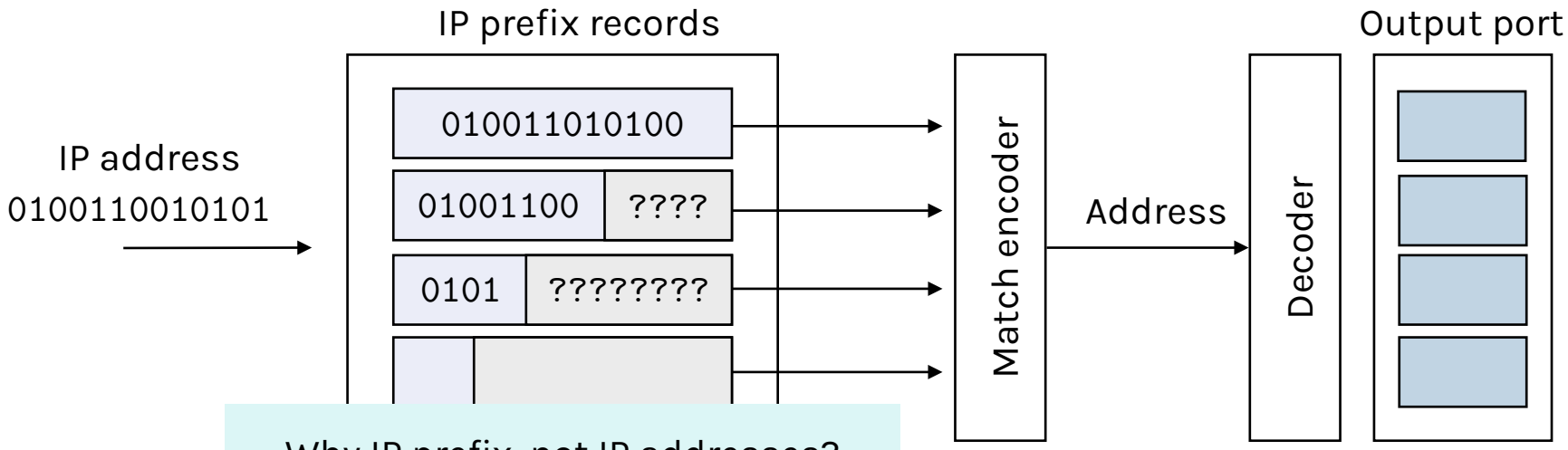
Address	Next hop	Output port
00	10.2.0.1	0
01	10.2.1.1	1
10	10.4.1.1	2
11	10.4.1.2	3

TCAM-based implementation

Ternary content addressable memory (TCAM)

Supports to match on a set of records in constant time (one iteration)

- CAM supports only two states (0/1) in each bit position: widely used in switches for MAC address matching
- TCAM extends CAM by allowing for 3 states (0/1/?) in each position: useful for IP prefix matching
- Disadvantages: expensive, power-consuming

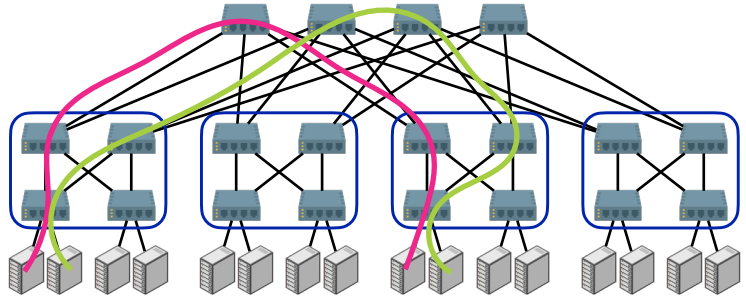


Routing on fat-tree

Prefixes in two-level lookup table prevent intra-pod traffic from leaving the pod

Inter-pod traffic is handled by suffix table

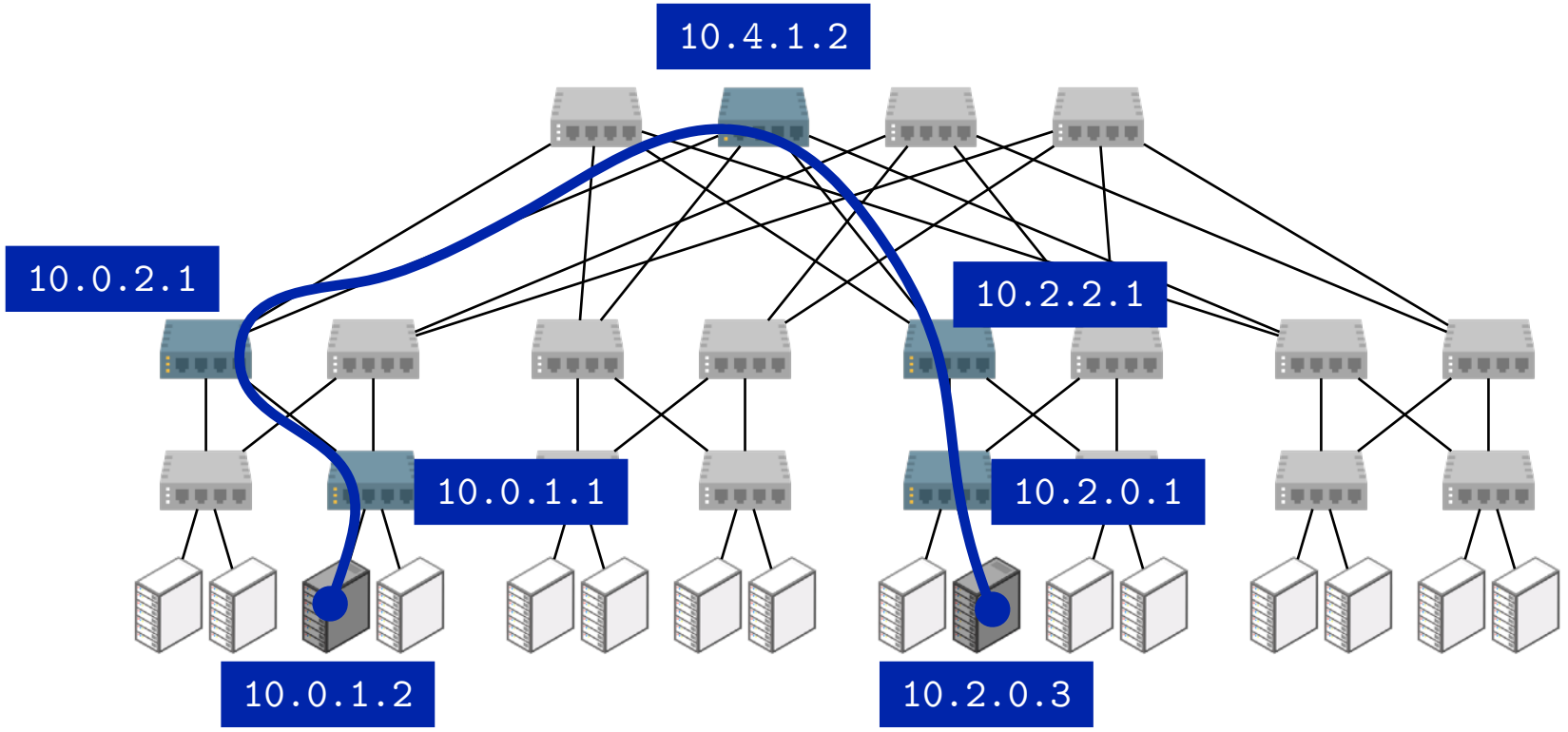
- **Suffixes** based on host IDs, ensuring spread of traffic across core switches
- Prevent packet reordering by having static path



Each host-to-host communication has a single static path

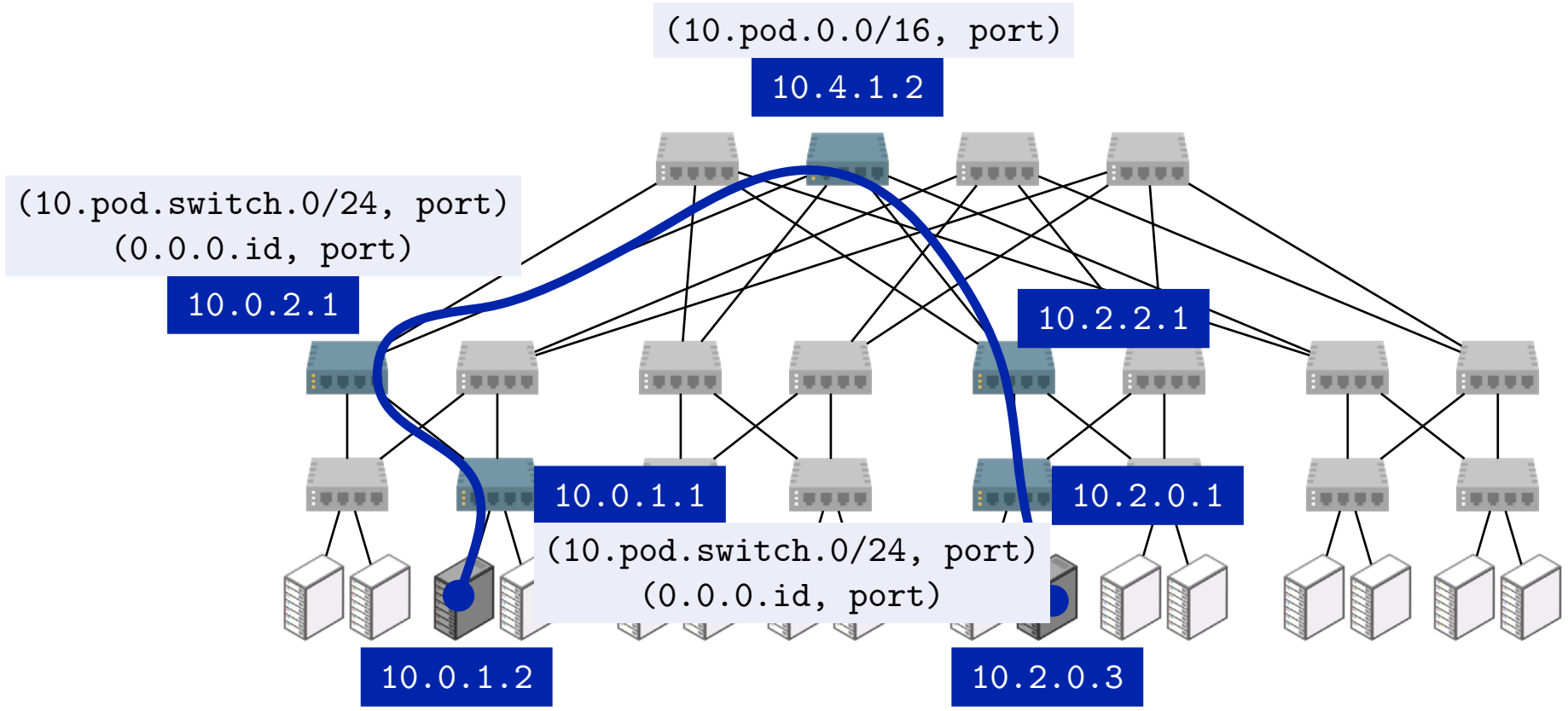
- Not perfect, but better than having a single static path between two subnets (as in OSPF)

Routing example

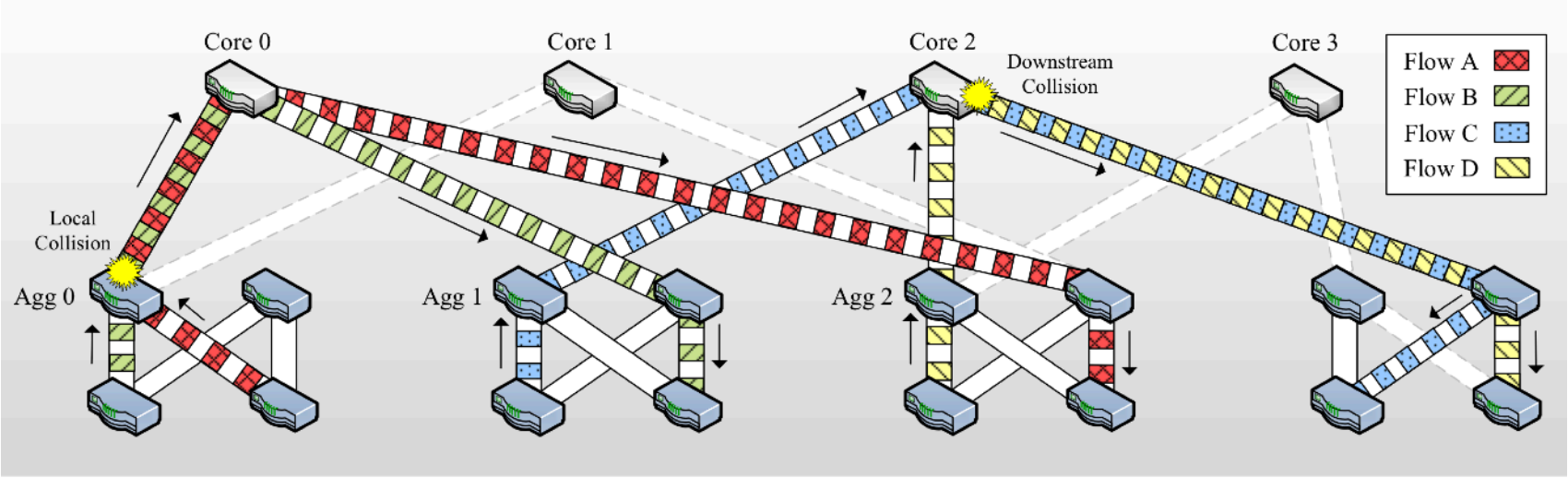


What are the forwarding rules to install on the switches?

Routing example



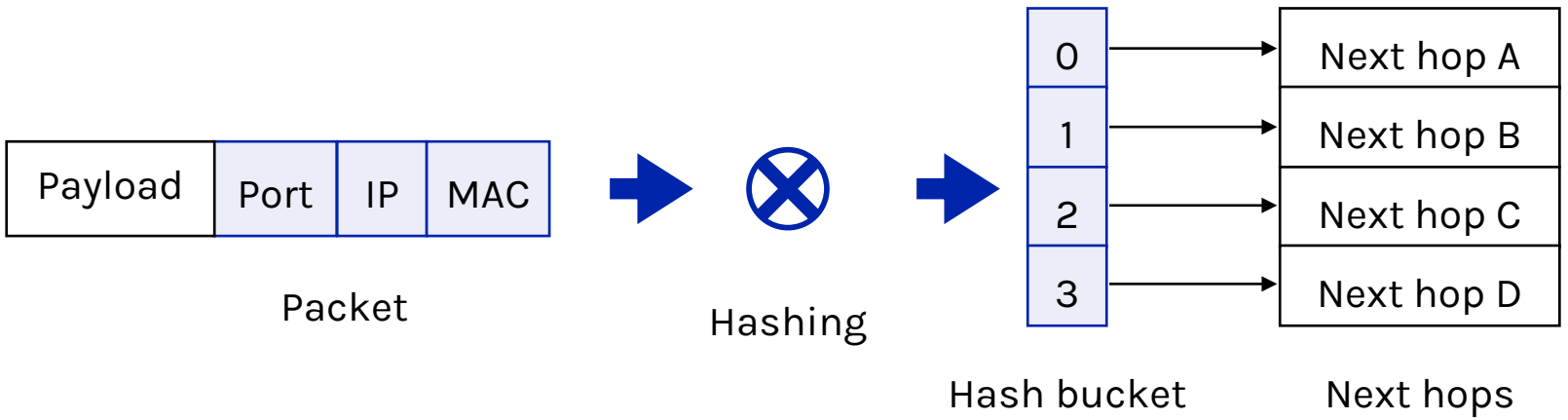
Flow collision



Hard-coded traffic diffusion can lead to bad collisions

→ performance bottleneck

Solutions to flow collisions

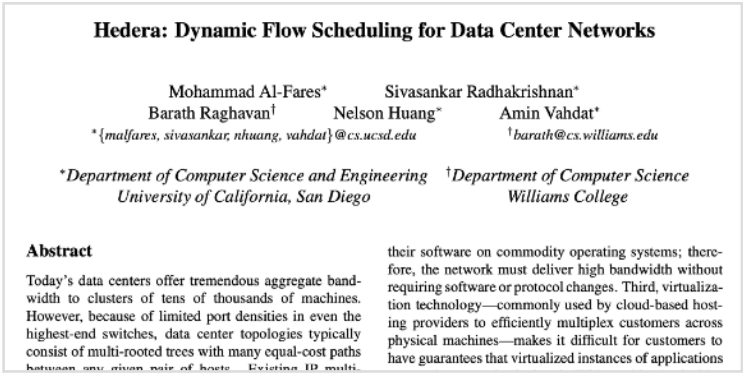


Equal-cost multi-path (ECMP)

- Static path between end-hosts → **static path for each flow**

Flow scheduling

- Have a centralized scheduler to assign flows to paths (leveraging software defined networking)



Hedera: Dynamic Flow Scheduling for Data Center Networks

Mohammad Al-Fares* Sivasankar Radhakrishnan*
Barath Raghavan† Nelson Huang* Amin Vahdat*
*{malfares, sivasankar, nhuang, vahdat}@cs.ucsd.edu †barath@cs.williams.edu

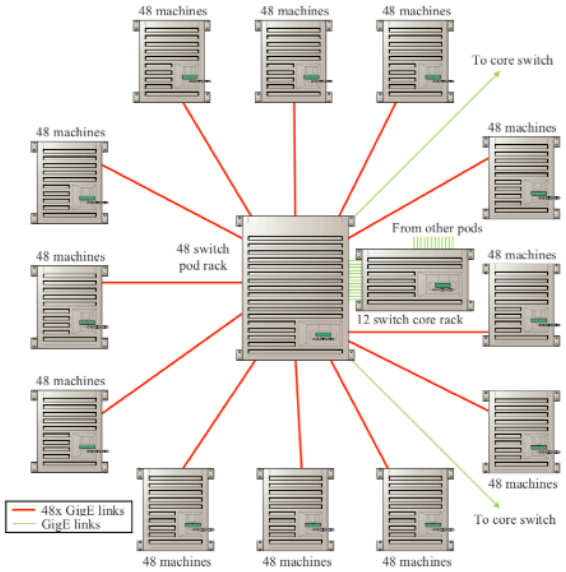
*Department of Computer Science and Engineering University of California, San Diego
†Department of Computer Science Williams College

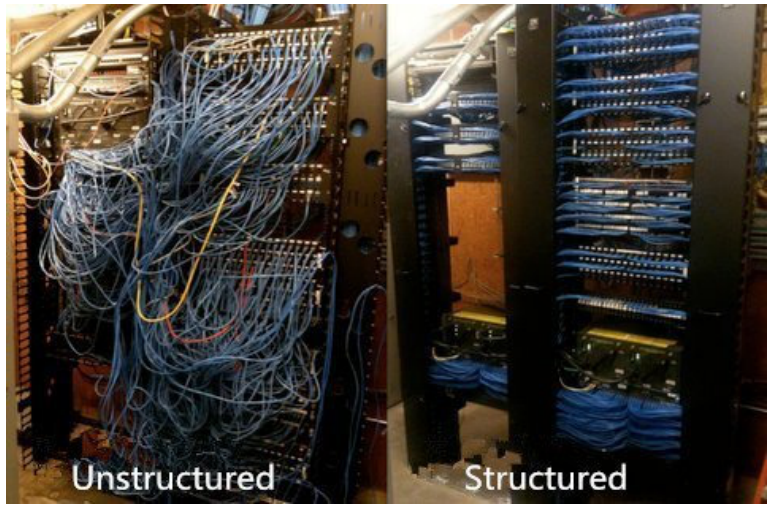
Abstract

Today's data centers offer tremendous aggregate bandwidth to clusters of tens of thousands of machines. However, because of limited port densities in even the highest-end switches, data center topologies typically consist of multi-rooted trees with many equal-cost paths

their software on commodity operating systems; therefore, the network must deliver high bandwidth without requiring software or protocol changes. Third, virtualization technology—commonly used by cloud-based hosting providers to efficiently multiplex customers across physical machines—makes it difficult for customers to have guarantees that virtualized instances of applications

Fat-tree cabling solution

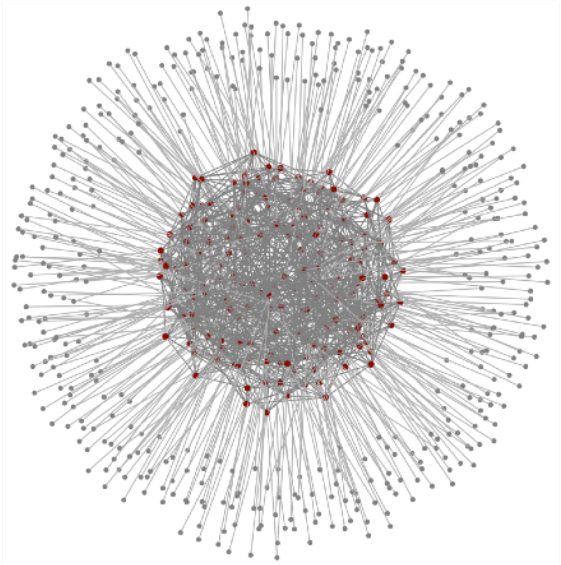




Organize switches into pod racks leveraging the regular structure of fat-tree

Fat-tree is quite regular, can we take the other extreme?

Can we general a completely random topology for the data center network?



Jellyfish: Networking Data Centers Randomly

Ankit Singla[†], Chi-Yao Hong[†], Lucian Popa[♯], P. Brighten Godfrey[†]
[†] University of Illinois at Urbana–Champaign
[♯] HP Labs

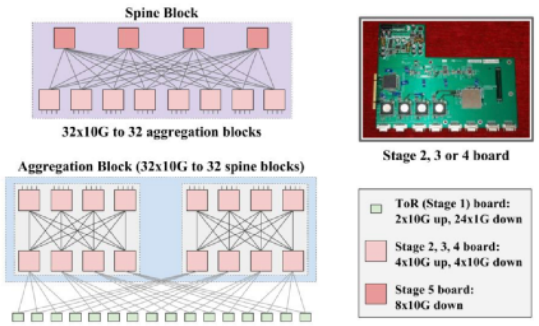
Abstract

Industry experience indicates that the ability to incrementally expand data centers is essential. However, existing high-bandwidth network designs have rigid structure that interferes with incremental expansion. We present Jellyfish, a high-capacity network interconnect which, by adopting a random graph topology, yields itself naturally to incremental expansion. Somewhat surprisingly, Jellyfish is more cost-efficient than a fat-tree, supporting as many as 25% more servers at full capacity using the same equipment at the scale of a few thousand nodes, and this advantage improves with scale. Jellyfish also allows great flexibility in building networks with different degrees of oversubscription. However, Jellyfish's unstructured design brings new challenges in routing, physical layout, and wiring. We describe approaches to resolve these challenges, and our evaluation suggests that Jellyfish could be deployed in today's data centers.

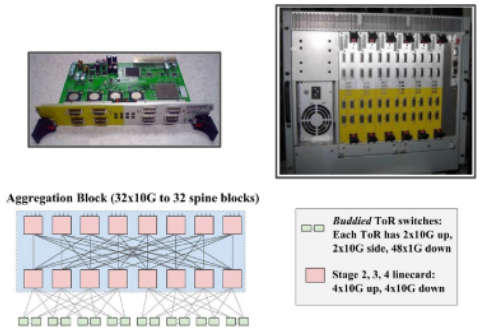
Industry experience indicates that incremental expansion is important. Consider the growth of Facebook's data center server population from roughly 30,000 in Nov. 2009 to >60,000 by June 2010 [34]. While Facebook has added entirely new data center facilities, much of this growth involves incrementally expanding existing facilities by "adding capacity on a daily basis" [33]. For instance, Facebook announced that it would double the size of its facility at Prineville, Oregon by early 2012 [16]. A 2011 survey [15] of 300 enterprises that run data centers of a variety of sizes found that 84% of firms would probably or definitely expand their data centers in 2012. Several industry products advertise incremental expandability of the server pool, including SGI's IceCube (marketed as "The Expandable Modular Data Center" [5]; expands 4 racks at a time) and HP's EcoPod [24] (a "pay-as-you-grow" enabling technology [23]).

Do current high-bandwidth data center network proposals allow incremental growth? Consider the fat-tree

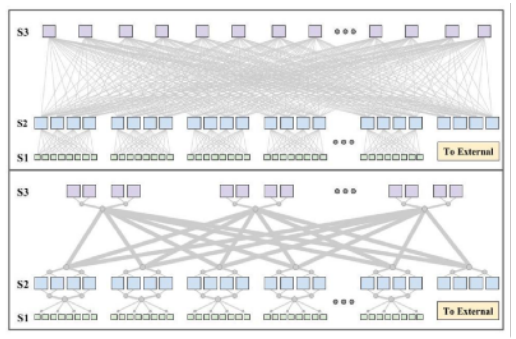
How does Google build its data center networks?



Firehose 1.0 (never in production)

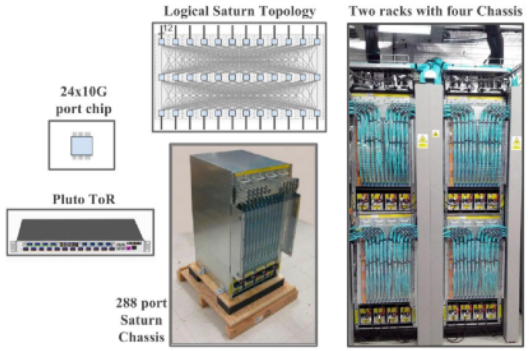


Firehose 1.1 (first production Clos, bag on the side)

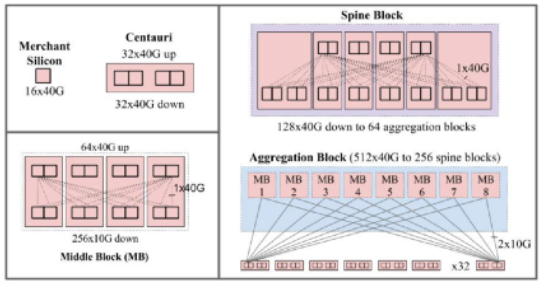


Watchtower (inter-cluster networking, depop)

How does Google build its data center networks?

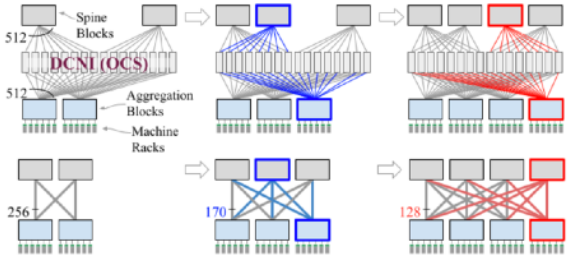


Saturn (first 10G possible between servers)



Jupiter, 2015 (uniform bandwidth, incremental deployment)

OCS: Optical Circuit Switches



Jupiter, 2022 (direct connect, reconfigurable)

Factors driven Google's designs

Motivation

- **Bandwidth demands** in the data center are doubling 12-15 months
- **Cost and operational complexity** become prohibitive
- **Availability requirements** not strict in data centers (due to abundant, cheap bandwidth)
- **Interoperability** is not a big concern (single-operator)

Design principles

- **Clos topologies** (can scale to nearly arbitrary size, in-built path diversity and redundancy)
- **Merchant silicon** (general purpose, commodity priced, exponential growth in bandwidth capacity)
- **Centralized control protocols** (to replace distributed protocols)

Google's data center network evolution

Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network

Arjun Singh, Joon Ong, Amit Agarwal, Glen Anderson, Ashby Armistead, Roy Bannon, Seb Boving, Gaurav Desai, Bob Felderman, Paulie Germano, Anand Kanagala, Jeff Provost, Jason Simmons, Eiichi Tanda, Jim Wanderer, Urs Hölzle, Stephen Stuart, and Amin Vahdat
Google, Inc.
jupiter-sigcomm@google.com

ABSTRACT

We present our approach for overcoming the cost, operational complexity, and limited scale endemic to datacenter networks a decade ago. Three themes unify the five generations of datacenter networks detailed in this paper. First, multi-stage Clos topologies built from commodity switch silicon can support cost-effective deployment of building-scale networks. Second, much of the general, but complex, decentralized network routing and management protocols supporting arbitrary deployment scenarios were overkill for single-operator, pre-planned datacenter networks. We built a centralized control mechanism based on a global configuration pushed to all datacenter switches. Third, modular hardware design coupled with simple, robust software allowed our design to also support inter-cluster and wide-area networks. Our datacenter networks run at dozens of sites across the planet, scaling in capacity by 100x over ten years to more than 1Pbps of bisection bandwidth.

abler for cloud computing. Bandwidth demands
datacenter are doubling every 12-15 months (Fig
even faster than the wide area Internet. A numbe
cent trends drive this growth. Dataset sizes are c
ing to explode with more photo/video content, lo
the proliferation of Internet-connected sensors. A
sult, network-intensive data processing pipeline
operate over ever-larger datasets. Next, Web s
can deliver higher quality results by accessing mo
on the critical path of individual requests. Finall
stellations of co-resident applications often shar
stantial data with one another in the same cluste
sider index generation, web search, and serving

Ten years ago, we found the cost and opera
complexity associated with traditional datacente
work architectures to be prohibitive. Maximu
work scale was limited by the cost and capacity
highest end switches available at any point in tin
These switches were engineering marvels, typica
cycled from products targeting wide area deploy
WAN switches were differentiated with hardwa

Jupiter Evolving: Transforming Google's Datacenter Network via Optical Circuit Switches and Software-Defined Networking

Leon Poutievski, Omid Mashayekhi, Joon Ong, Arjun Singh, Mukarram Tariq, Rui Wang, Jianan Zhang, Virginia Beauregard, Patrick Conner, Steve Gribble, Rishi Kapoor, Stephen Kratzer, Nanfang Li, Hong Liu, Karthik Nagaraj, Jason Ornstein, Samir Sawhney, Ryohei Urata, Lorenzo Vicisano, Kevin Yasumura, Shidong Zhang, Junlan Zhou, Amin Vahdat
Google
sigcomm-jupiter-evolving@google.com

WORDS

enter network, Software-defined networking, Traffic
cering, Topology engineering, Optical circuit switches.

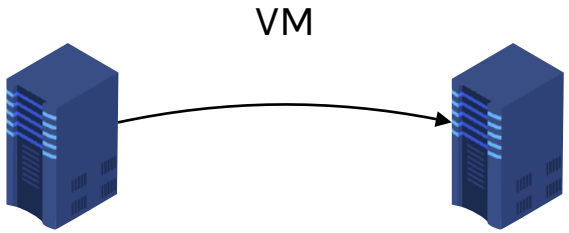
Reference Format:

outievski, Omid Mashayekhi, Joon Ong, Arjun Singh, Mukar-
ariq, Rui Wang, Jianan Zhang, Virginia Beauregard, Patrick
Rishi Kapoor, Stephen Kratzer, Nanfang Li,
rstein, Samir Sawhney, Ryohei
Shidong Zhang, Junlan
olving@google.com
Datacenter Net-
efined Network-
22). ACM, New
544216.3544265

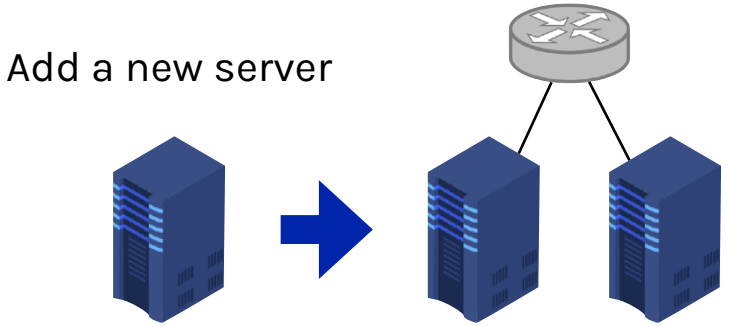
Clos topologies [1, 2, 14,
silicon have enabled cost effec-
enable building-scale datacenter networks as the basis
oud infrastructure. A range of networked services, ma-
learning workloads, and storage infrastructure lever-
niform, high bandwidth connectivity among tens of
ands of servers to great effect.
ile there is tremendous progress, managing the het-

How to achieve flexible management of data center networks?

Issues in fat-tree



No support for seamless VM migration
IP addresses are location-dependent and migration would break the TCP connection



Plug-and-play not possible
IP addresses have to be pre-assigned to both switches and hosts

It seems that the location-dependent IP address is the culprit. How to address this issue?

L2 vs L3 data center network fabric

Technique	Plug-and-play	Scalability	Small switch state	Seamless VM migration
Layer 2: flat MAC addresses				
Layer 3: IP addresses				

L2 vs L3 data center network fabric

Technique	Plug-and-play	Scalability	Small switch state	Seamless VM migration
Layer 2: flat MAC addresses	+	− Broadcast strom	− Exact match leads to too many entries	+
Layer 3: IP addresses	− Location-dependent addresses mandate manual configuration	+−	+	− IP endpoint changes

Switch state: L2 vs L3

Commodity switches have ~640KB of low latency, power hungry, expensive on chip memory (e.g., TCAM): can store 32-64K forwarding entries

In a data center with 500K servers, there could be 10 million virtual endpoints that need to be addressed

- Flat address (MAC address)

10 million address mappings	→	~100 MB on-chip memory	→	~150x over the limit

- Hierarchical address (IP address)

100-1000 address mappings (using prefix/suffix matching)	→	~10 KB of memory	→	easily accommodated in today's switches

PortLand

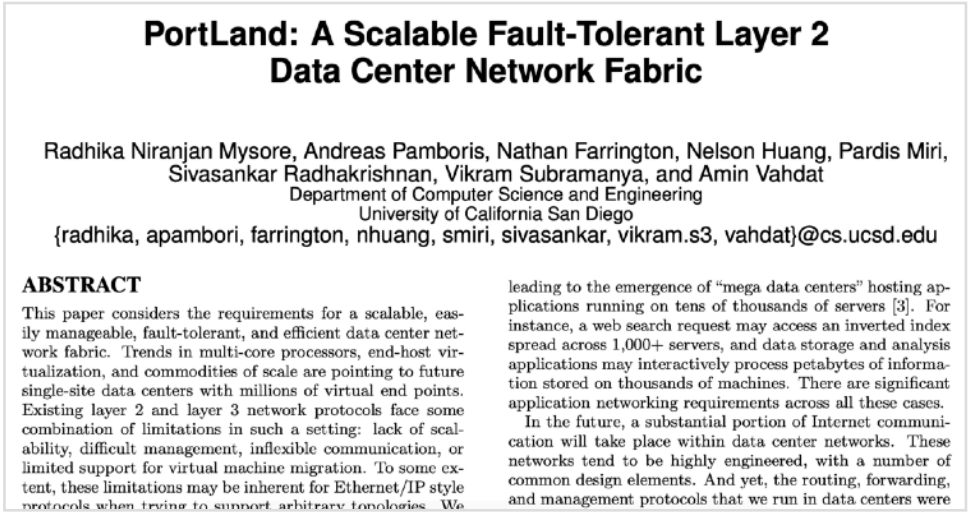
Main idea: separate node location from node identifier

- Host IP: node identifier
- Pseudo MAC (PMAC): node location

Fabric manager

- Maintains IP → PMAC mapping for ARP
- Facilitates fault tolerance

PMAC sufficient for positional forwarding



PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric

Radhika Niranjan Mysore, Andreas Pamboris, Nathan Farrington, Nelson Huang, Pardis Miri, Sivasankar Radhakrishnan, Vikram Subramanya, and Amin Vahdat
Department of Computer Science and Engineering
University of California San Diego
{radhika, apambori, farrington, nhuang, smiri, sivasankar, vikram.s3, vahdat}@cs.ucsd.edu

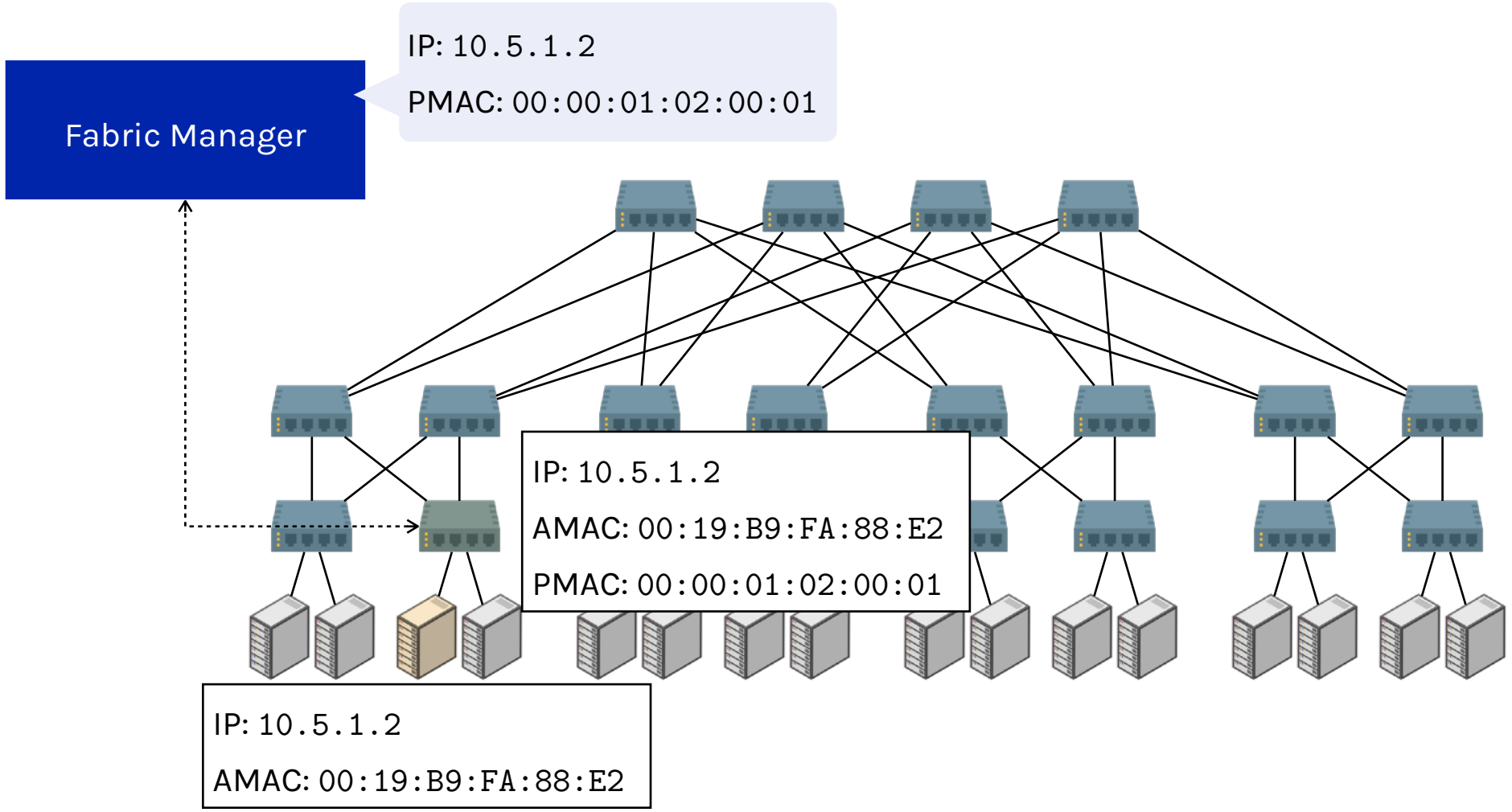
ABSTRACT

This paper considers the requirements for a scalable, easily manageable, fault-tolerant, and efficient data center network fabric. Trends in multi-core processors, end-host virtualization, and commodities of scale are pointing to future single-site data centers with millions of virtual end points. Existing layer 2 and layer 3 network protocols face some combination of limitations in such a setting: lack of scalability, difficult management, inflexible communication, or limited support for virtual machine migration. To some extent, these limitations may be inherent for Ethernet/IP style protocols when trying to support arbitrary topologies. We

leading to the emergence of "mega data centers" hosting applications running on tens of thousands of servers [3]. For instance, a web search request may access an inverted index spread across 1,000+ servers, and data storage and analysis applications may interactively process petabytes of information stored on thousands of machines. There are significant application networking requirements across all these cases.

In the future, a substantial portion of Internet communication will take place within data center networks. These networks tend to be highly engineered, with a number of common design elements. And yet, the routing, forwarding, and management protocols that we run in data centers were

PortLand design



PMAC and location discovery

PMAC: `pod.position.port.vmid`

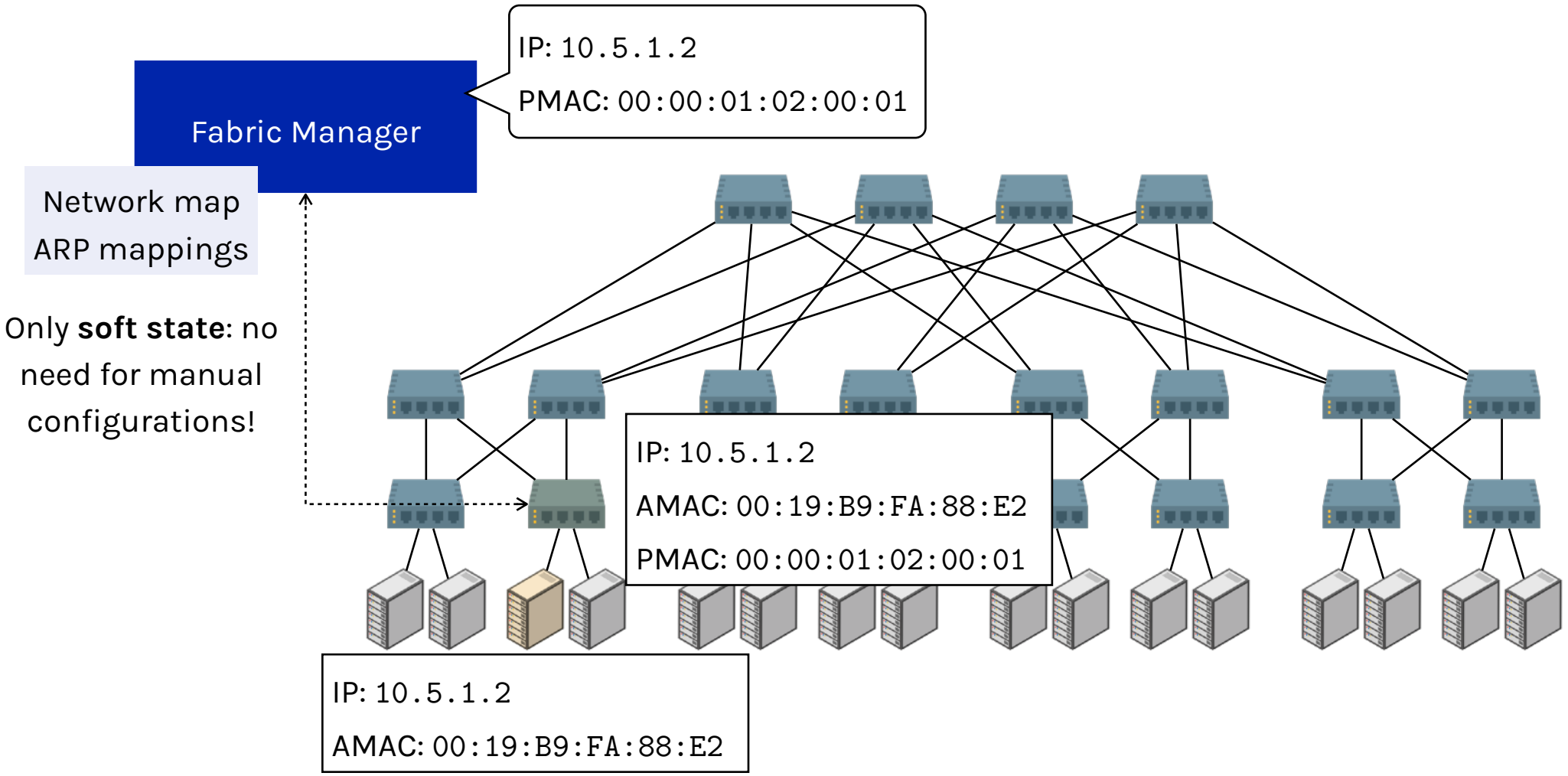
Switches self-discover location by exchanging Location Discovery Messages (LDMs):

- Tree-level/role: based on neighbor identity
- Pod number: fetch from the Fabric manager
- Position number: aggregation switches help ToR switches choose unique position number

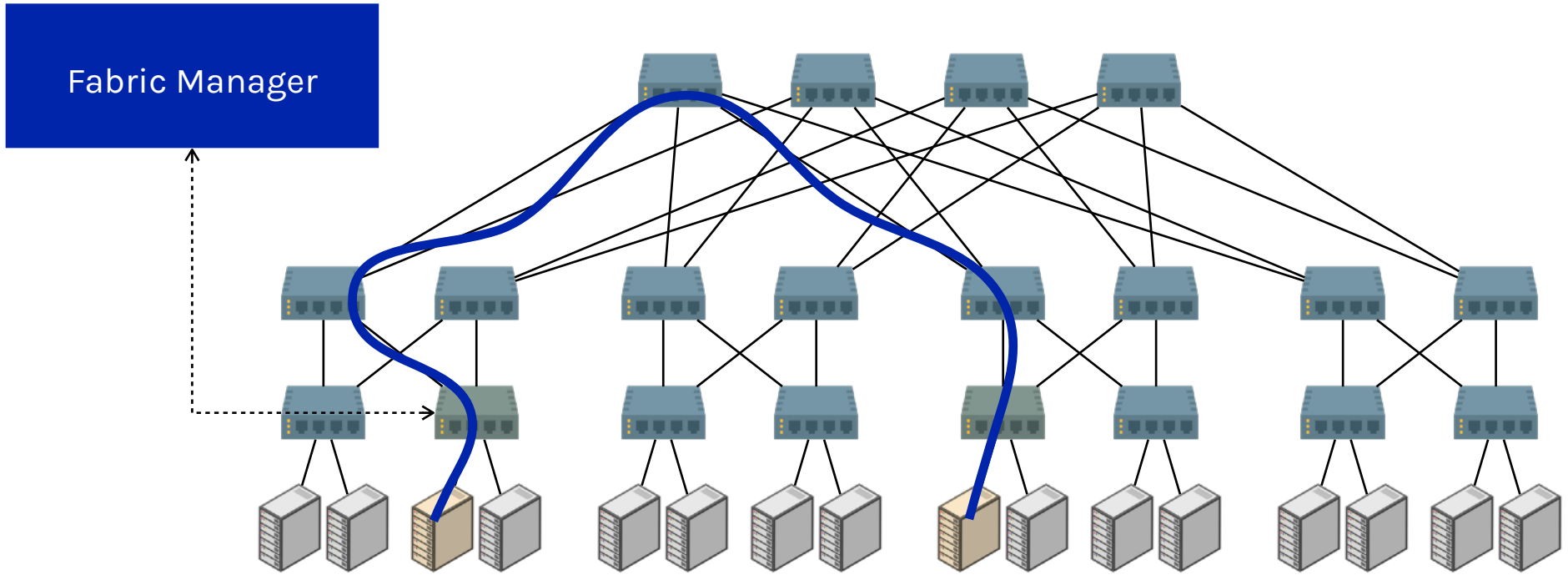
Advantages

- Plug-and-play
- Small switch state

Fabric manager

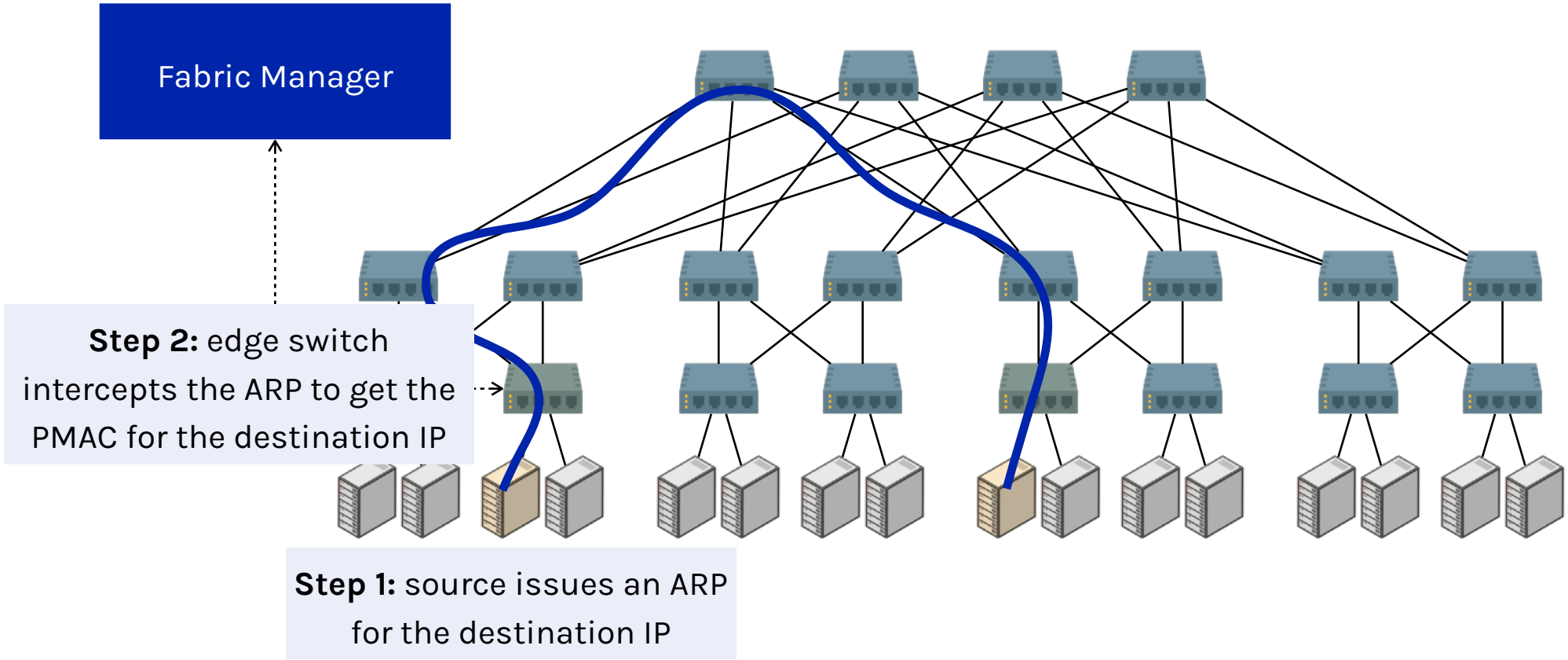


PortLand workflow



Step 1: source issues an ARP

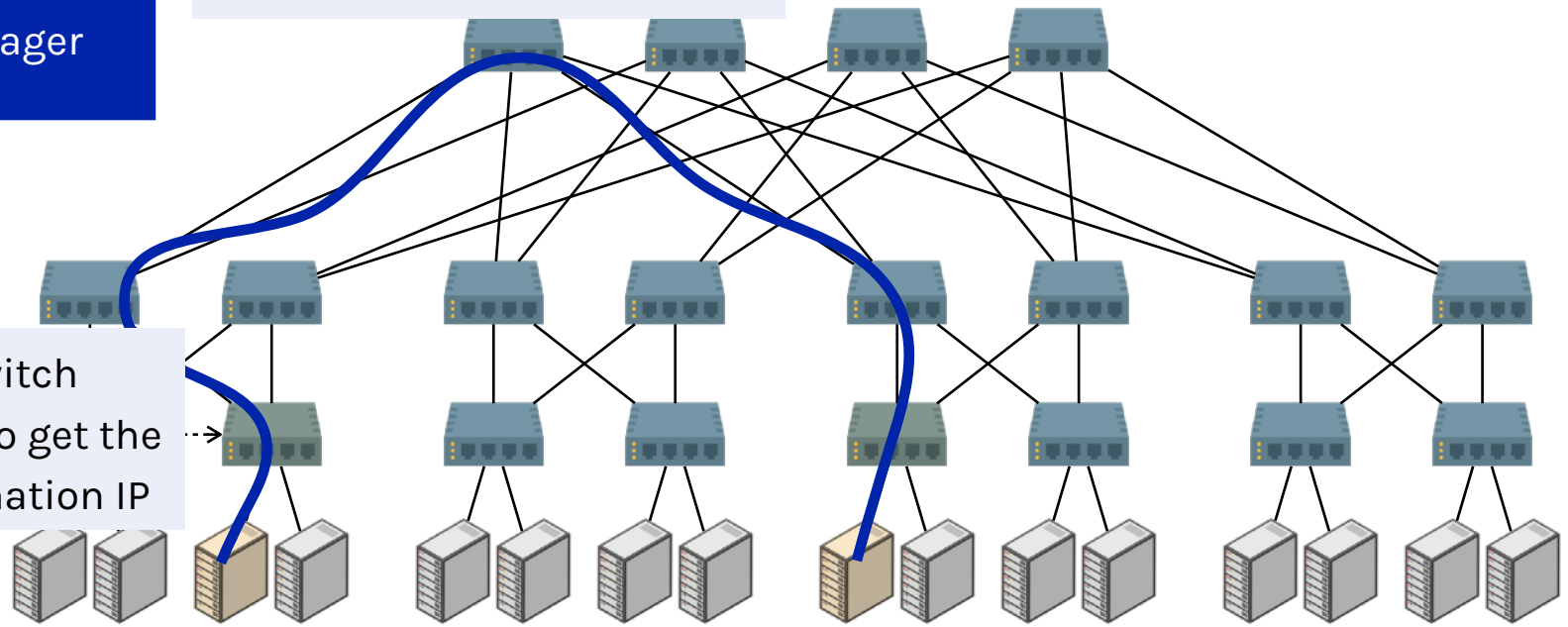
PortLand workflow



PortLand workflow

In case of no matching entries, FM broadcasts the ARP to all core switches

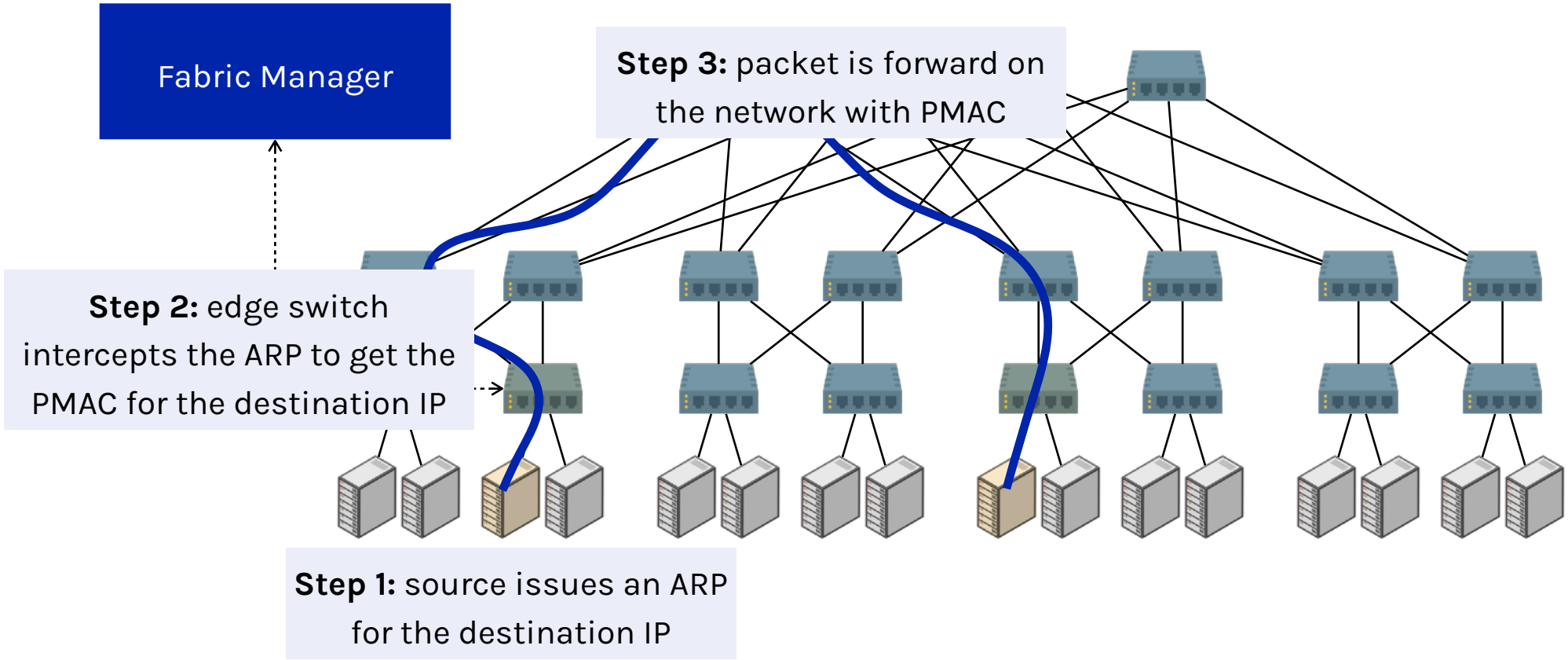
Fabric Manager



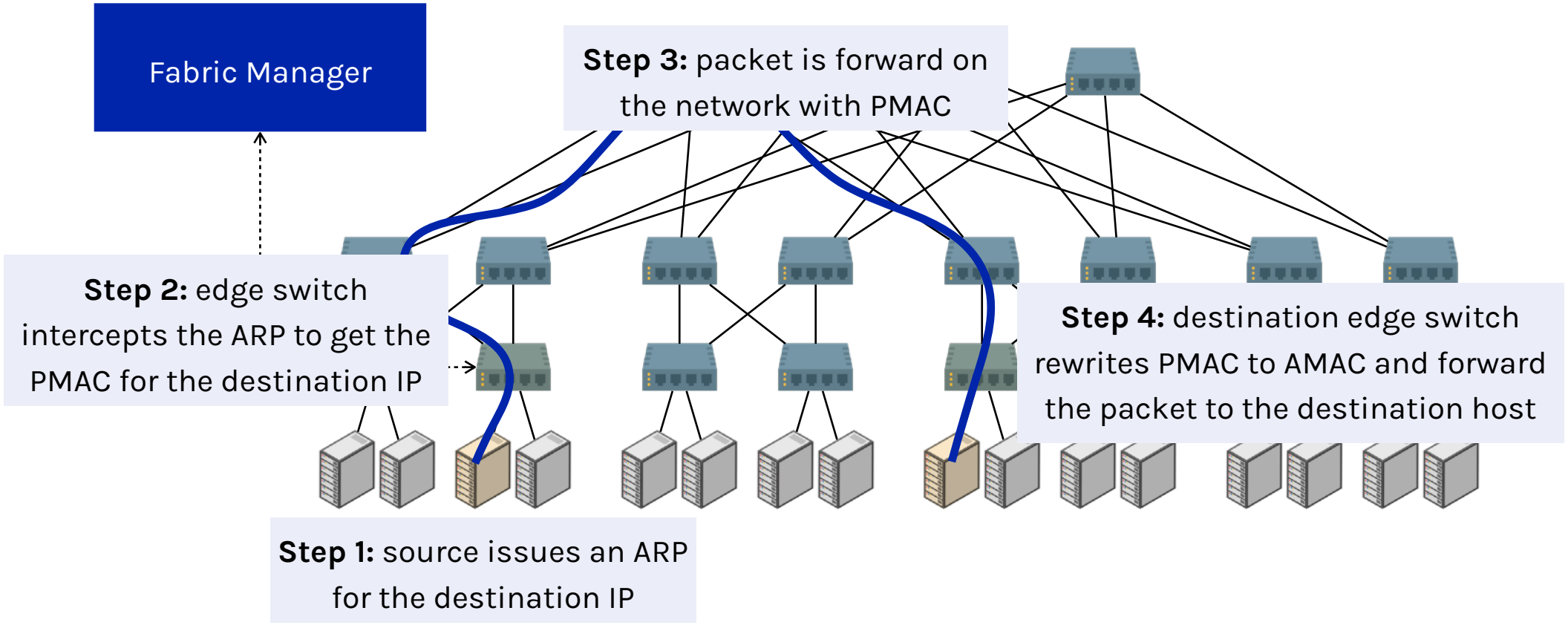
Step 2: edge switch intercepts the ARP to get the PMAC for the destination IP

Step 1: source issues an ARP for the destination IP

PortLand workflow



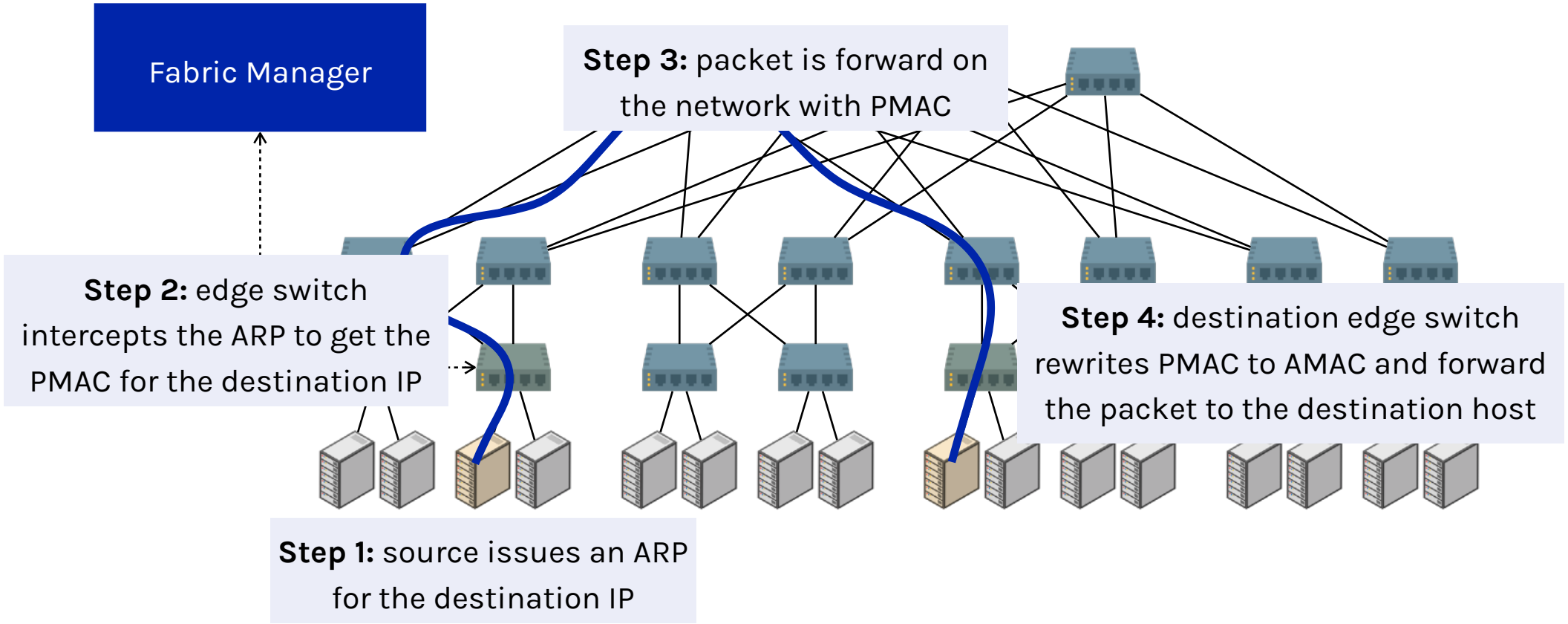
PortLand workflow



PortLand workflow

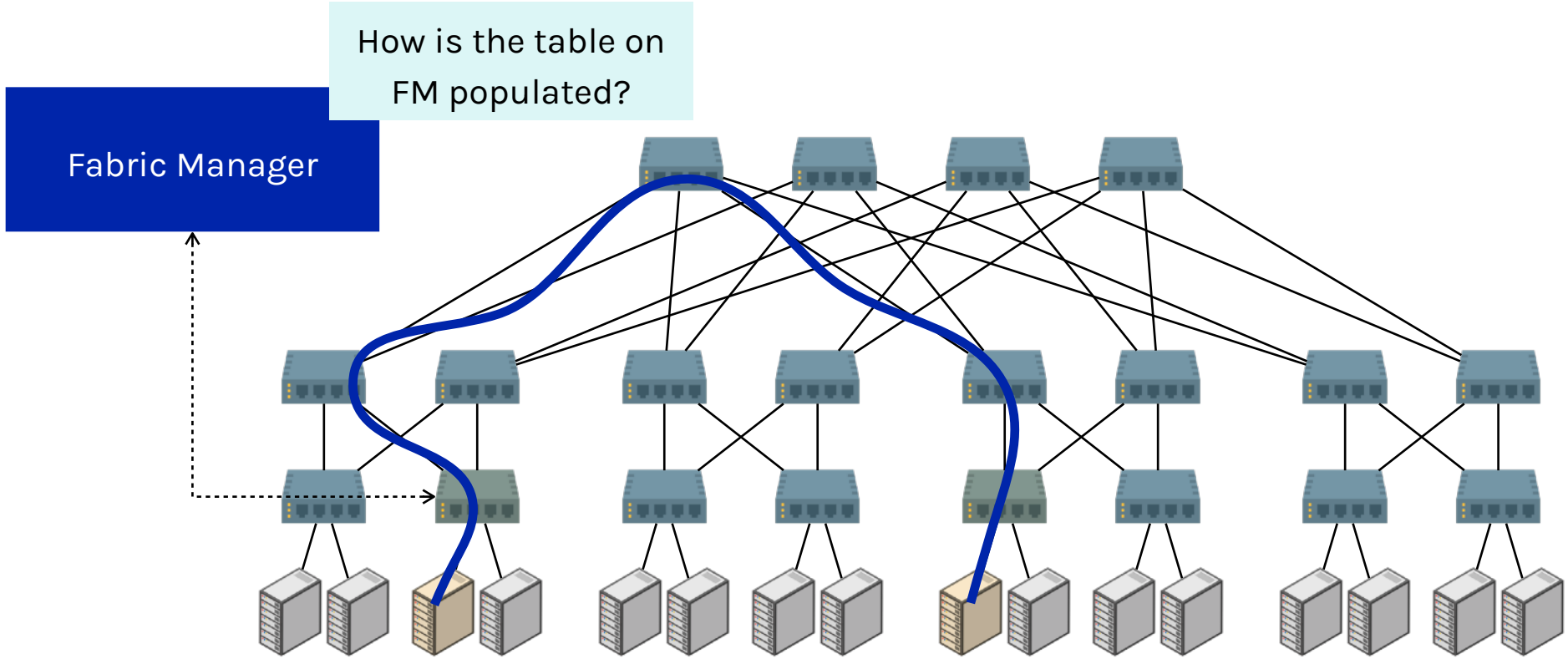
Straightforward hardware support:

- No modification needed for hosts
- PMAC <-> AMAC translation on edge switches
- Other switches forward based on prefix-matching on PMAC



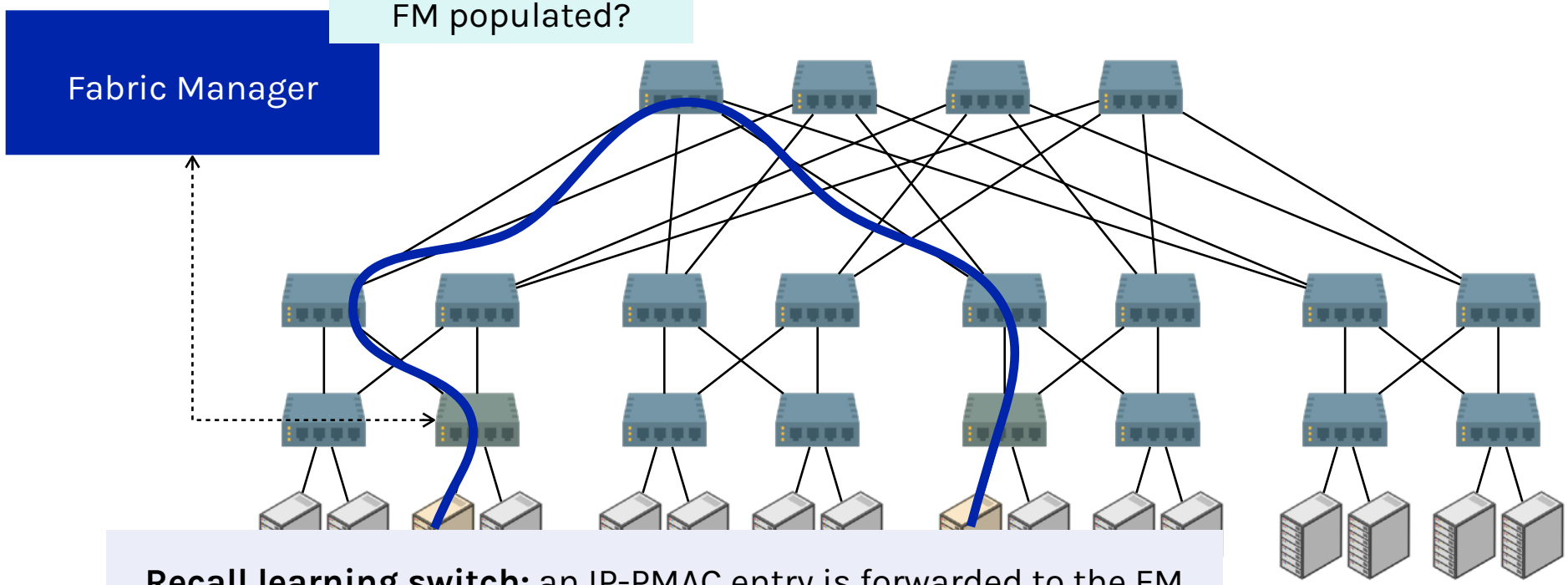
PortLand workflow

How is the table on FM populated?



PortLand workflow

How is the table on FM populated?

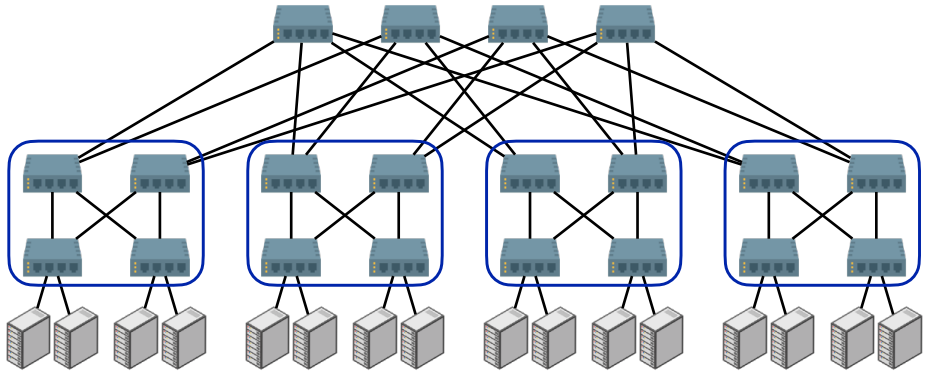


Recall learning switch: an IP-PMAC entry is forwarded to the FM every time the edge switch sees a new IP

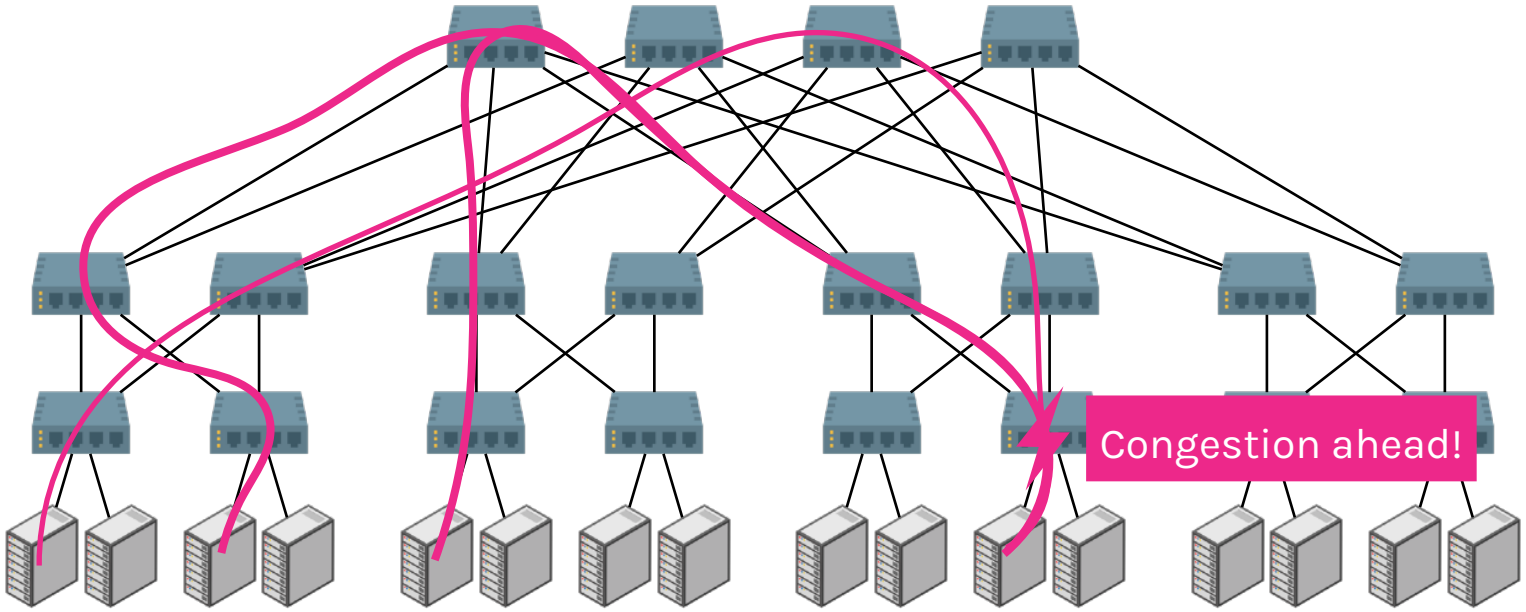
Summary

Data center networking

- Topology, performance (bisection bandwidth, over-subscription ratio)
- Architecture design: fat-tree
- Routing in fat-tree
- L2 vs. L3 addressing for data center networking
- PortLand design
- Forwarding and routing in PortLand



Next time: data center transport



How to deal with congestions in data center networks?