Institut für Kommunikations-Technik

Data-Center Network Management

Network Management

Prof. Dr. Panagiotis Papadimitriou

Leibniz Universität Hannover

Data-Center Networks

Data Centers (1)

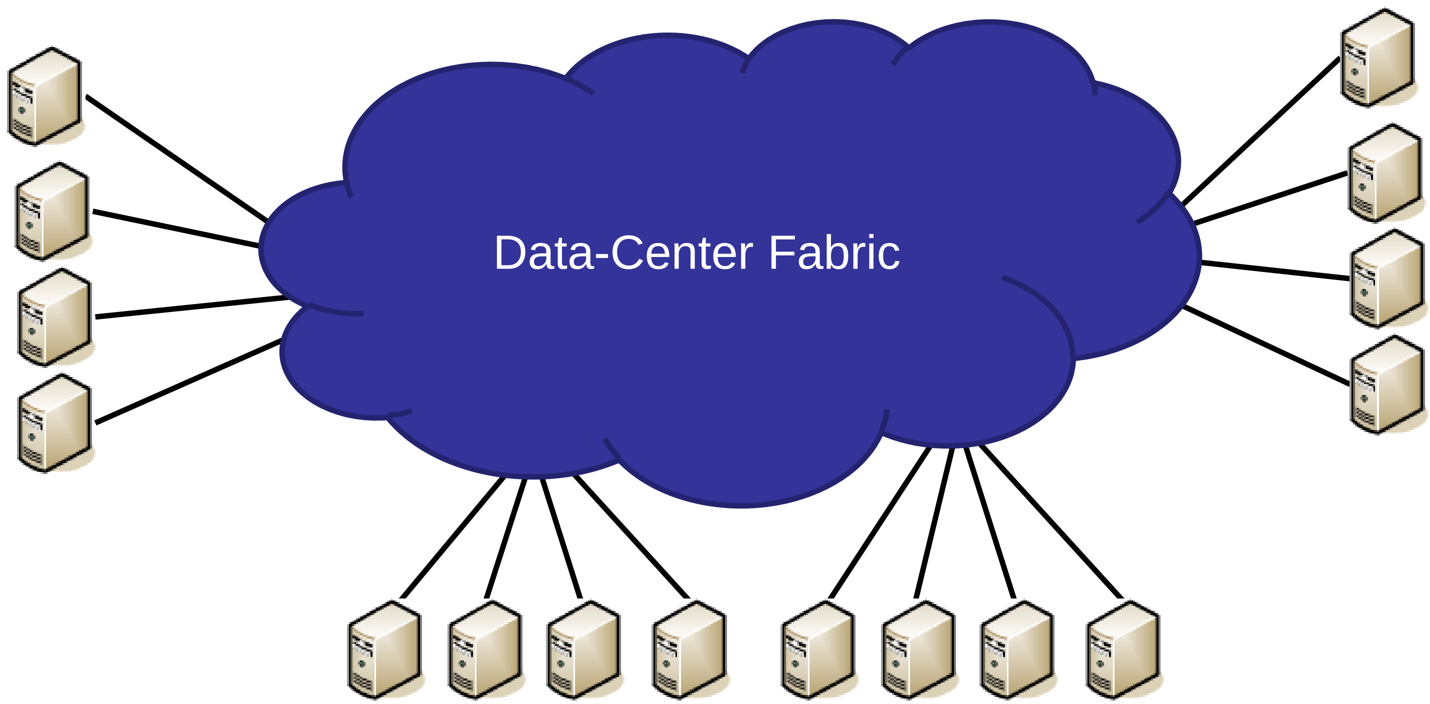
- Major players build and operate data centers:
  - Amazon, Google, HP, Microsoft, Facebook, etc.

- Features:
  - Massive scale:
    - Usually tens of thousands servers (up to hundreds of thousands servers
  - Commoditization:
    - Wide use of commodity (inexpensive) hardware (i.e., servers and switches)
  - Server virtualization:
    - Widespread adoption of server virtualization to maximize resource utilization
    - A large number of virtual machines may be hosted on a single server using technologies such as VMWare, Xen, etc.

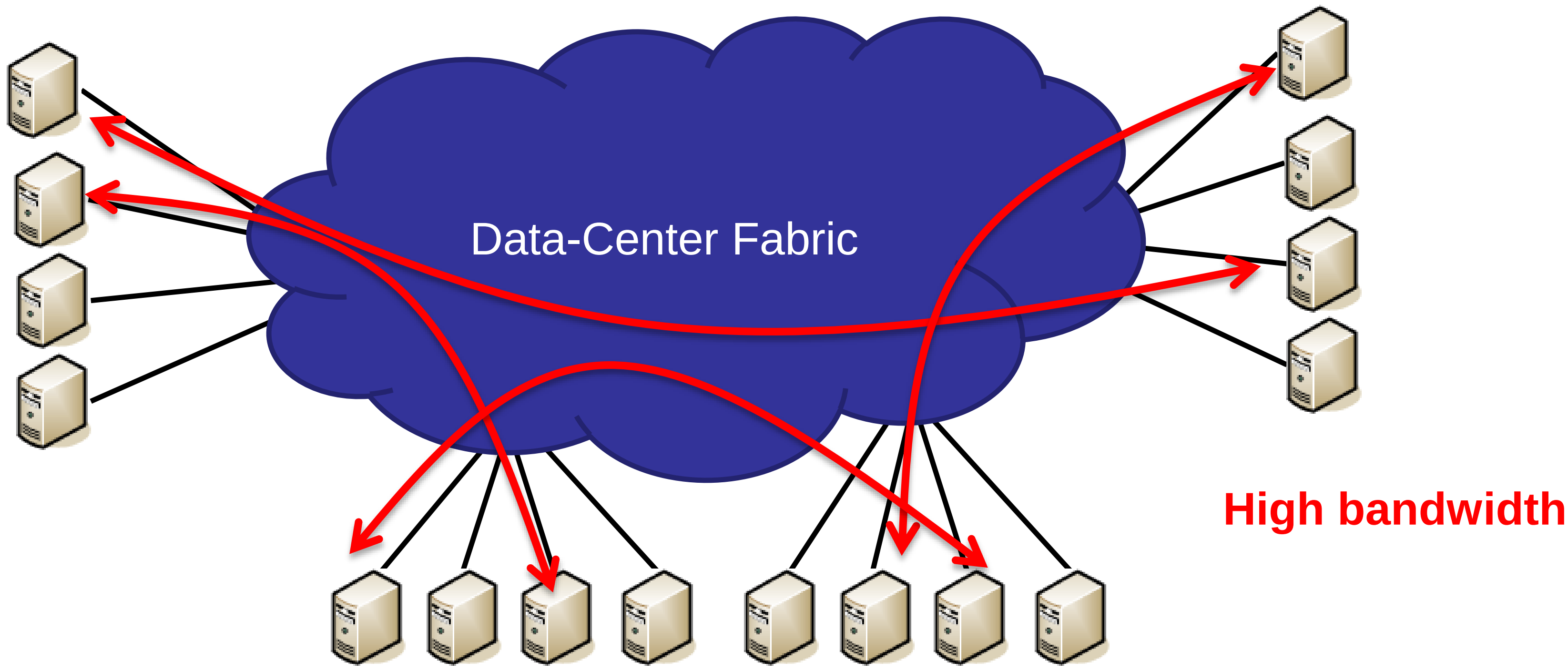
Data Centers (2)

- Main Goals:
  - Performance:
    - Fast execution of applications (e.g., Map-Reduce jobs)
  - Energy efficiency:
    - Increased consolidation (e.g., shut down redundant servers) to achieve energy savings
  - Minimal configuration overhead:
    - Host/service discovery
    - Mobility support (e.g., VM migration)
  - Scalability:
    - Scaling data-center network to tens or hundreds of servers

Data-Center Network

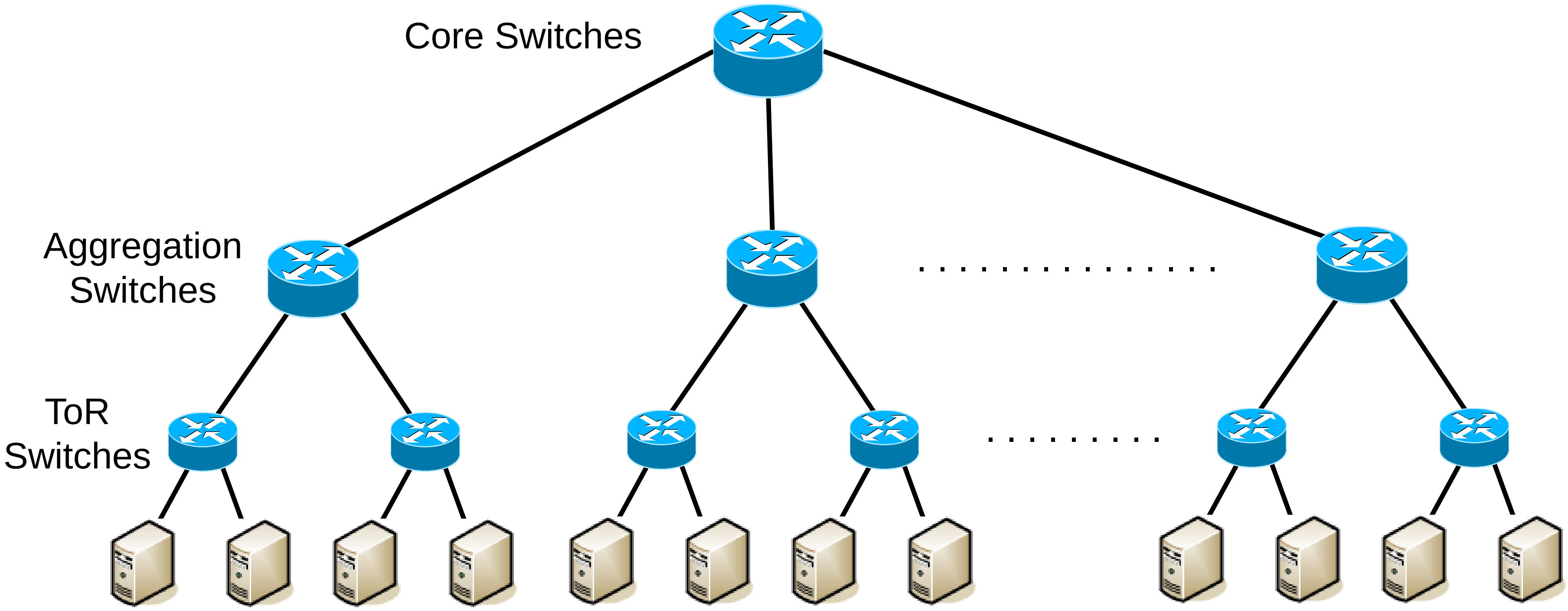


Data-Center Network



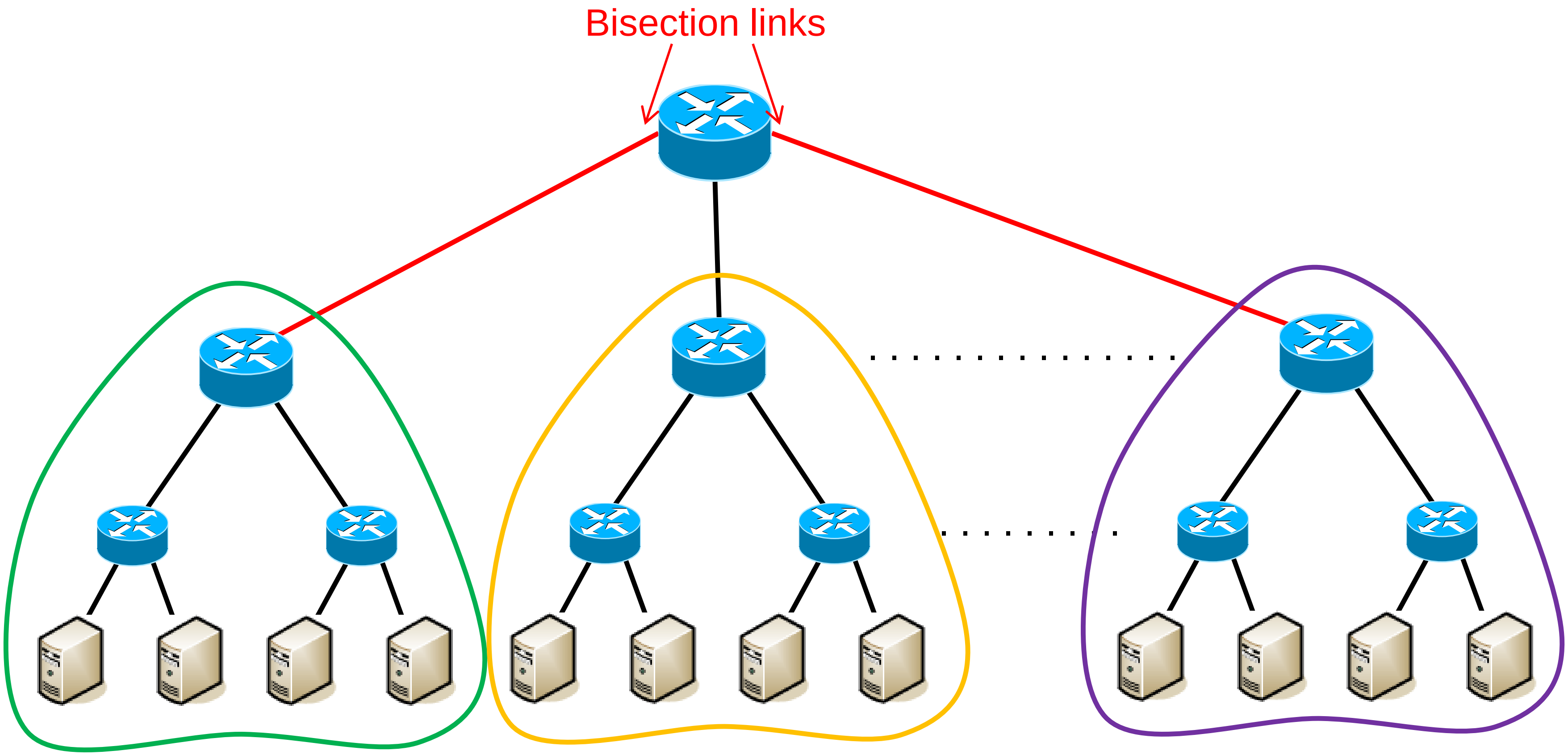
Traditional DC Network Topology

- Tree-based topologies:
  - Suitable either for flows that enter/leave the DC or for data transfers within the same rack
  - Poor performance with data transfers among different racks
  - Not fault-tolerant



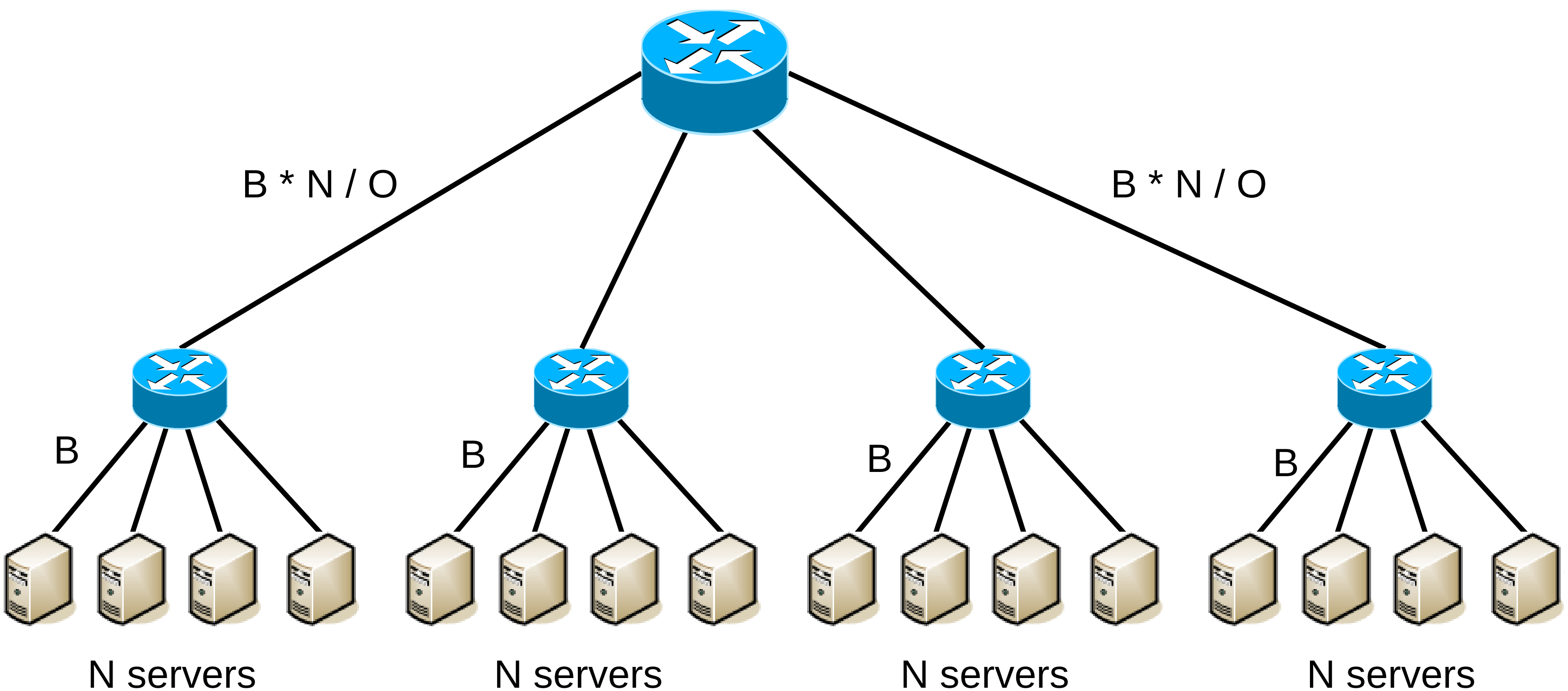
Bisection Bandwidth

- Bisection bandwidth is the sum of the core link bandwidths
- Full bisection bandwidth:
  - The total bandwidth of the core links should be equal to the sum of all server links bandwidth



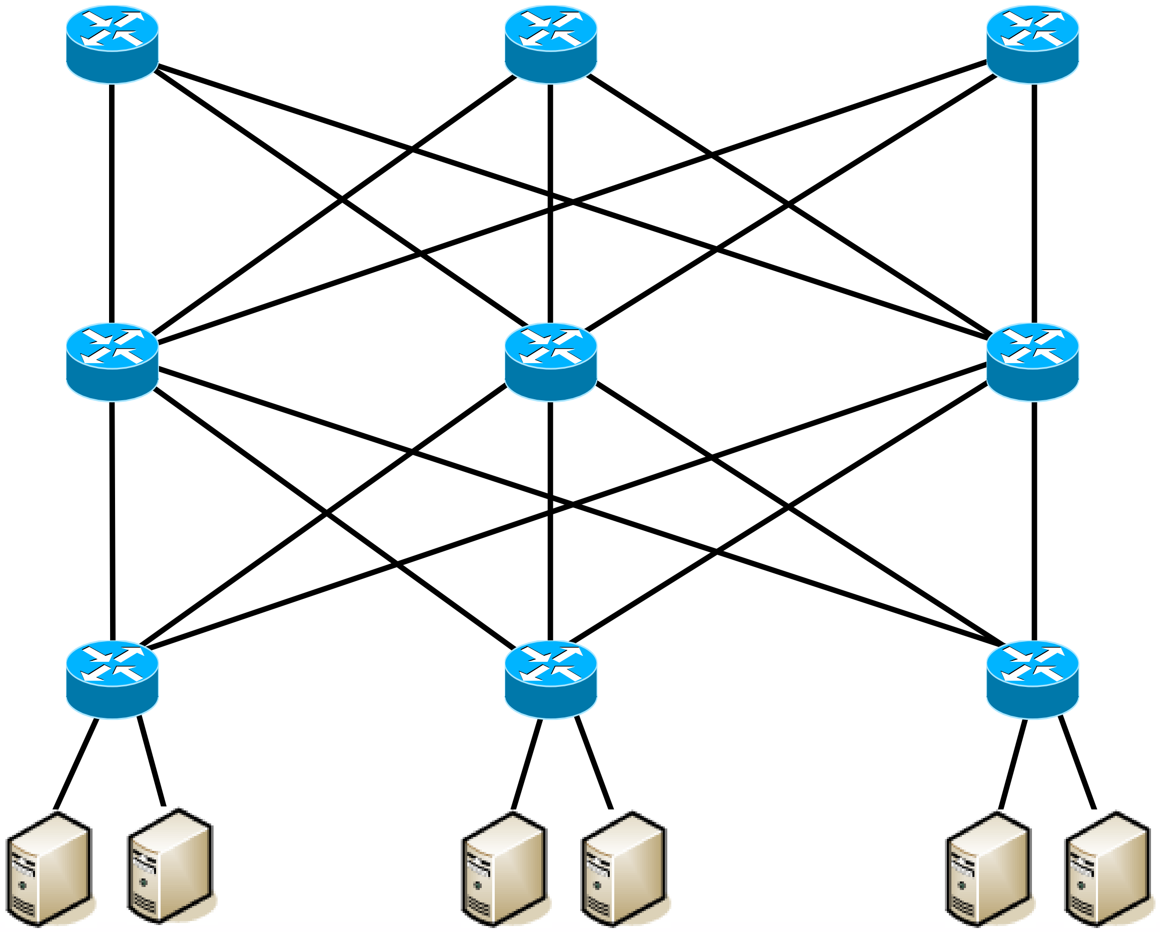
Data-Center Network Oversubscription

- Oversubscription factor (O):
  - is adjusted taking into account the locality of data transfers and the traffic rate/pattern



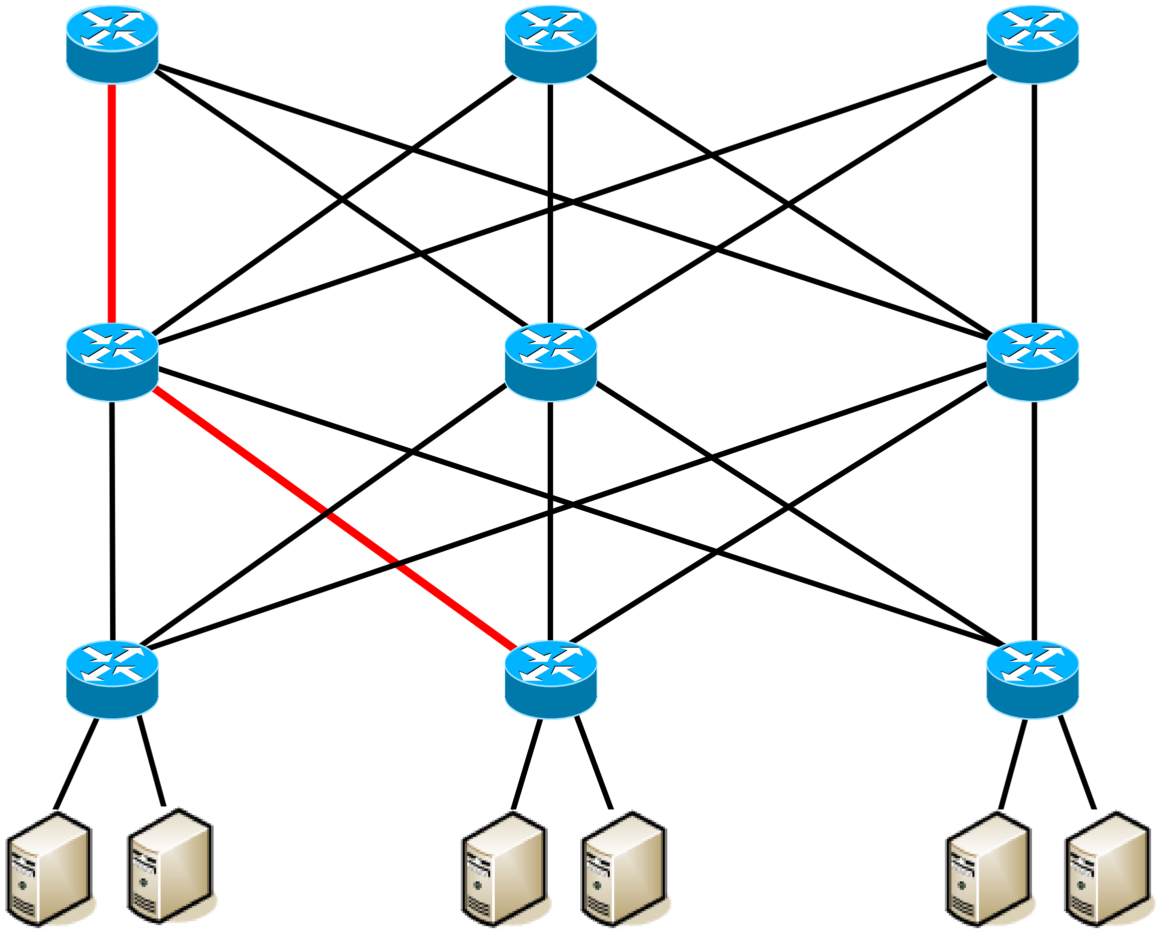
Clos Topology

- Multiple Equal Cost Paths (ECMP)
- Full bisection bandwidth



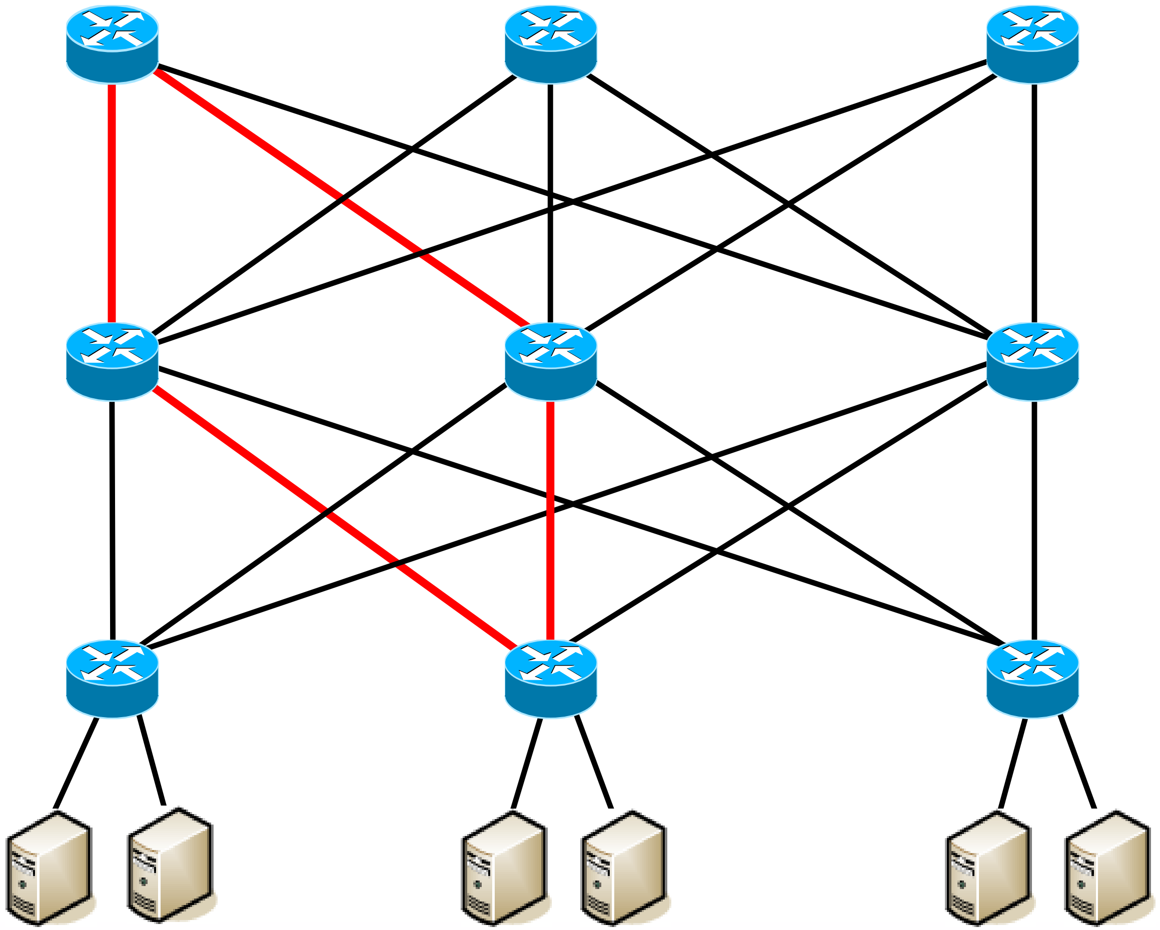
Clos Topology

- Multiple Equal Cost Paths (ECMP)
- Full bisection bandwidth



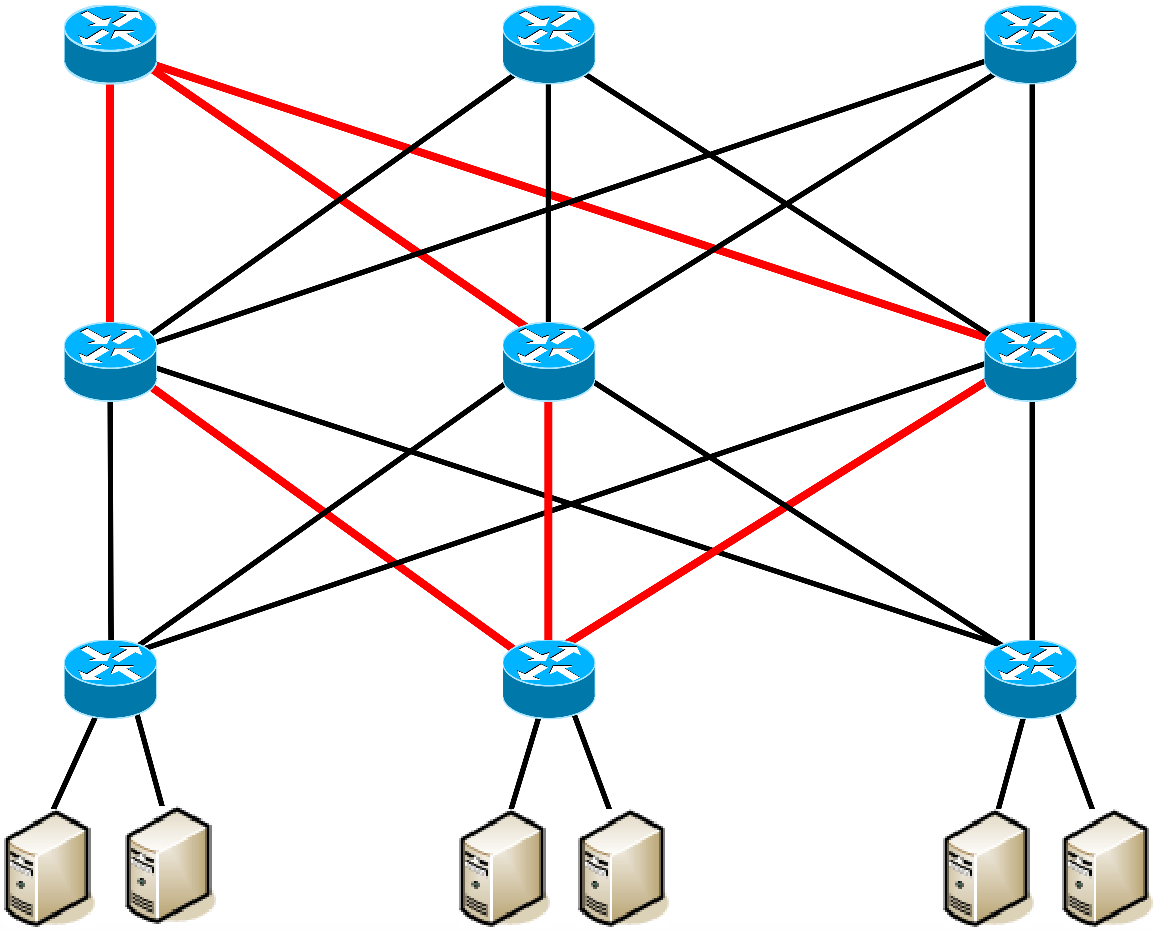
Clos Topology

- Multiple Equal Cost Paths (ECMP)
- Full bisection bandwidth



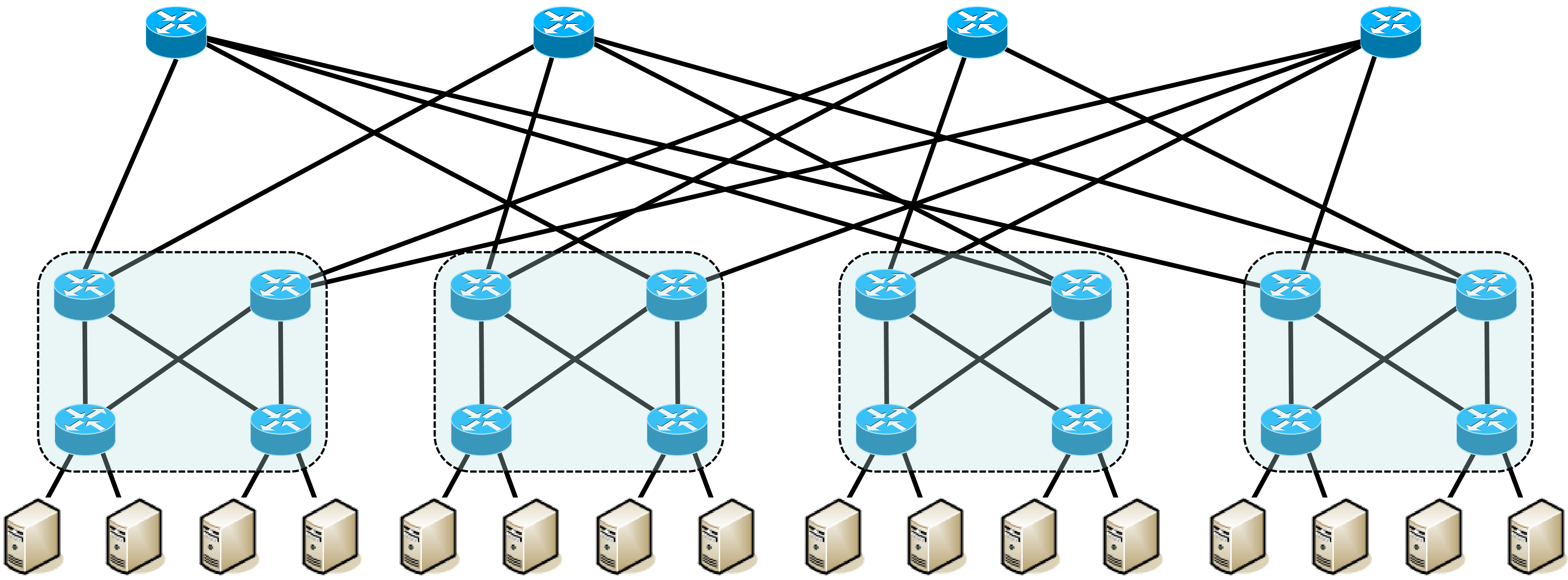
Clos Topology

- Multiple Equal Cost Paths (ECMP)
- Full bisection bandwidth



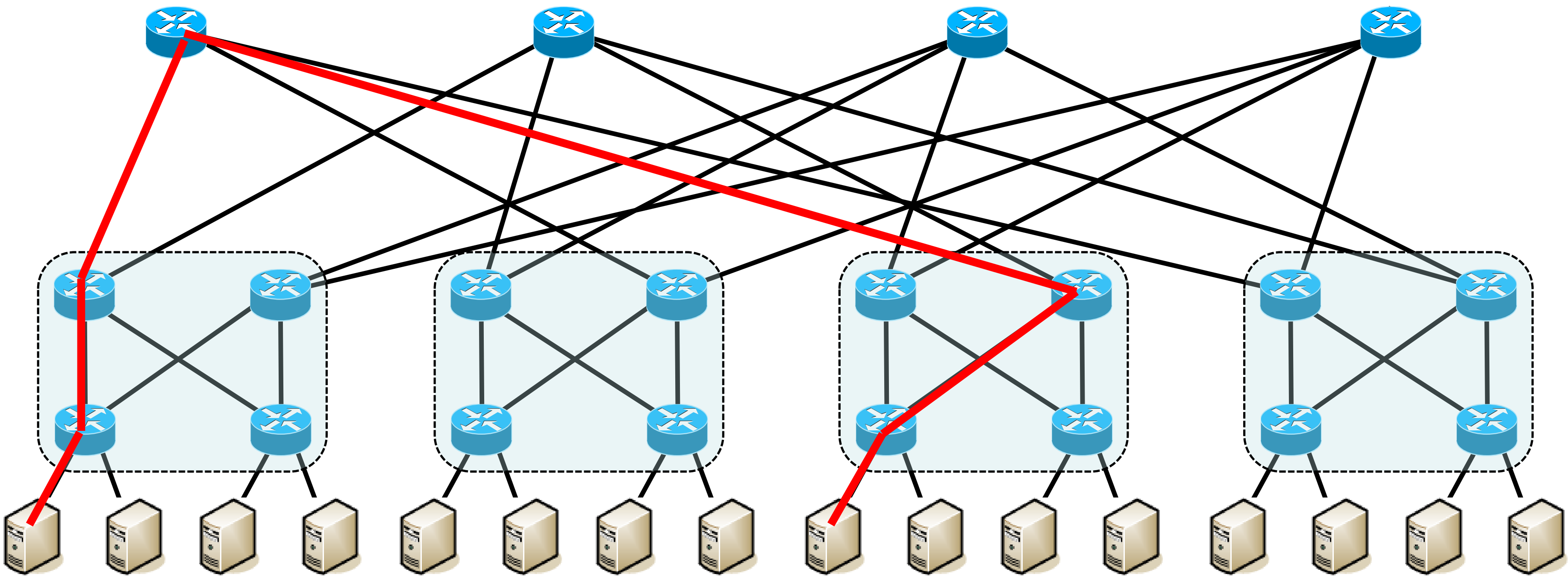
Fat-Tree Topology

- Special case of a Clos topology:
  - $(\mathbf{k}/2)^2$ **k**-port core switches, **k** pods each one with two layers of **k**/2 switches and $\mathbf{k}^2/4$ servers, $\mathbf{k}^3/4$ servers in total



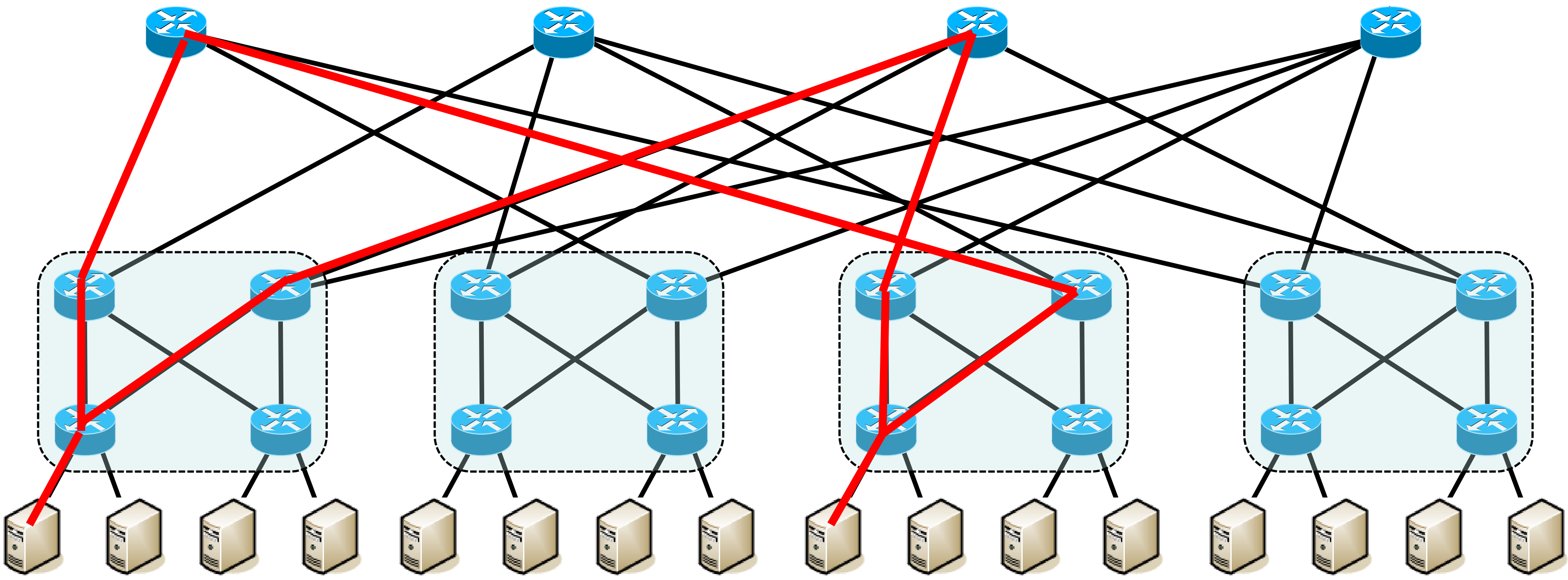
Fat-Tree Topology

- Special case of a Clos topology:
  - ($\mathbf{k}$/2)^2 **k**-port core switches, **k** pods each one with two layers of **k**/2 switches and $\mathbf{k}^2$/4 servers, $\mathbf{k}^3$/4 servers in total



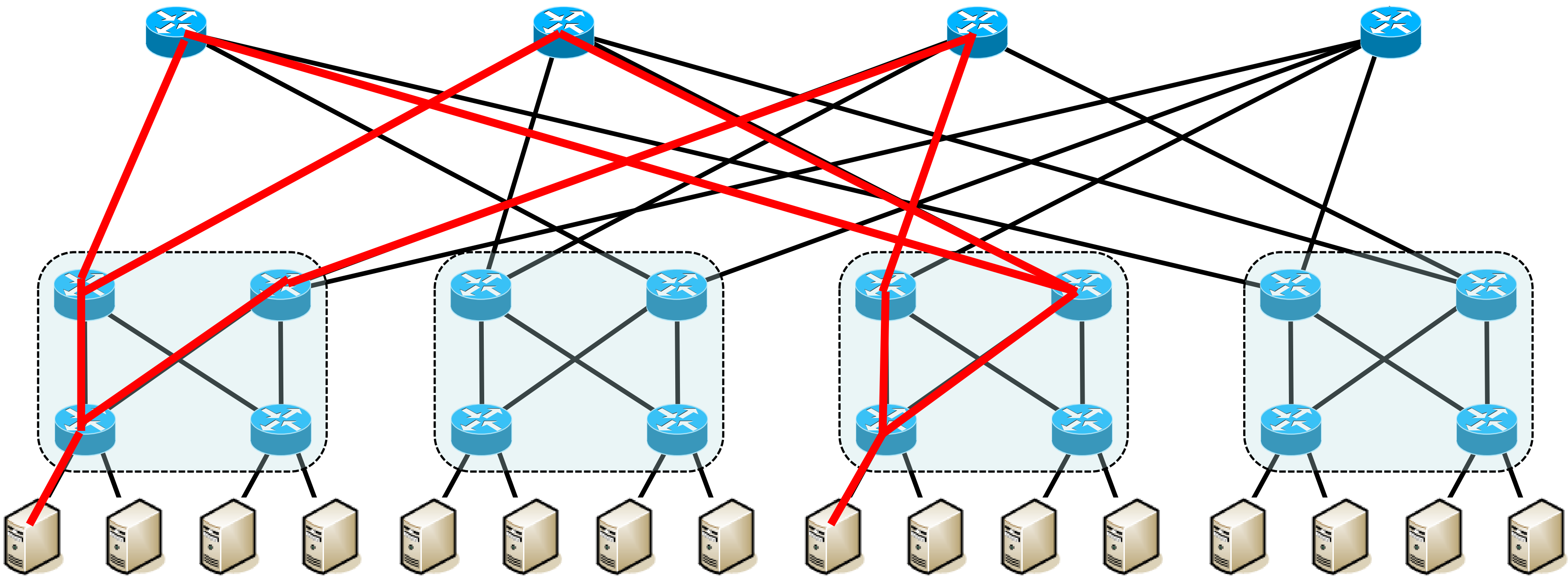
Fat-Tree Topology

- Special case of a Clos topology:
  - $(\mathbf{k}/2)^2$ **k**-port core switches, **k** pods each one with two layers of **k**/2 switches and $\mathbf{k}^2/4$ servers, $\mathbf{k}^3/4$ servers in total



Fat-Tree Topology

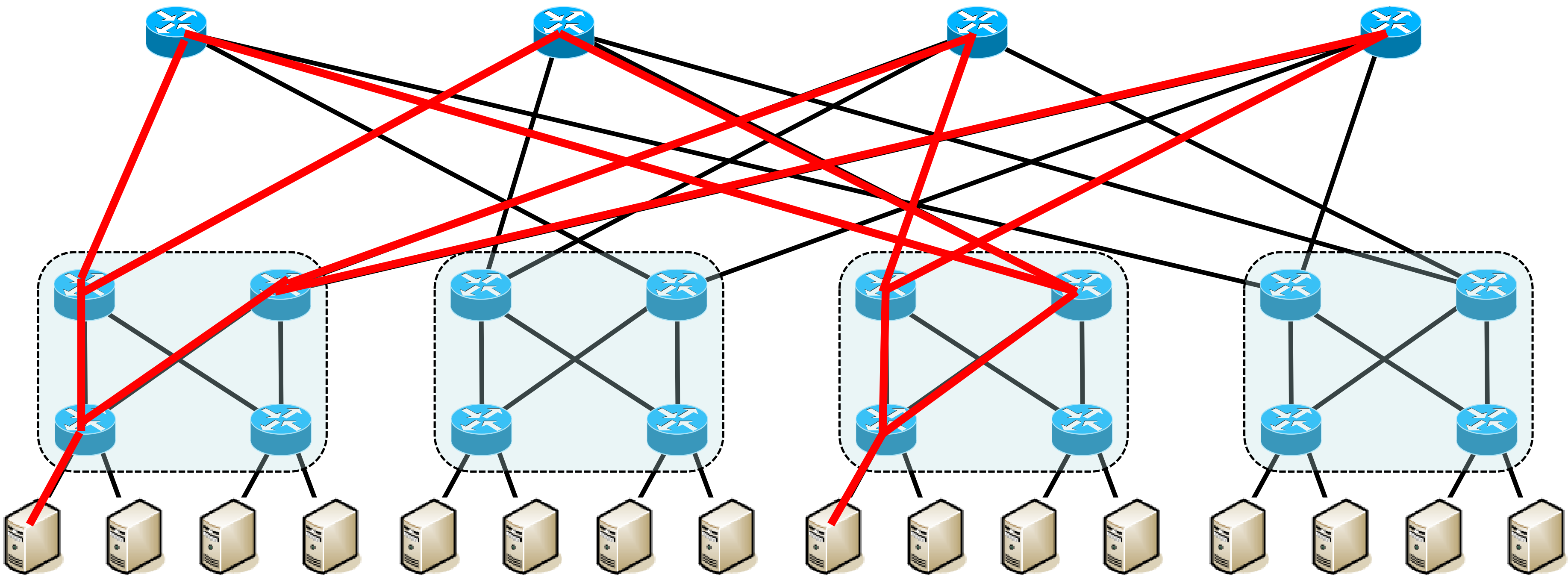
- Special case of a Clos topology:
  - $(\mathbf{k}/2)^2$ **k**-port core switches, **k** pods each one with two layers of **k**/2 switches and $\mathbf{k}^2/4$ servers, $\mathbf{k}^3/4$ servers in total



Fat-Tree Topology

Institut für Kommunikations-Technik

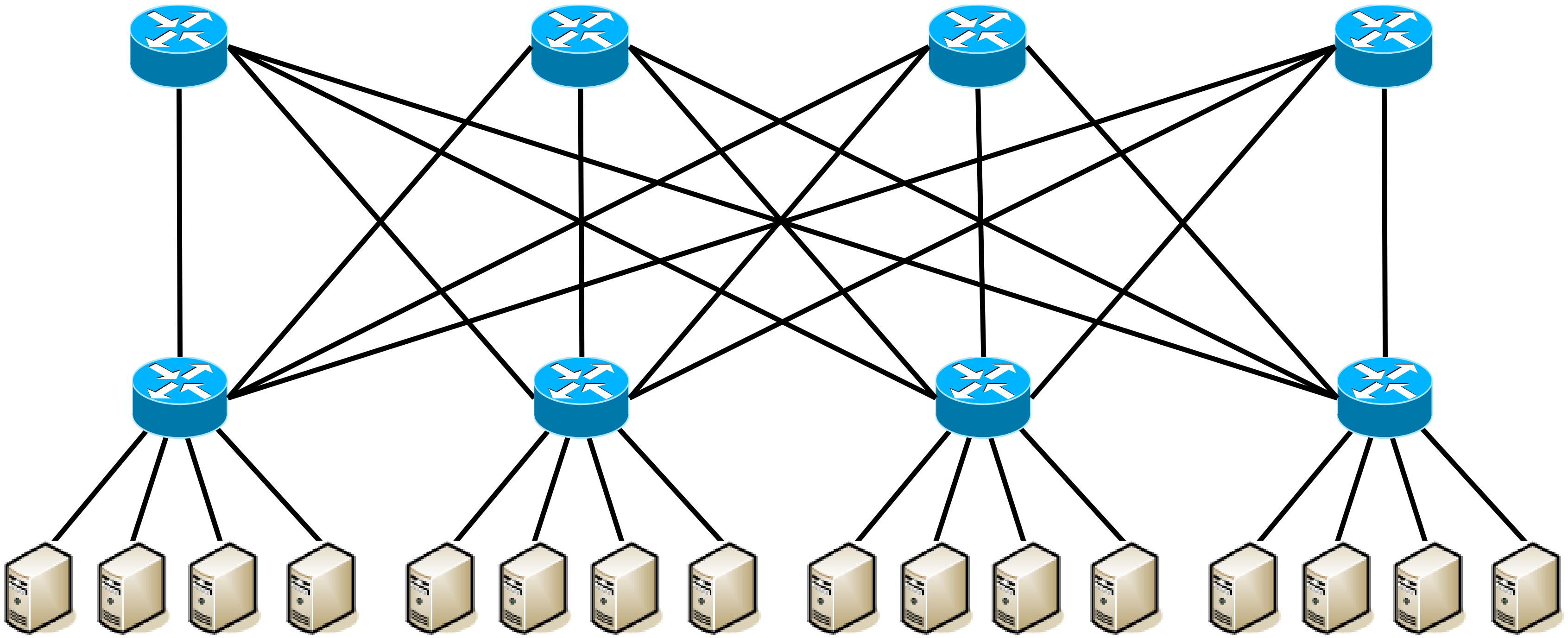
- Special case of a Clos topology:
  - ($\mathbf{k}$/2)^2 **k**-port core switches, **k** pods each one with two layers of **k**/2 switches and $\mathbf{k}^2$/4 servers, $\mathbf{k}^3$/4 servers in total



Limitations of Ethernet Architecture

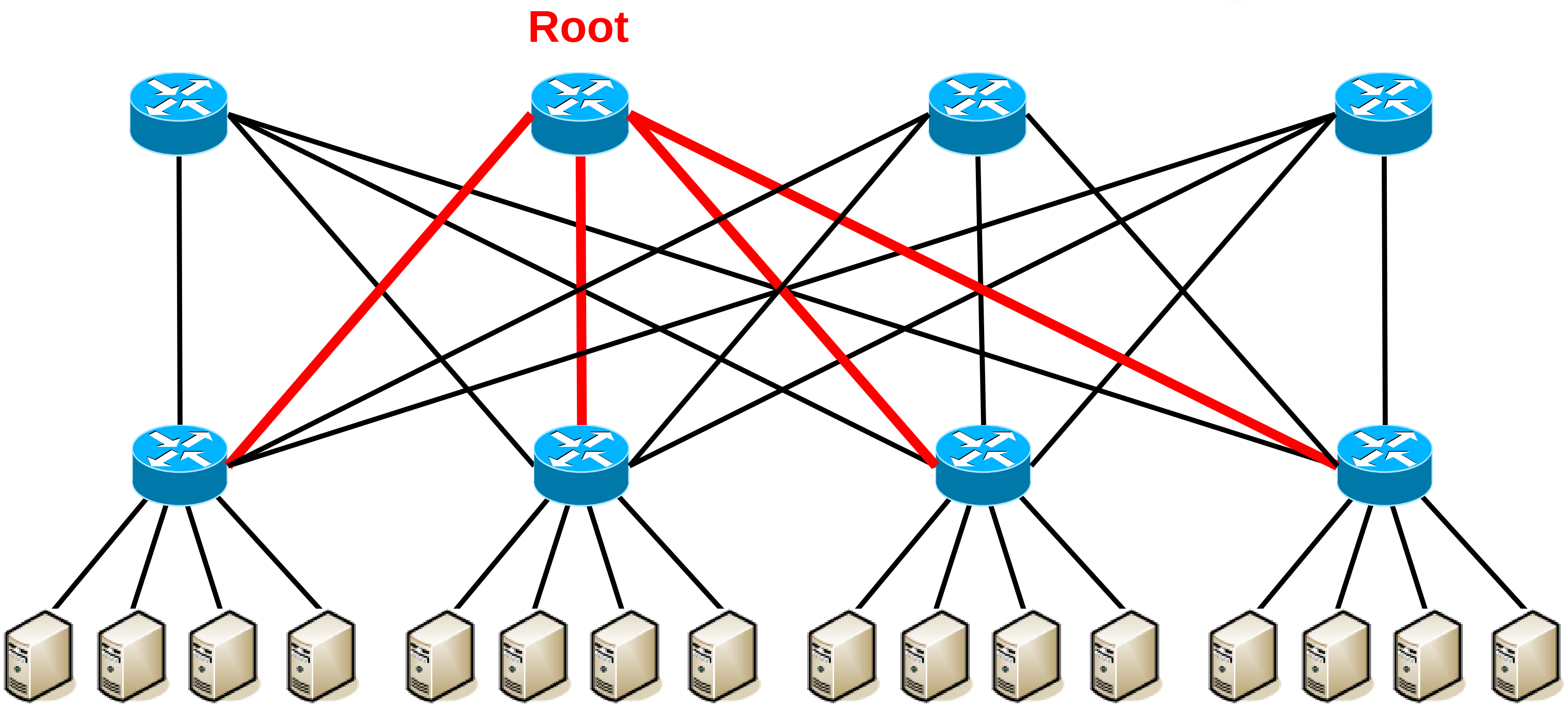
- Network-wide flooding:
  - Large number of control messages while disseminating host information
- Large forwarding tables:
  - Switches have to maintain large forwarding tables
    - Forwarding table size is proportional to the number of hosts due to flat addressing
- Broadcast traffic:
  - ARP and DHCP broadcasts consume bandwidth and processing resources at hosts and switches
- Spanning tree:
  - Waste of bandwidth at unused links
  - Load imbalance

Spanning Trees



- Spanning-tree inefficiencies:
  - Limited bandwidth utilization, since many links are not used
  - Requires very expensive core switches with large switching capacity

Spanning Trees



- Spanning-tree inefficiencies:
  - Limited bandwidth utilization, since many links are not used
  - Requires very expensive core switches with large switching capacity

Limitations of Multiple IP Subnets

- Configuration overhead:
  - Administrators should assign manually prefixes to subnets
  - Host IP address assignment via DCHP should be consistent with prefix assignment
  - Network topology changes require manual reconfiguration

- Limited mobility support:
  - Host mobility is restricted within a subnet where the host can maintain its IP address
  - Mobility across subnets requires reconfiguration (e.g., IP address reassignment) and may cause service disruption

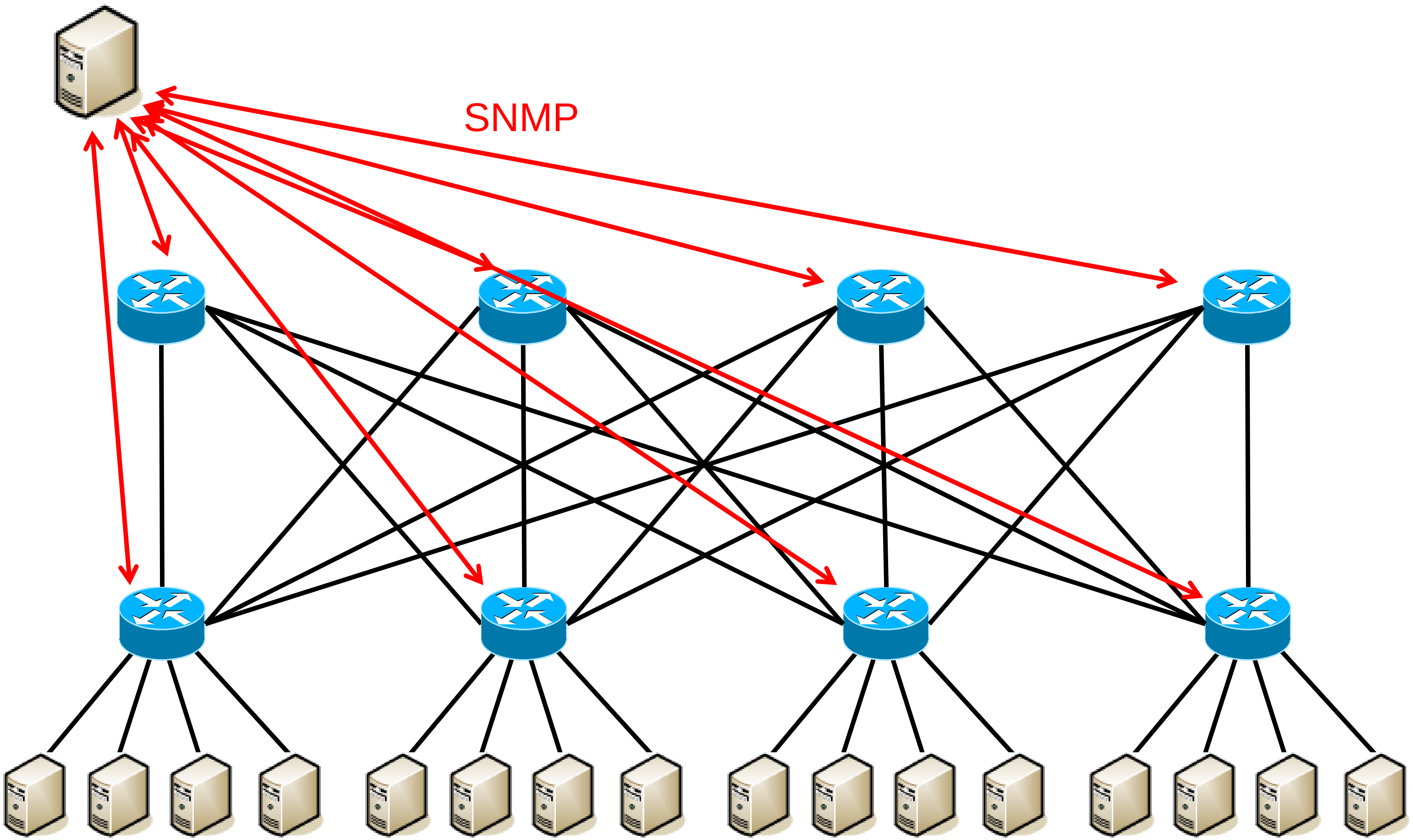
Smart Path Assignment in Networks (SPAIN)

Multi-Path Routing with SPAIN

- Goal:
  - Provide multi-path routing using VLANs
  - Should work on arbitrary DC network topologies

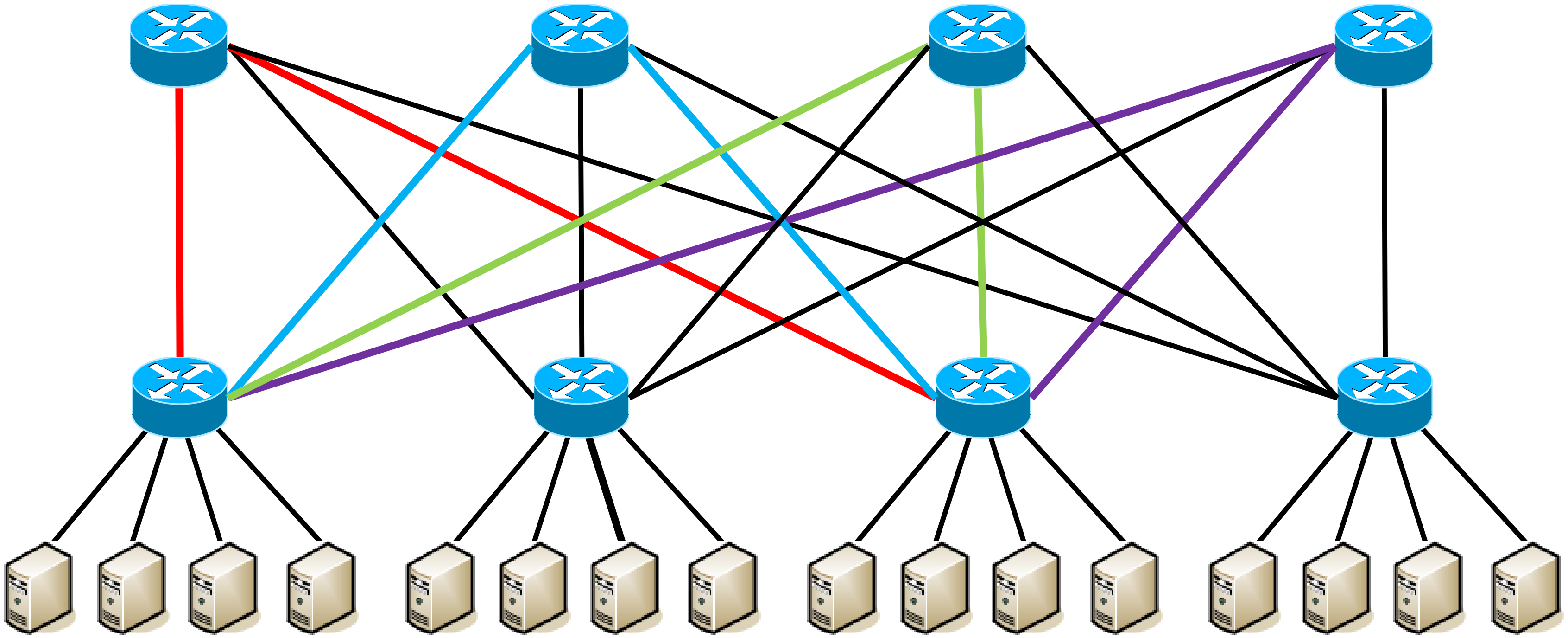
- Offline computation of the network:
  - Topology discovery
  - Path computation
  - Assignment of paths to VLANs

Topology Discovery



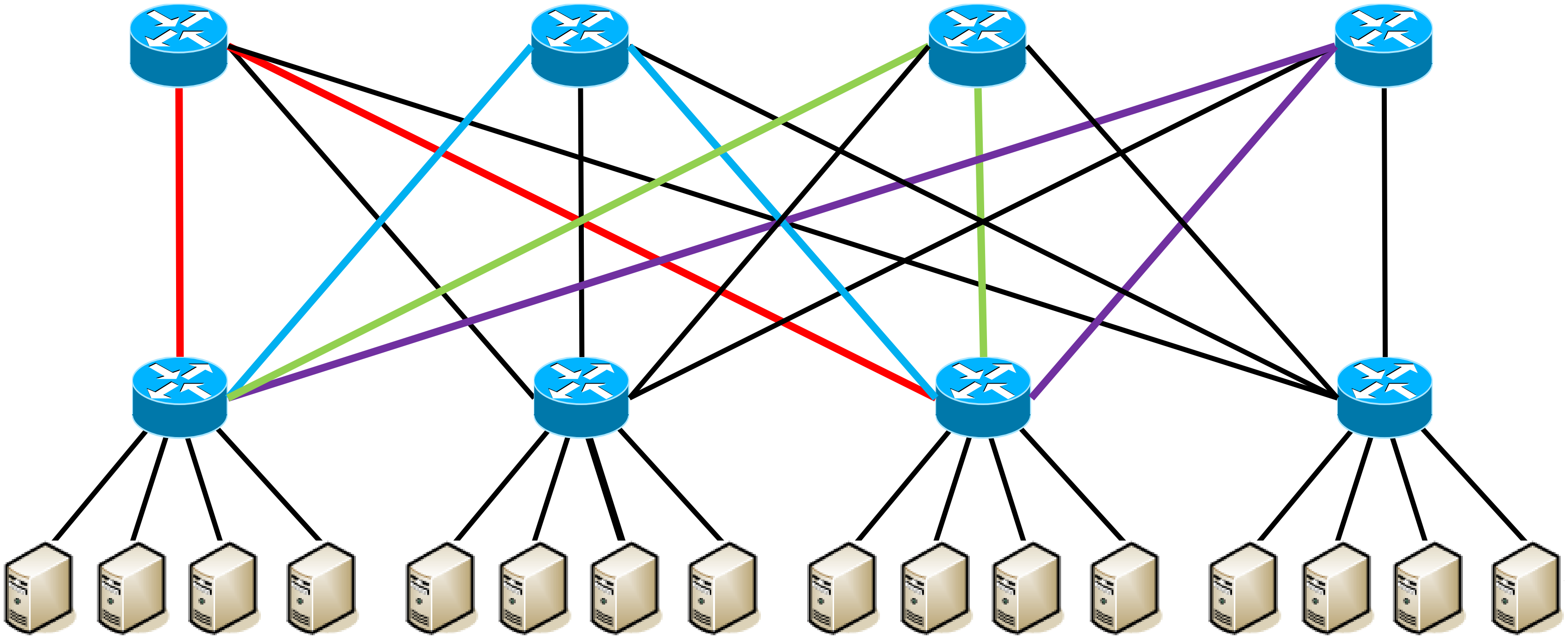
Path Computation

- Compute the smallest set of paths that exploit all redundancy
- Consider only paths between edge switches



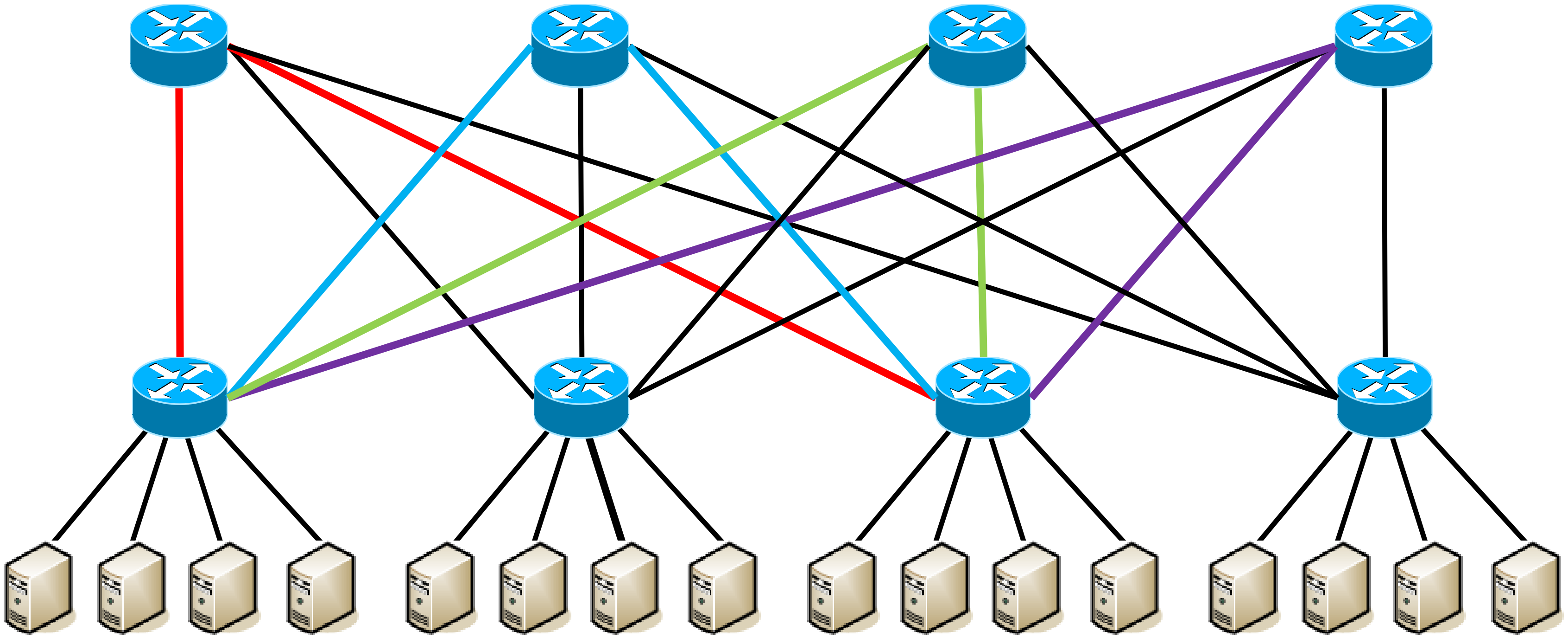
Path to VLAN Assignment (1)

- Simple assignment:
  - Each path is assigned to a separate VLAN
  - Limited by the maximum number of VLANs (4096)
  - Scales only to a small number of switches



Path to VLAN Assignment (2)

- Assignment proposed by SPAIN:
  - 1 VLAN is used for a set of paths
  - Greedy VLAN packing algorithm for optimizing path assignment



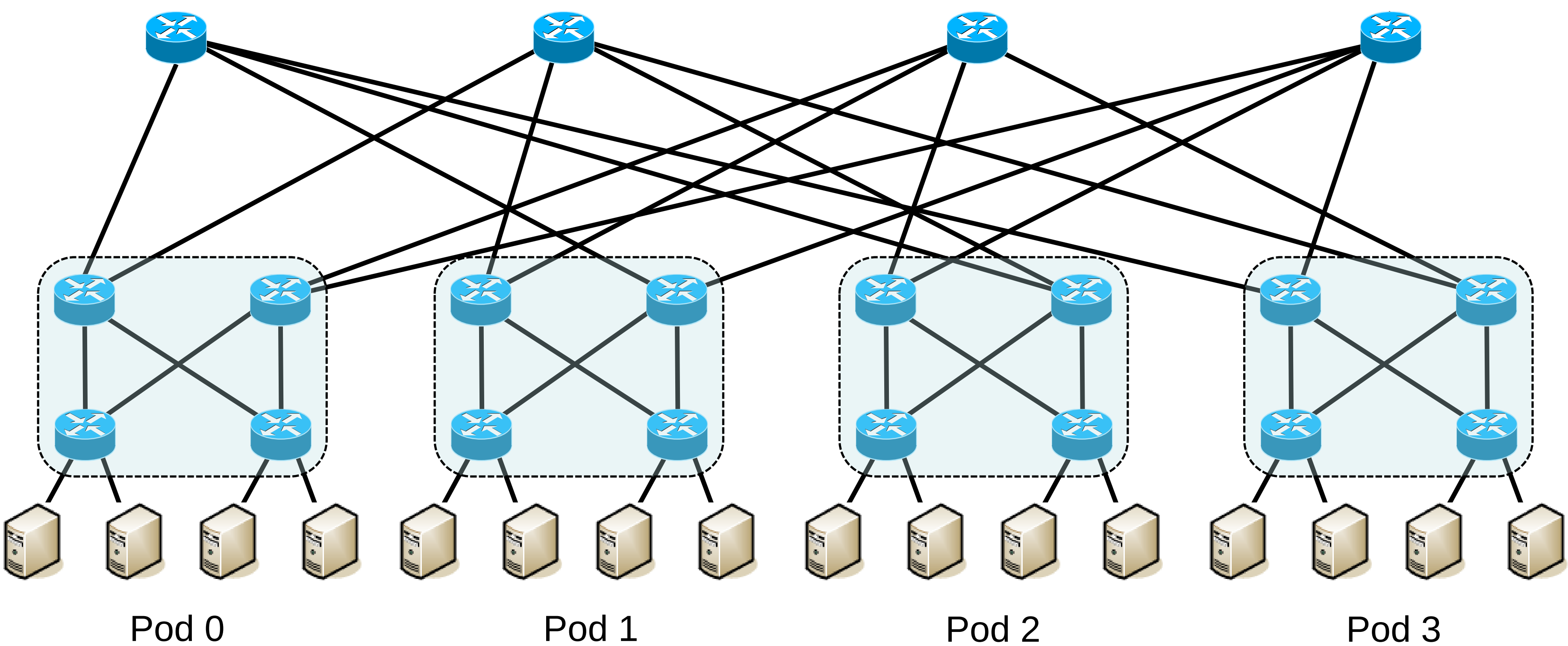
PortLand

Main Features

- PortLand is a single logical layer-2 data center network fabric that scales to millions of (virtual) end-points
- PortLand internally separates host identity from host location:
  - uses IP address as host identifier
  - introduces “Pseudo MAC” (PMAC) addresses internally to encode endpoint location
- PortLand runs on commodity switch hardware with unmodified hosts

Fat-Tree Topology

- PortLand assumes hierarchical structure of data center networks:
  - e.g., fat-tree topology (multi-rooted tree)

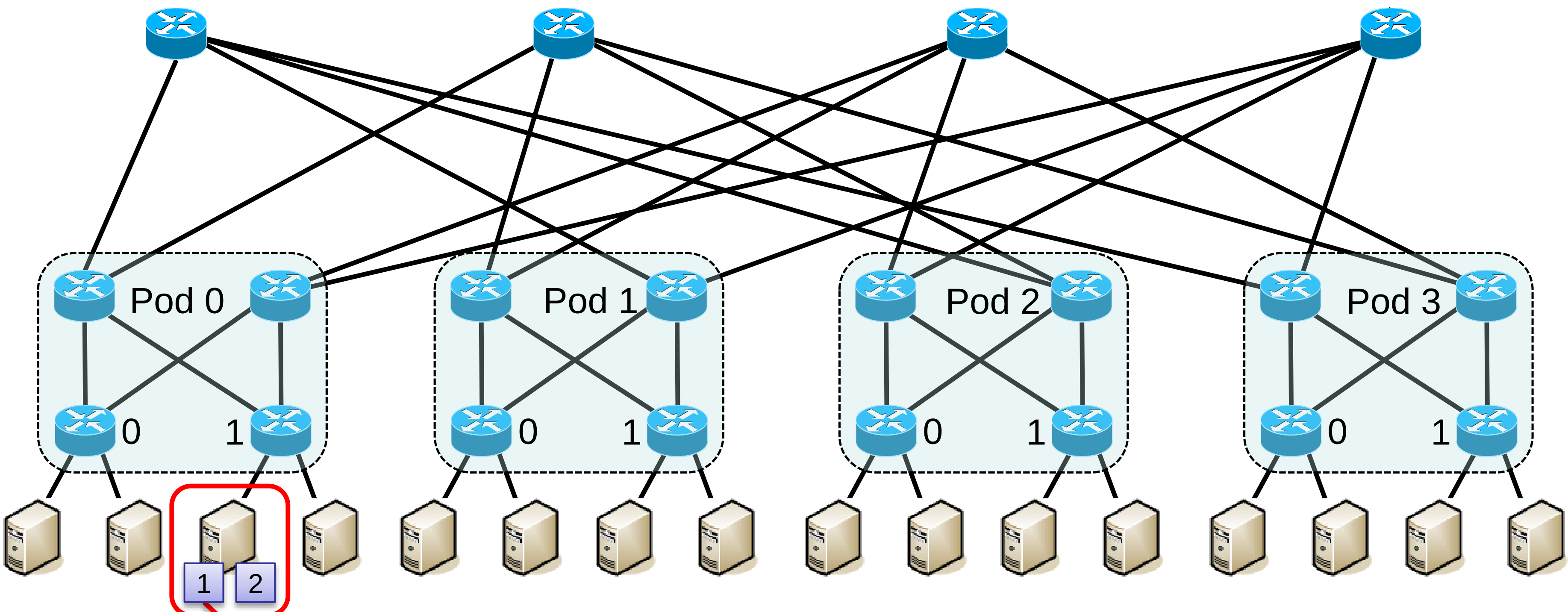


Pseudo MAC Addressing

- PMAC addressing for packet forwarding and routing:
  - Besides IP and MAC, each end-host is assigned with a unique PMAC address
  - PMAC encodes the location of each end-host

- PMAC address format: `pod.position.port.vmid`
  - `pod`: pod number of the edge switch (16 bits)
  - `position`: position of the edge switch within the pod (8 bits)
  - `port`: switch port number (8 bits)
  - `vmid`: virtual machine ID for demultiplexing (16 bits)

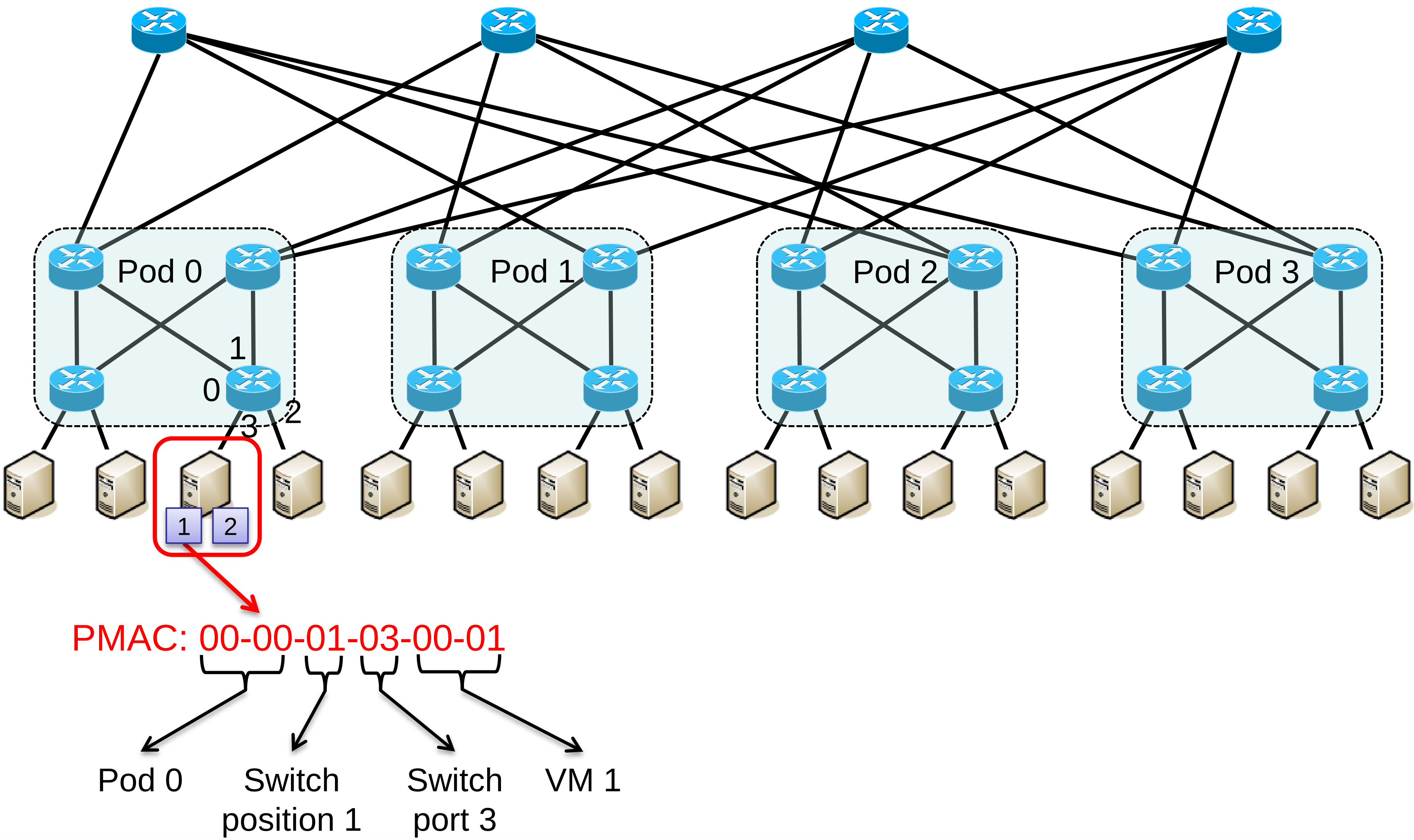
PMAC Example



PMAC: $\underbrace{00\text{-}00}_{\text{Pod 0}}\text{-}\underbrace{01}_{\text{Switch position 1}}\text{-}\underbrace{03}_{\text{Switch port 3}}\text{-}\underbrace{00\text{-}01}_{\text{VM 1}}$

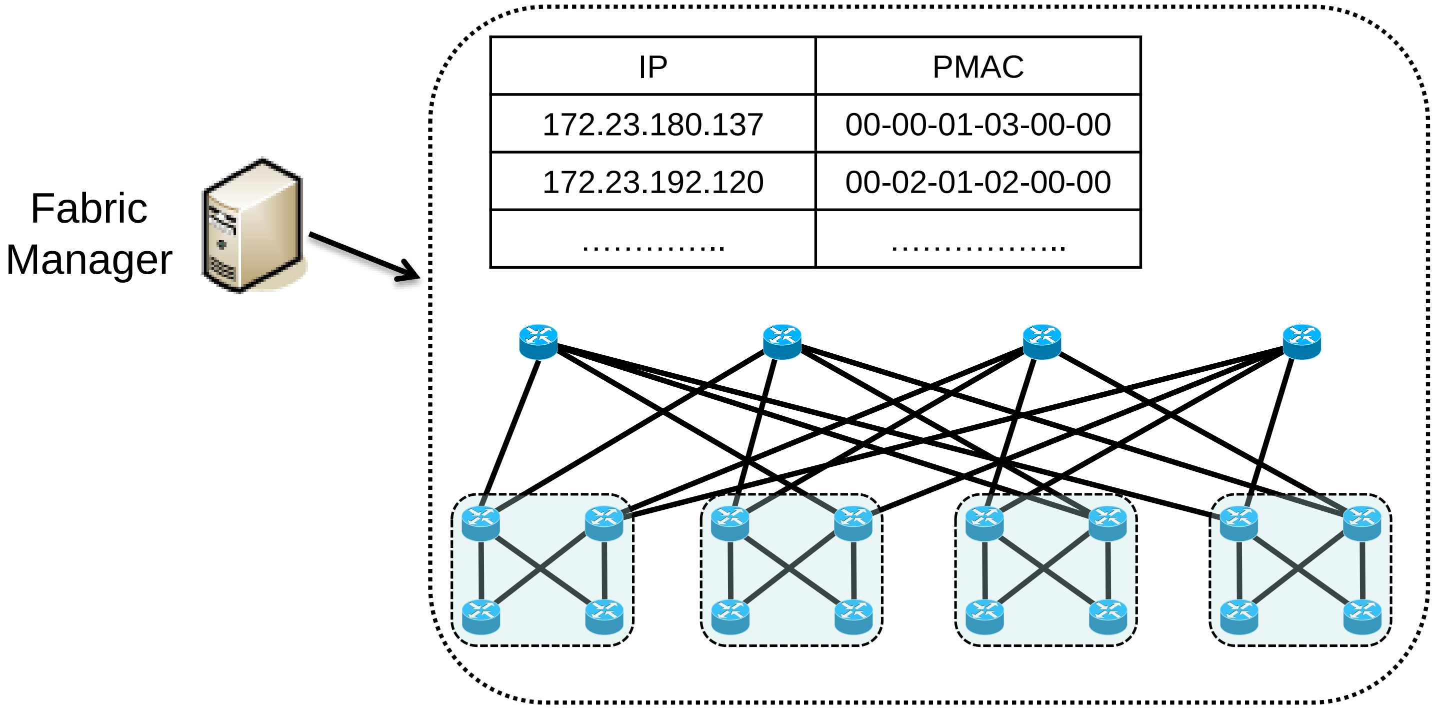
PMAC Example

Institut für Kommunikations-Technik



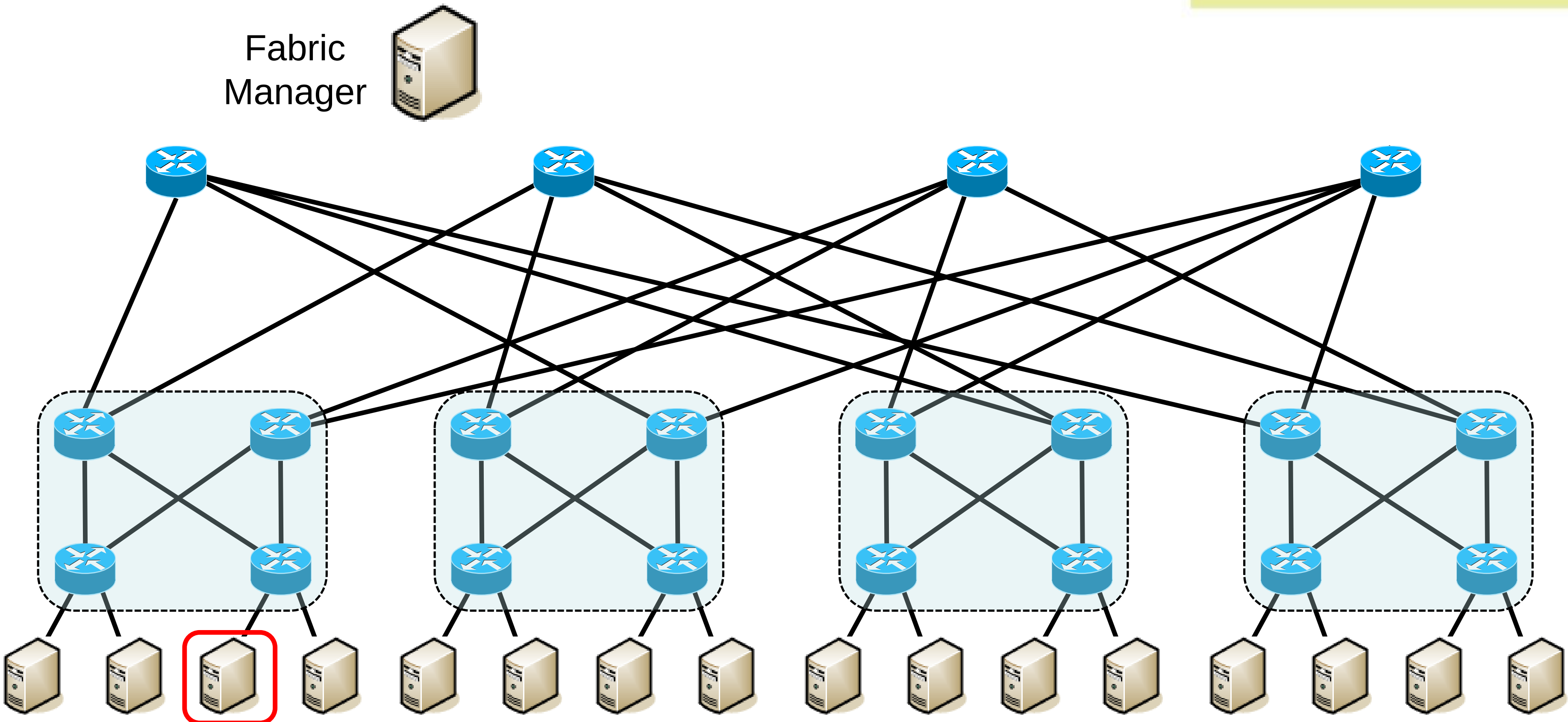
Fabric Manager

- Centralized Fabric Manager:
  - maintains (IP, PMAC) bindings, assisting ARP resolution
  - maintains the switch-level topology, facilitating fault-tolerant routing
  - maintains soft state, eliminating the need for manual configuration



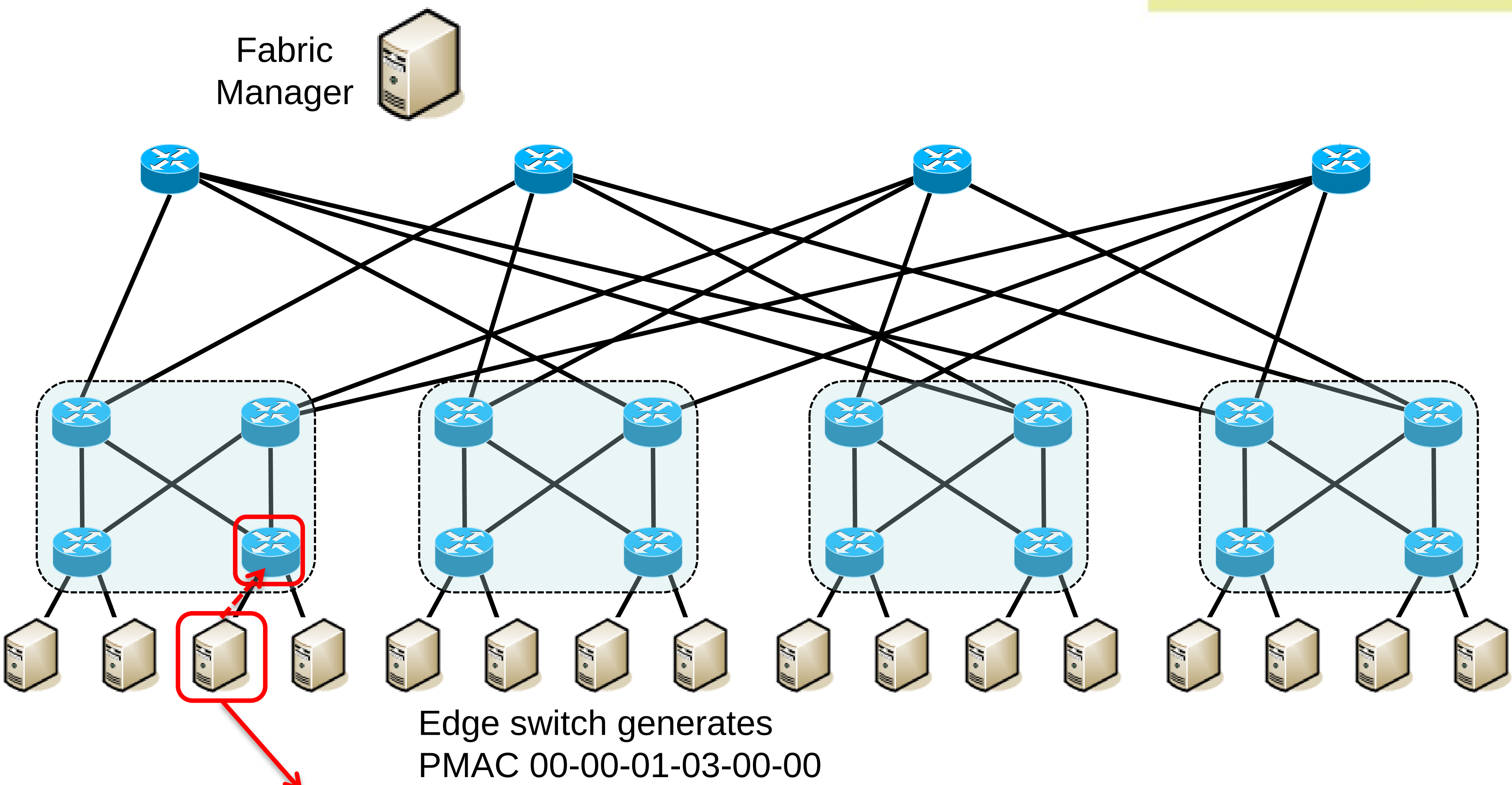
IP	PMAC
172.23.180.137	00-00-01-03-00-00
172.23.192.120	00-02-01-02-00-00
…………..	……………..

PMAC Assignment



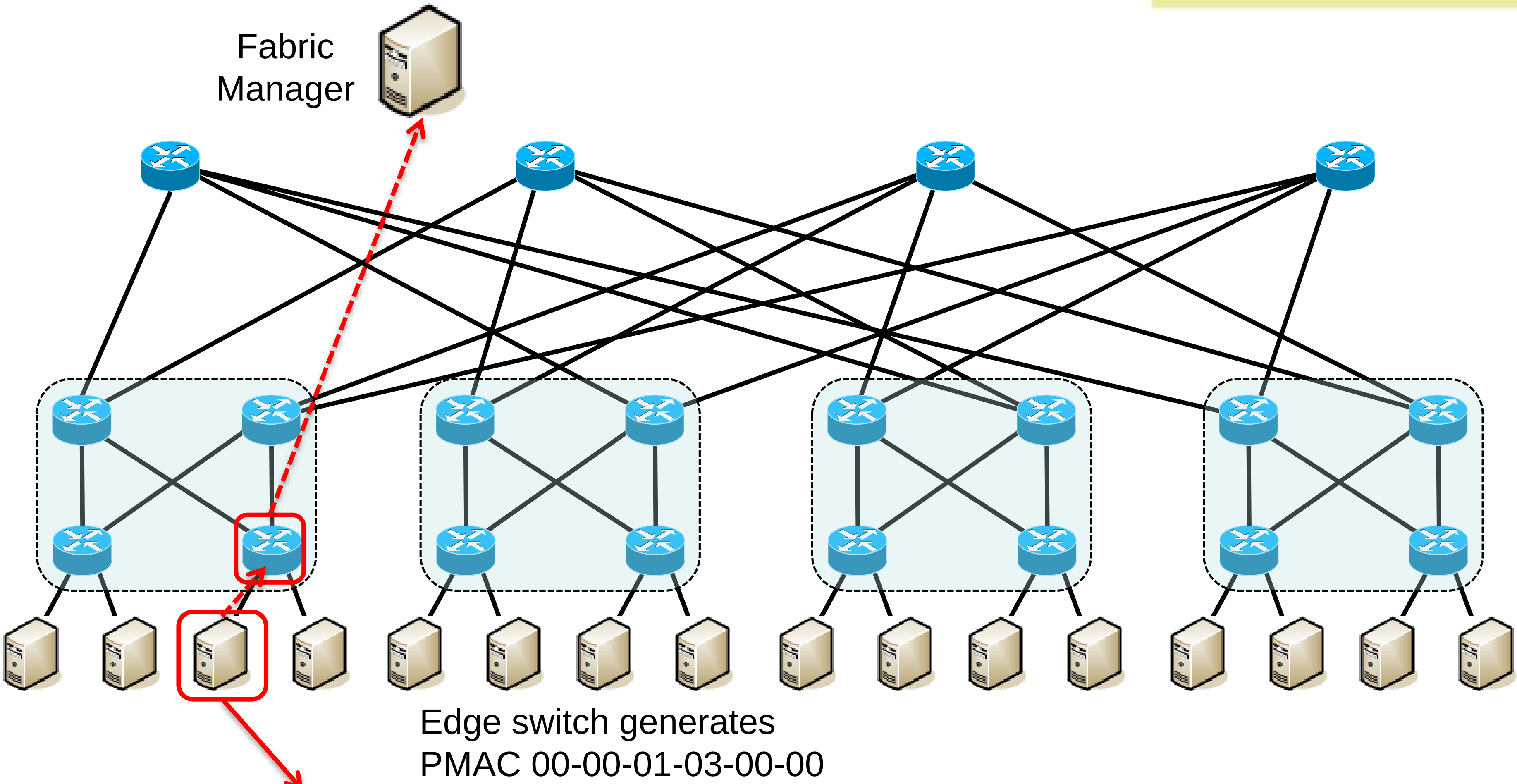
IP	MAC
172.23.180.137	84-2B-2B-A5-A5-77

PMAC Assignment



IP	MAC
172.23.180.137	84-2B-2B-A5-A5-77

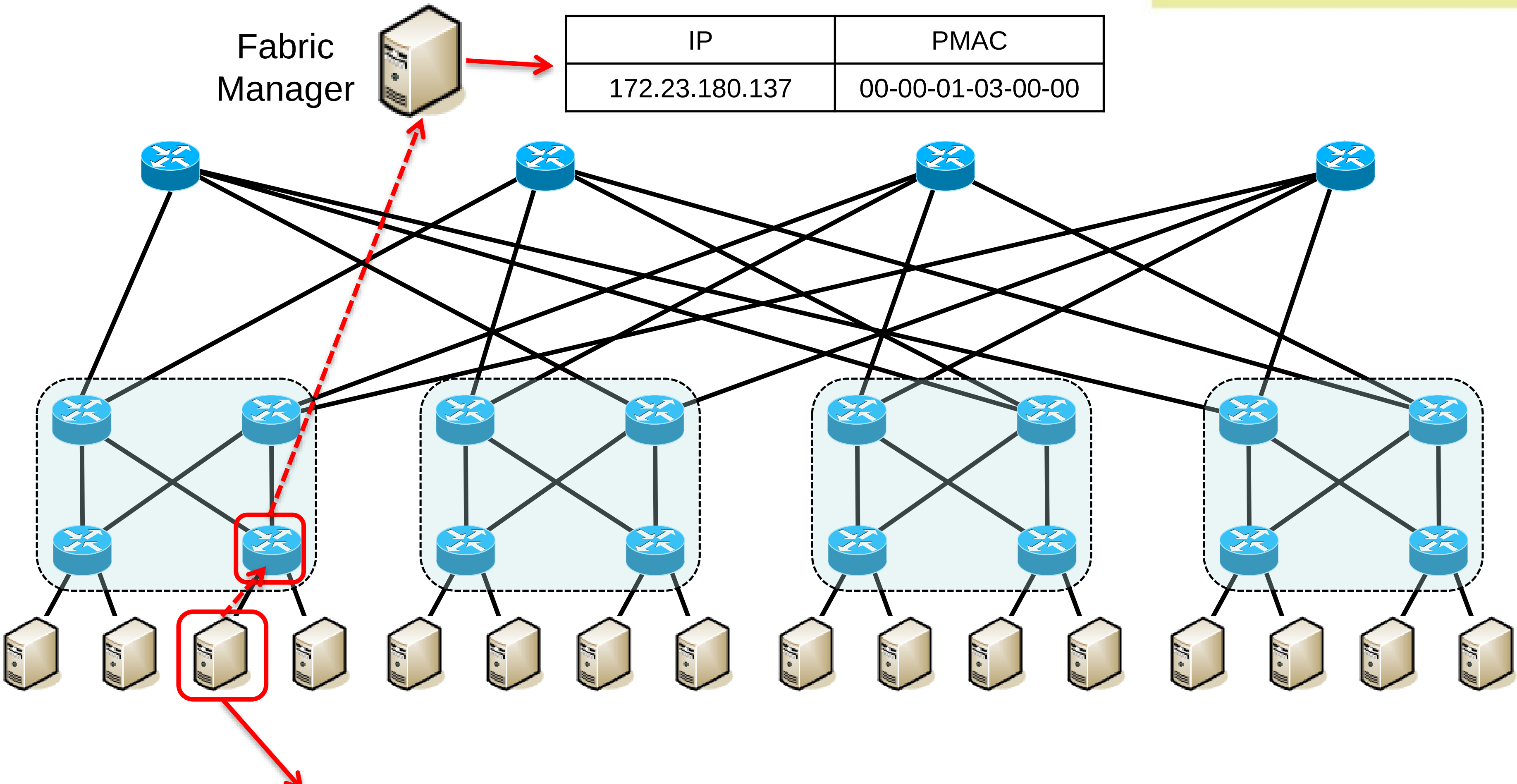
PMAC Assignment



Edge switch generates
PMAC 00-00-01-03-00-00

IP	MAC
172.23.180.137	84-2B-2B-A5-A5-77

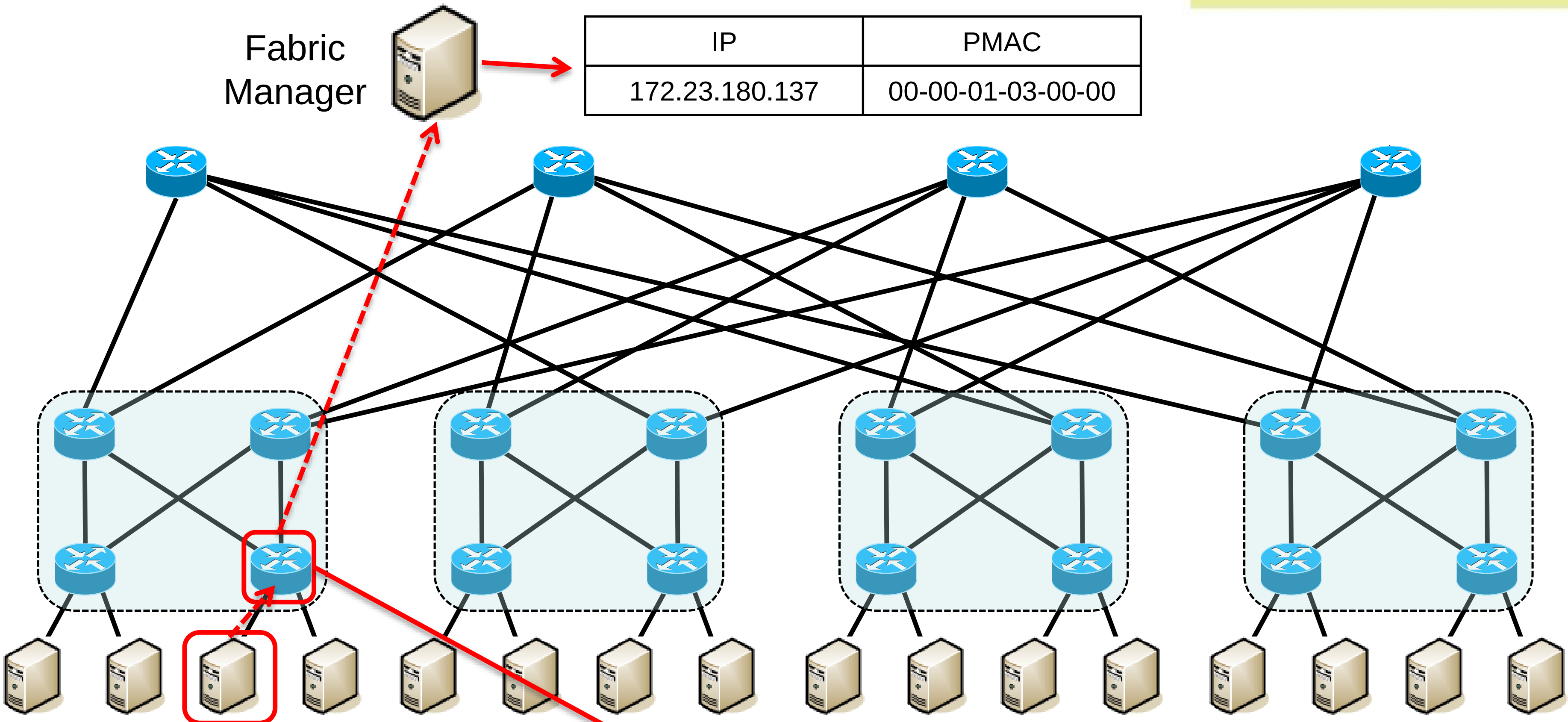
PMAC Assignment



IP	PMAC
172.23.180.137	00-00-01-03-00-00

IP	MAC
172.23.180.137	84-2B-2B-A5-A5-77

PMAC Assignment



IP	PMAC
172.23.180.137	00-00-01-03-00-00

IP	MAC
172.23.180.137	84-2B-2B-A5-A5-77

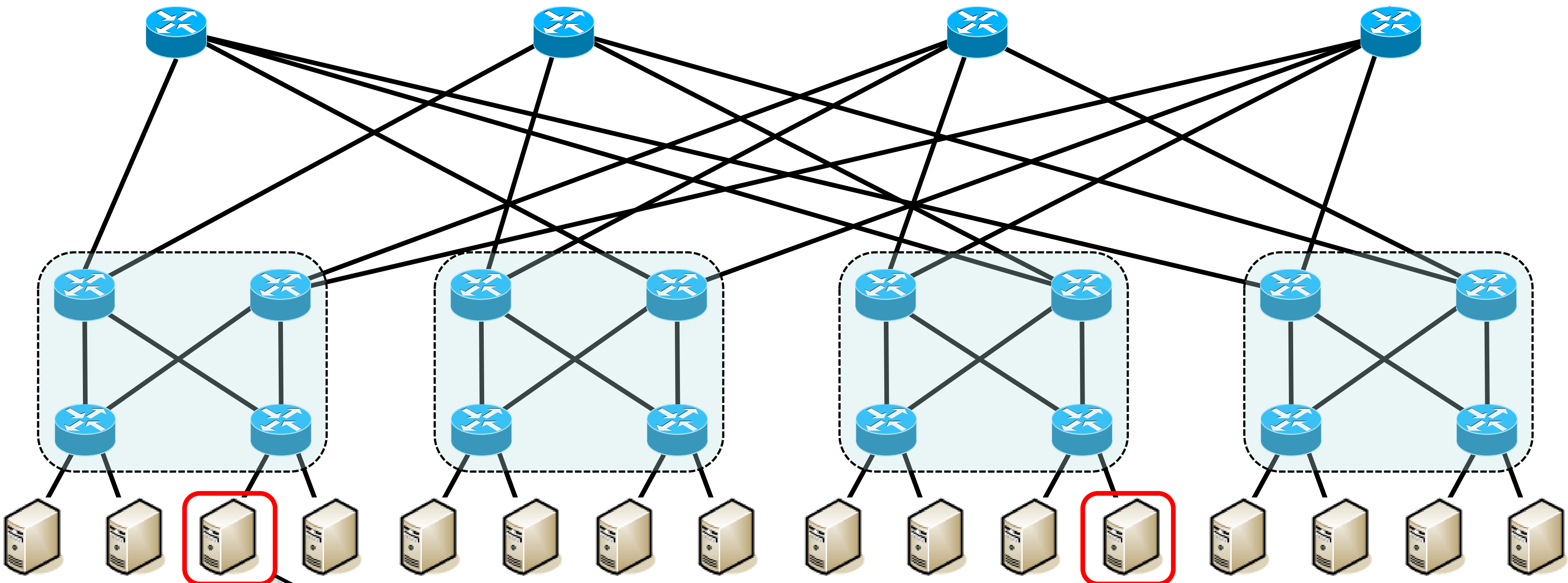
IP	MAC	PMAC
172.23.180.137	84-2B-2B-A5-A5-77	00-00-01-03-00-00

Proxy-based Address Resolution

Institut für Kommunikations-Technik

Fabric Manager

IP	PMAC
172.23.180.137	00-00-01-03-00-00
172.23.192.120	00-02-01-02-00-00



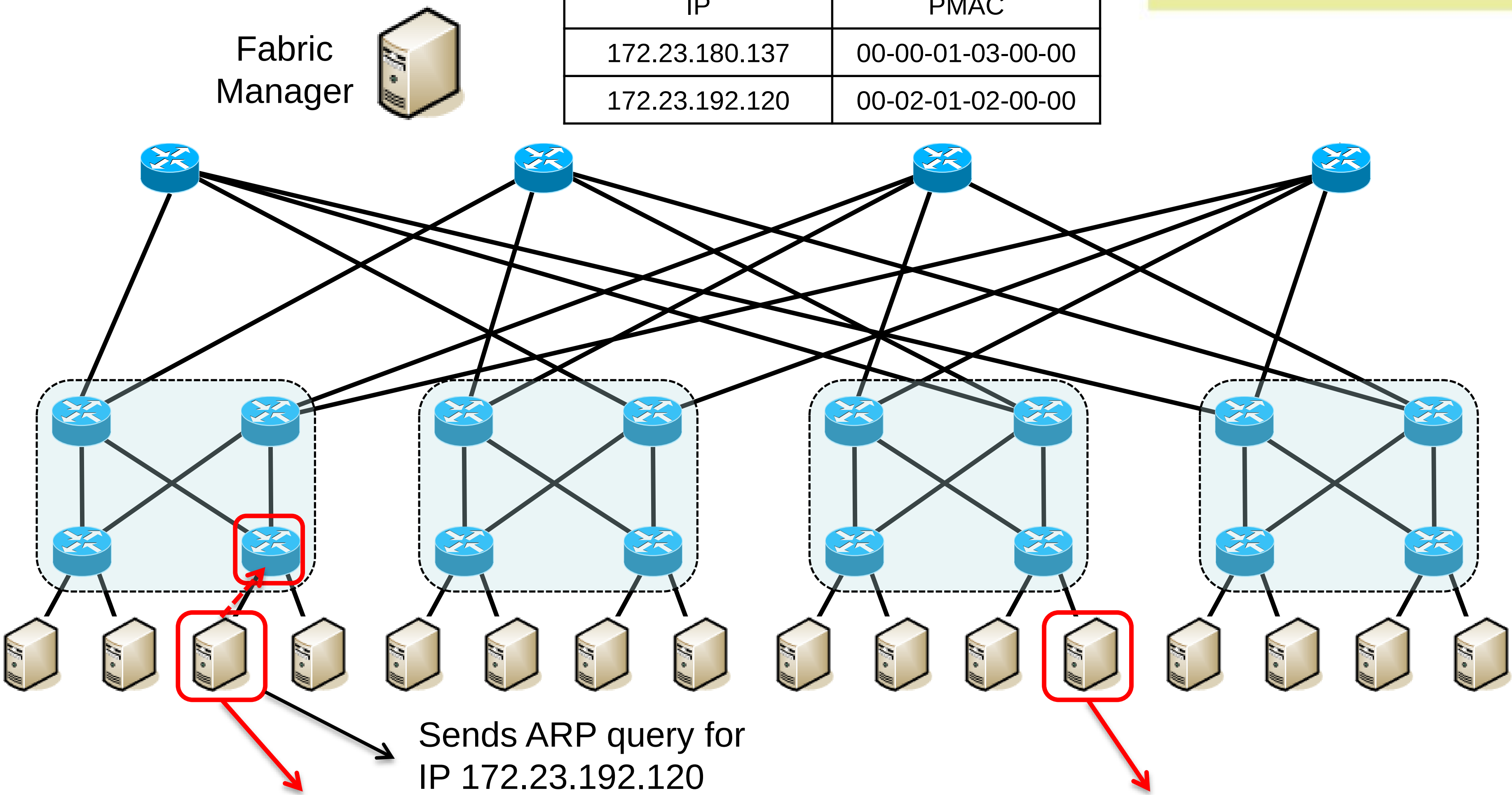
Wants to forward packets to IP 172.23.192.120. PMAC?

IP	MAC
172.23.180.137	84-2B-2B-A5-A5-77

IP	MAC
172.23.192.120	A3-B4-21-87-D4-12

Proxy-based Address Resolution

IP	PMAC
172.23.180.137	00-00-01-03-00-00
172.23.192.120	00-02-01-02-00-00



Sends ARP query for
IP 172.23.192.120

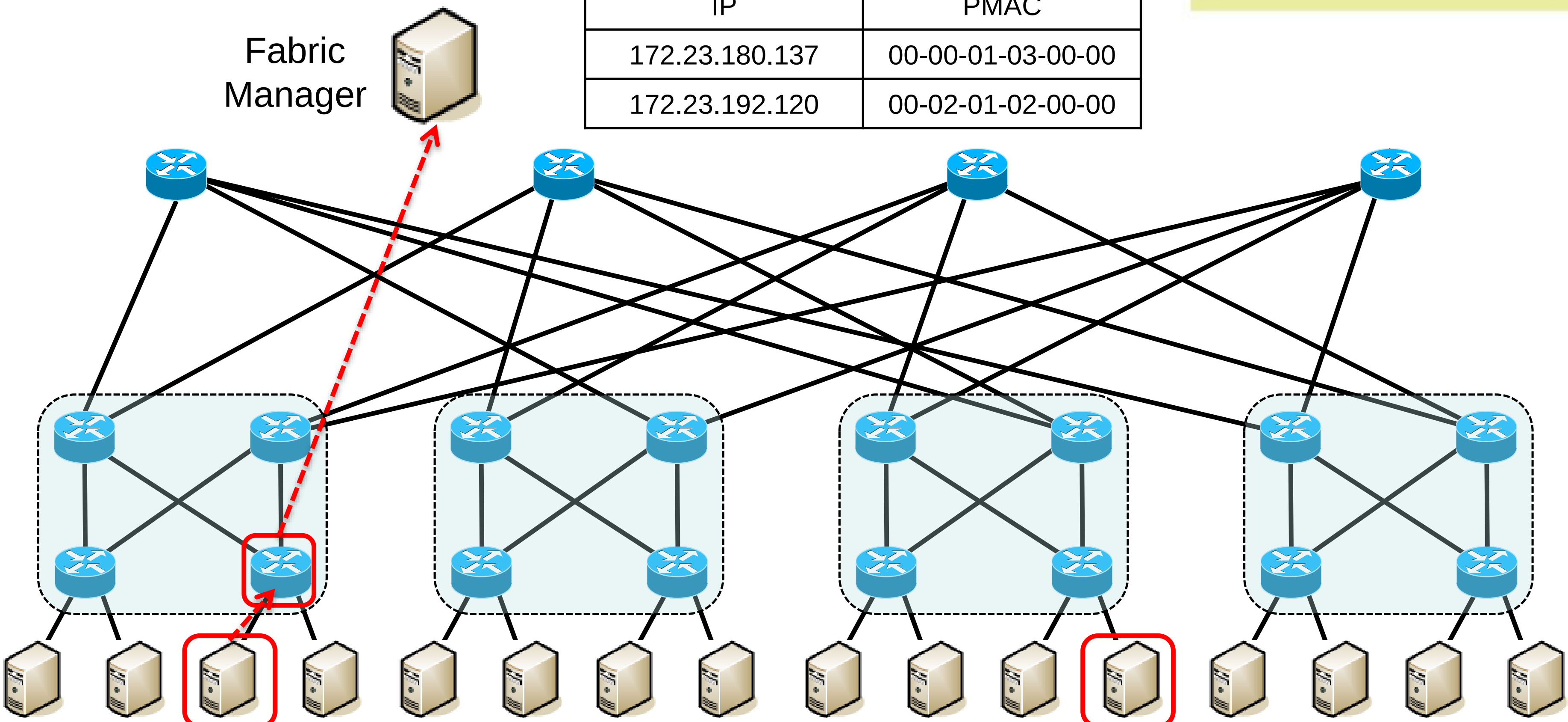
IP	MAC
172.23.180.137	84-2B-2B-A5-A5-77

IP	MAC
172.23.192.120	A3-B4-21-87-D4-12

Proxy-based Address Resolution

Fabric Manager

IP	PMAC
172.23.180.137	00-00-01-03-00-00
172.23.192.120	00-02-01-02-00-00



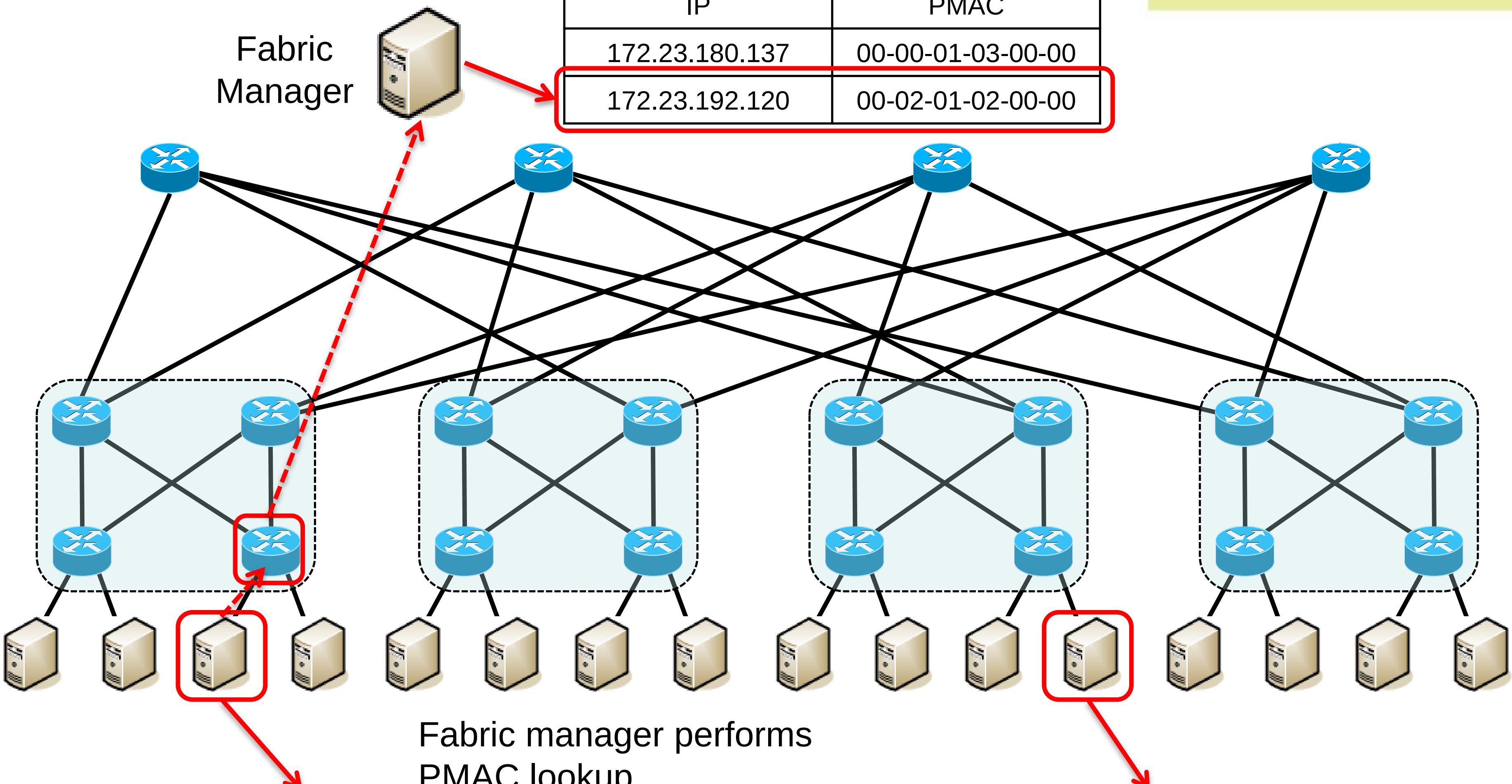
Edge switch intercepts ARP query and initiates PMAC resolution

IP	MAC
172.23.180.137	84-2B-2B-A5-A5-77

IP	MAC
172.23.192.120	A3-B4-21-87-D4-12

Proxy-based Address Resolution

IP	PMAC
172.23.180.137	00-00-01-03-00-00
172.23.192.120	00-02-01-02-00-00



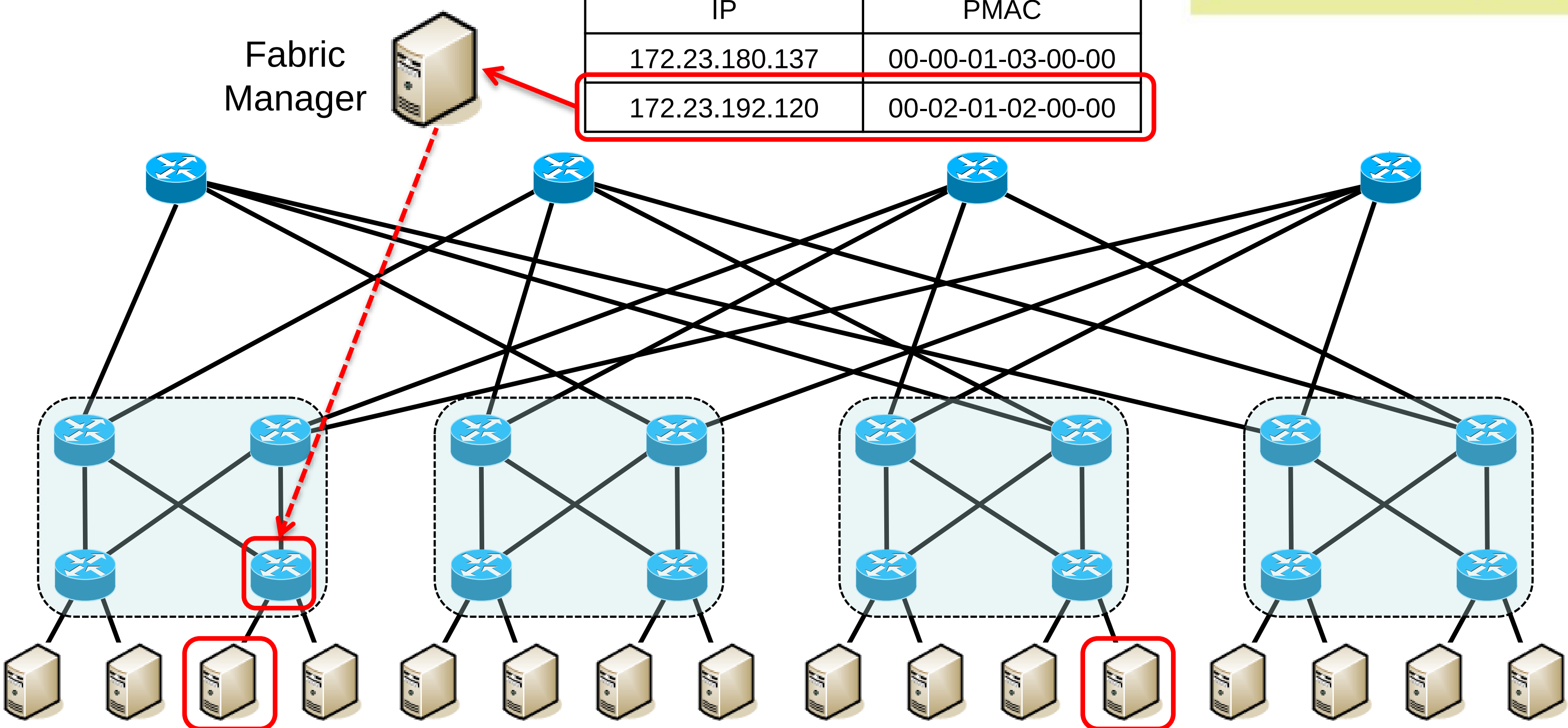
Fabric manager performs PMAC lookup

IP	MAC
172.23.180.137	84-2B-2B-A5-A5-77

IP	MAC
172.23.192.120	A3-B4-21-87-D4-12

Proxy-based Address Resolution

IP	PMAC
172.23.180.137	00-00-01-03-00-00
172.23.192.120	00-02-01-02-00-00



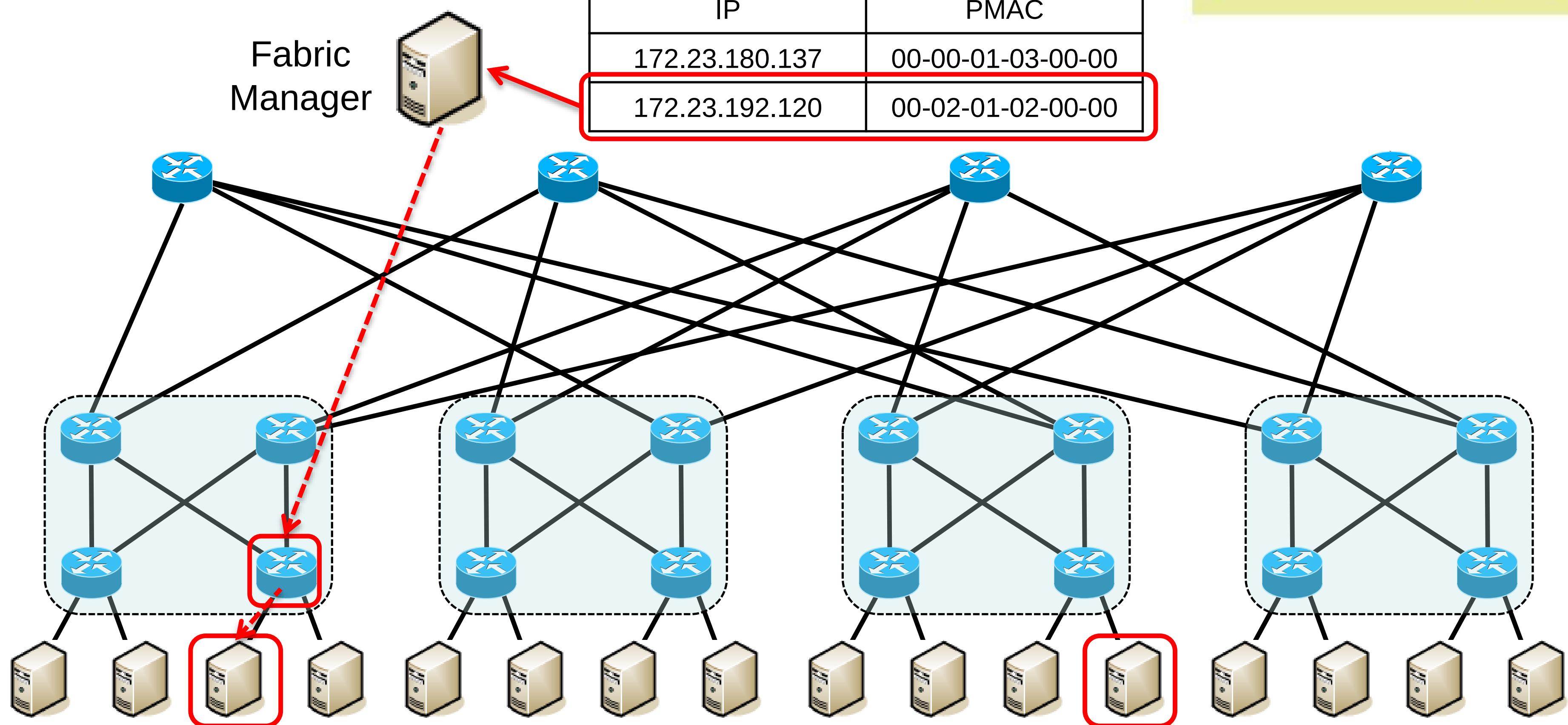
Fabric manager forwards the (IP, PMAC) binding to the edge switch

IP	MAC
172.23.180.137	84-2B-2B-A5-A5-77

IP	MAC
172.23.192.120	A3-B4-21-87-D4-12

Proxy-based Address Resolution

IP	PMAC
172.23.180.137	00-00-01-03-00-00
172.23.192.120	00-02-01-02-00-00

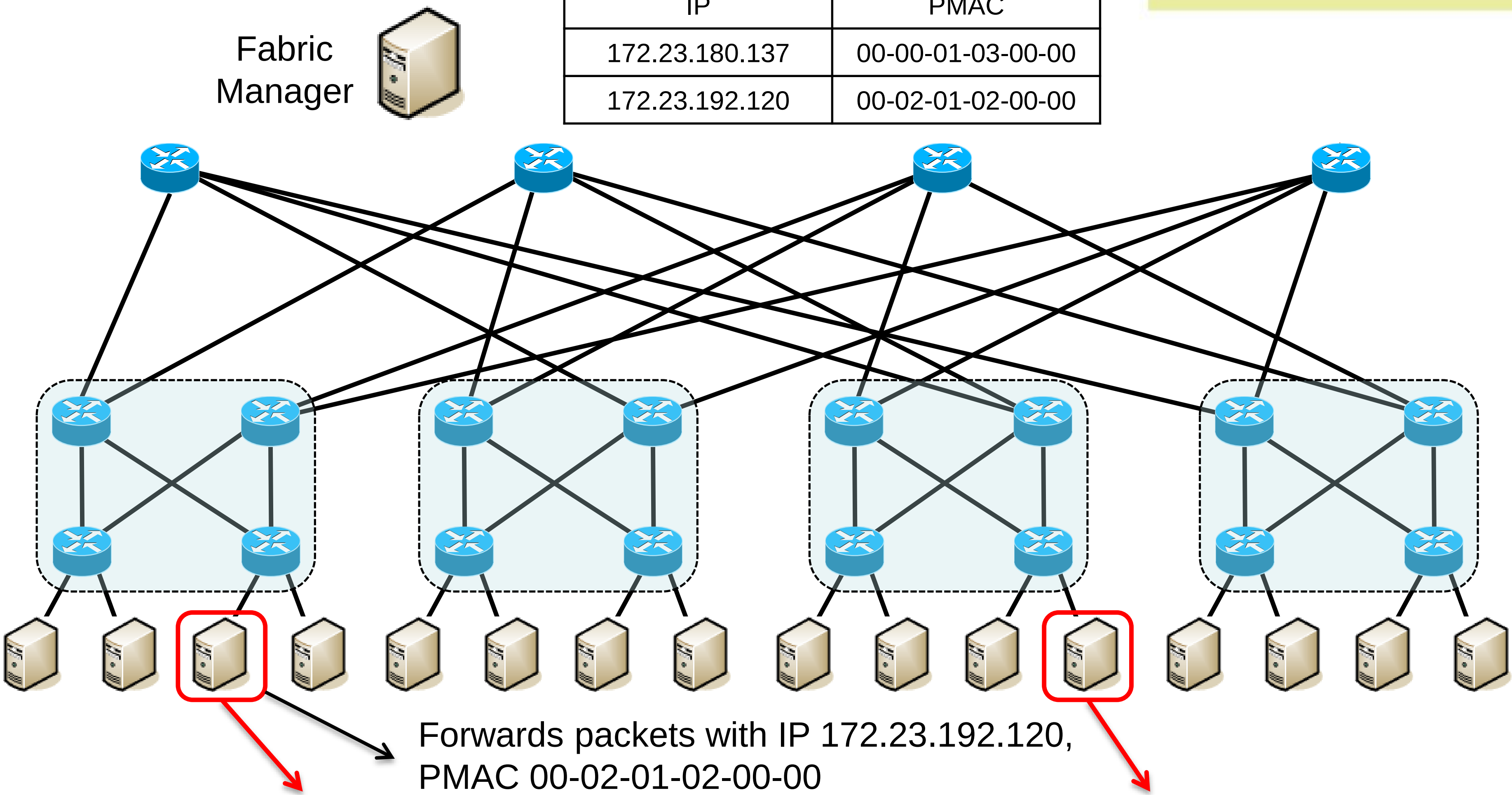


Edge switch returns PMAC to the end-host

IP	MAC
172.23.180.137	84-2B-2B-A5-A5-77

IP	MAC
172.23.192.120	A3-B4-21-87-D4-12

Packet Forwarding



IP	PMAC
172.23.180.137	00-00-01-03-00-00
172.23.192.120	00-02-01-02-00-00

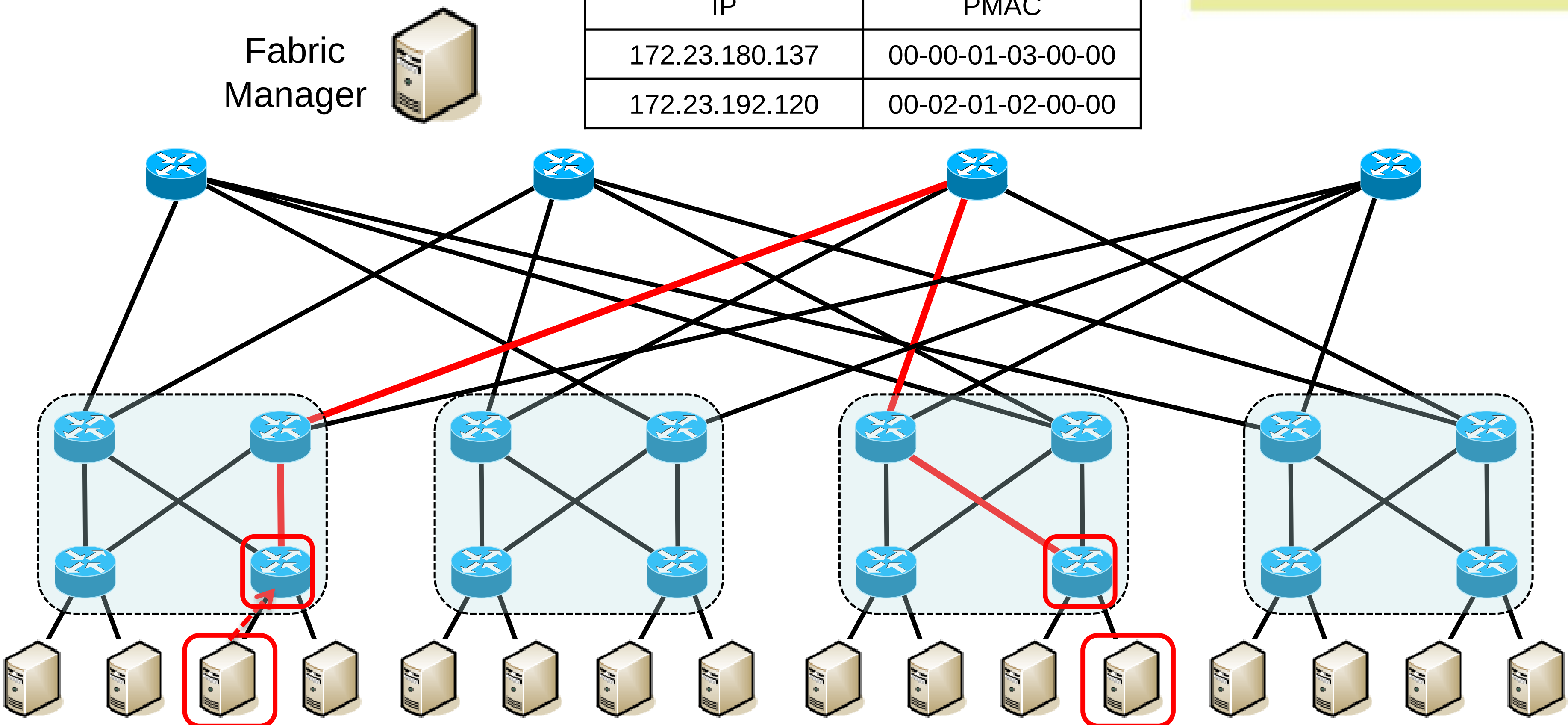
Forwards packets with IP 172.23.192.120, PMAC 00-02-01-02-00-00

IP	MAC
172.23.180.137	84-2B-2B-A5-A5-77

IP	MAC
172.23.192.120	A3-B4-21-87-D4-12

Packet Forwarding

IP	PMAC
172.23.180.137	00-00-01-03-00-00
172.23.192.120	00-02-01-02-00-00



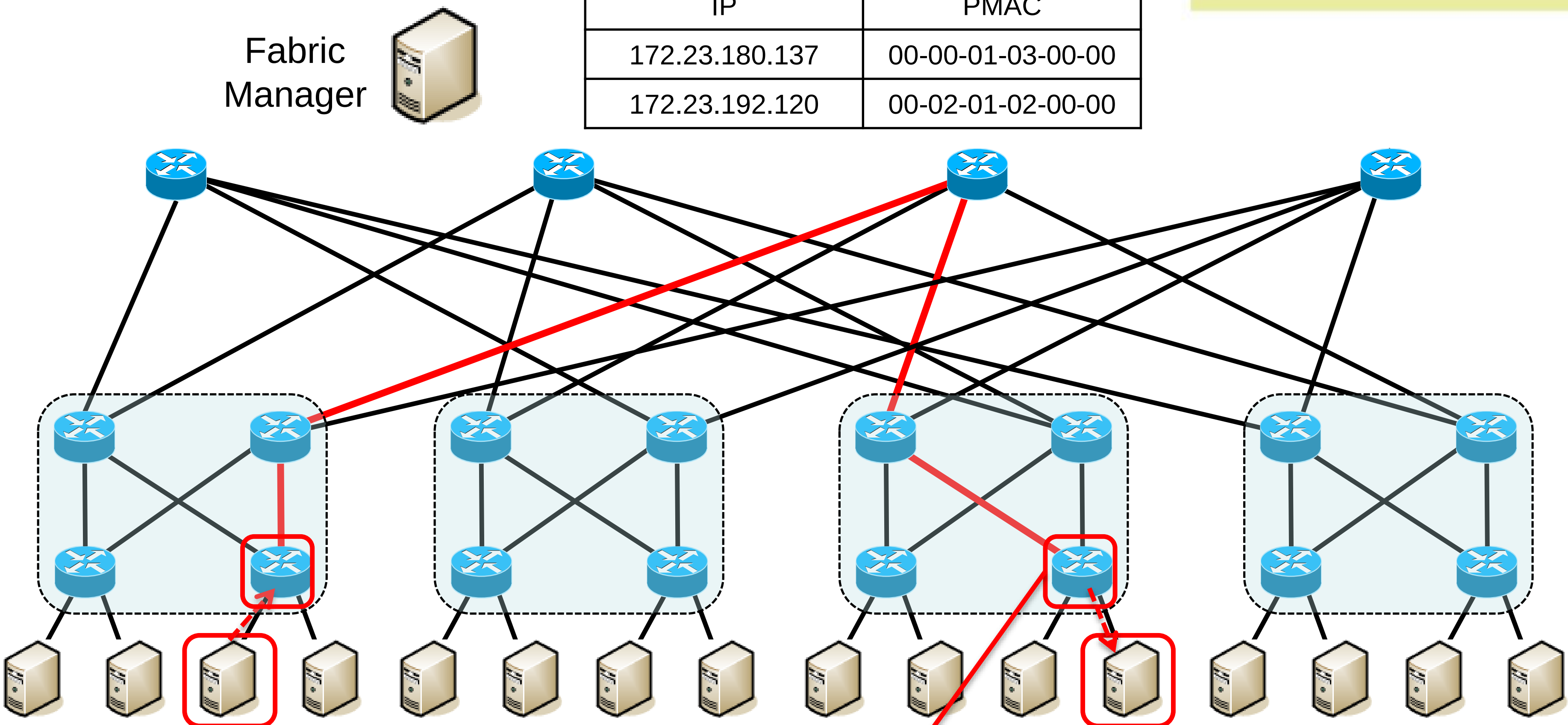
Edge switch routes packets to the destination edge switch based on PMAC

IP	MAC
172.23.180.137	84-2B-2B-A5-A5-77

IP	MAC
172.23.192.120	A3-B4-21-87-D4-12

Packet Forwarding

Institut für Kommunikations-Technik



IP	PMAC
172.23.180.137	00-00-01-03-00-00
172.23.192.120	00-02-01-02-00-00

Destination edge switch replaces PMAC with MAC and forwards packets to the end-host

IP	MAC	PMAC
172.23.192.120	A3-B4-21-87-D4-12	00-02-01-02-00-00

IP	MAC
172.23.192.120	A3-B4-21-87-D4-12

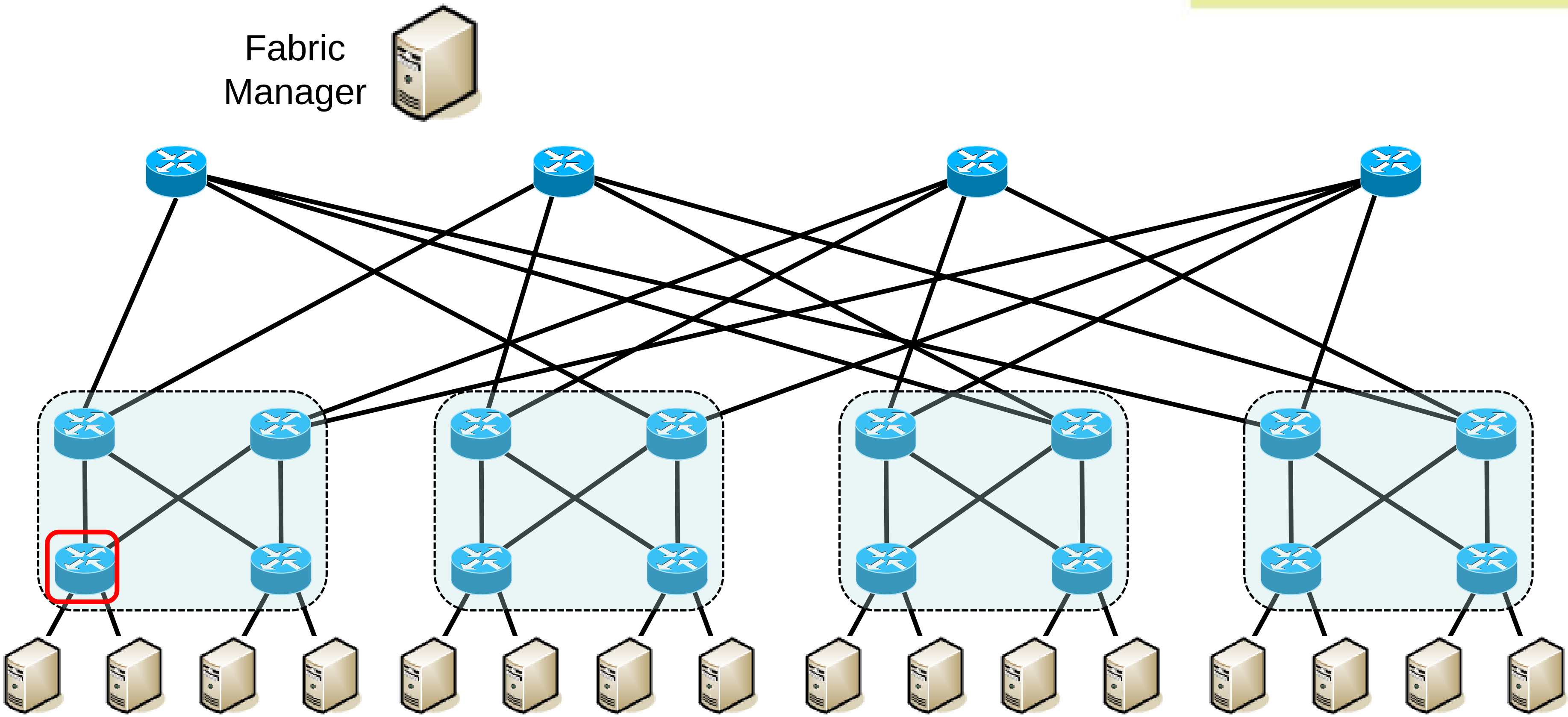
Distributed Location Discovery

- Location Discovery Messages (LDMs) are exchanged between neighboring switches
- Switches self-discover the following location information on boot:
  - Tree level / role:
    - Based on neighbor identity
  - Position number:
    - Aggregation switches assist edge switches in choosing a unique position number
  - Pod number:
    - Fabric manager assists switches in choosing pod number

Location Discovery Messages (LDM)

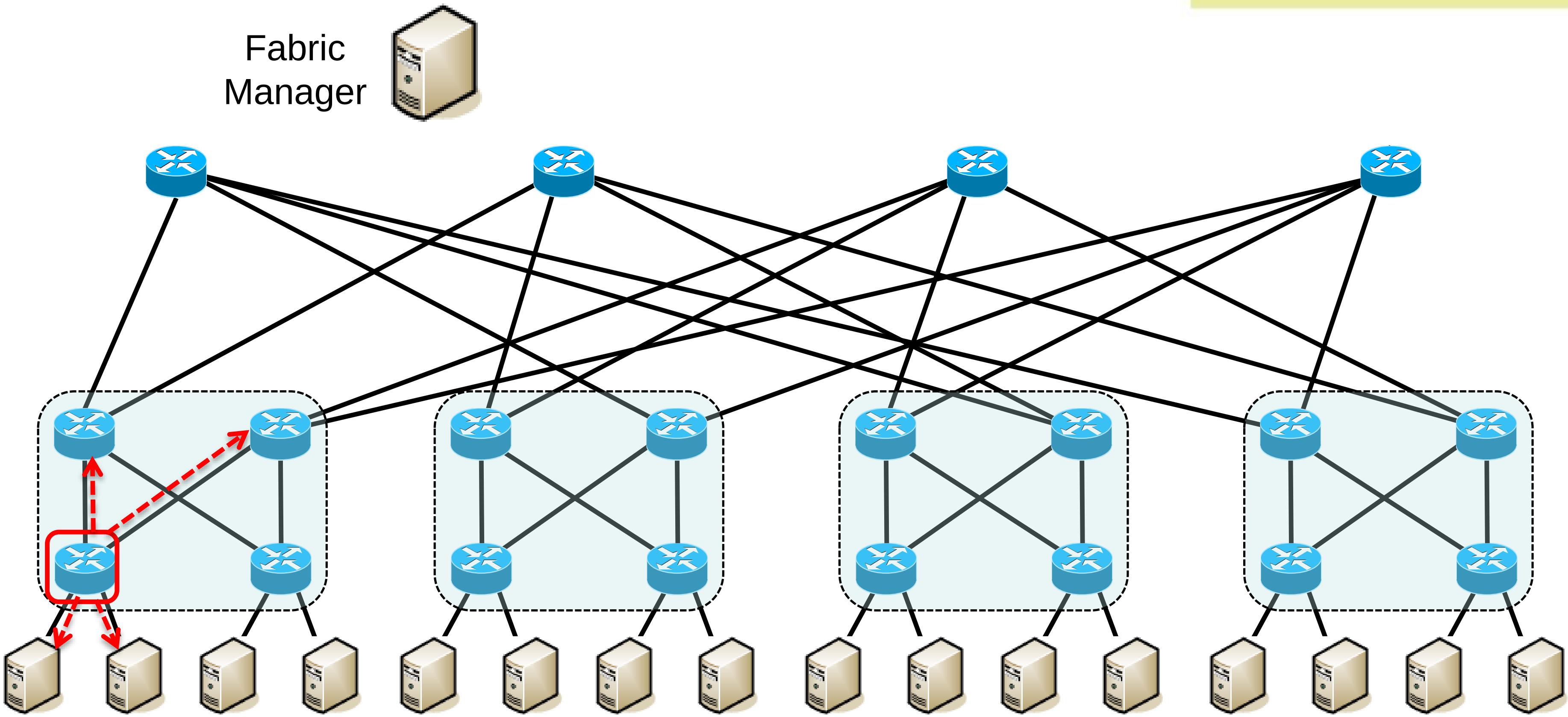
- LDMs include the following information:
  - Switch identifier:
    - Globally unique ID for each switch (e.g., the lowest MAC address of all local ports)
  - Pod number:
    - Unique pod number shared by all switches in the same pod
  - Position number:
    - Unique number for each switch in the same pod
  - Tree level:
    - Number that indicates whether a switch is an edge (0), aggregation (1) or core switch (2)
  - Up/down:
    - A bit that indicates whether a switch port is facing downward or upward in the fat tree

Location Discovery Protocol



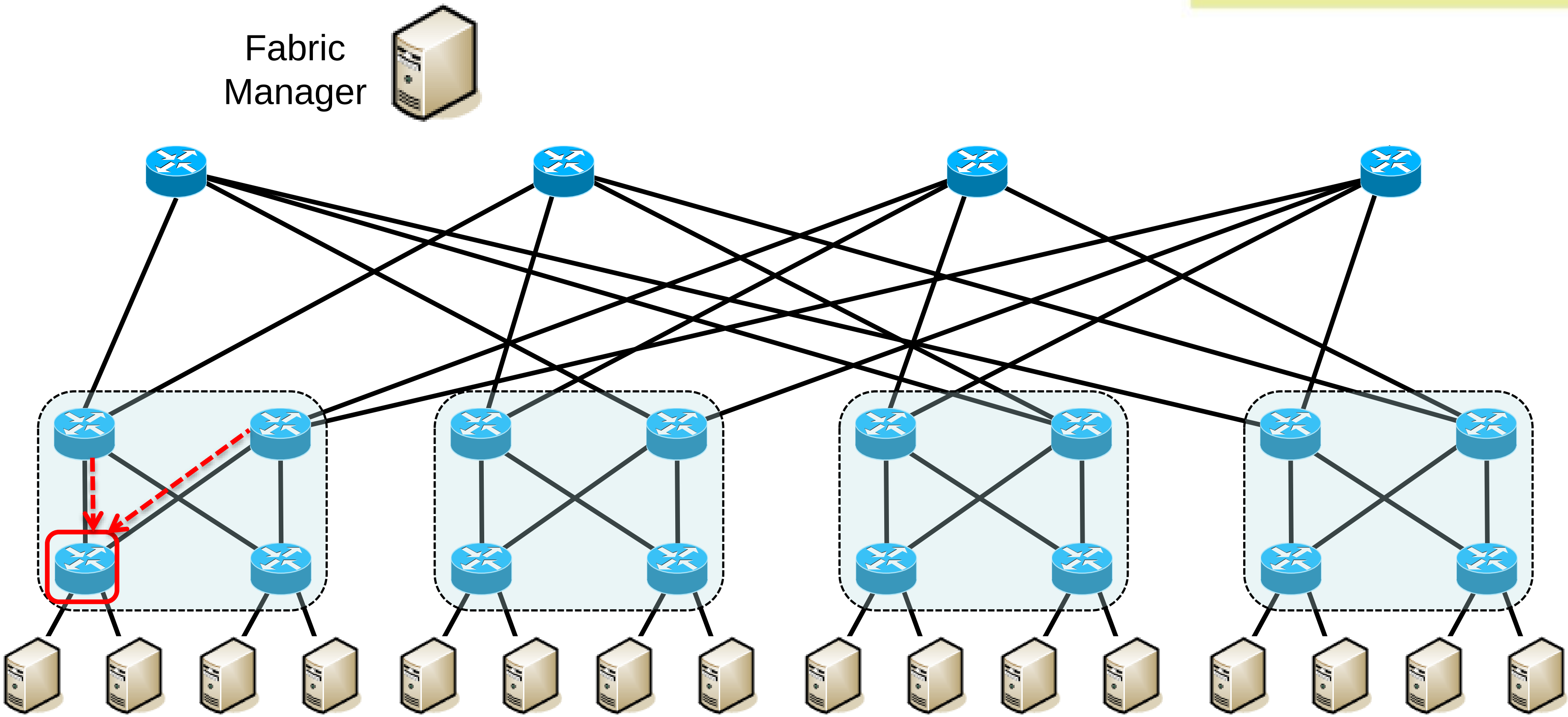
Switch ID	Pod Number	Position	Tree Level
A0-2B-FB-23-34-01			

Location Discovery Protocol



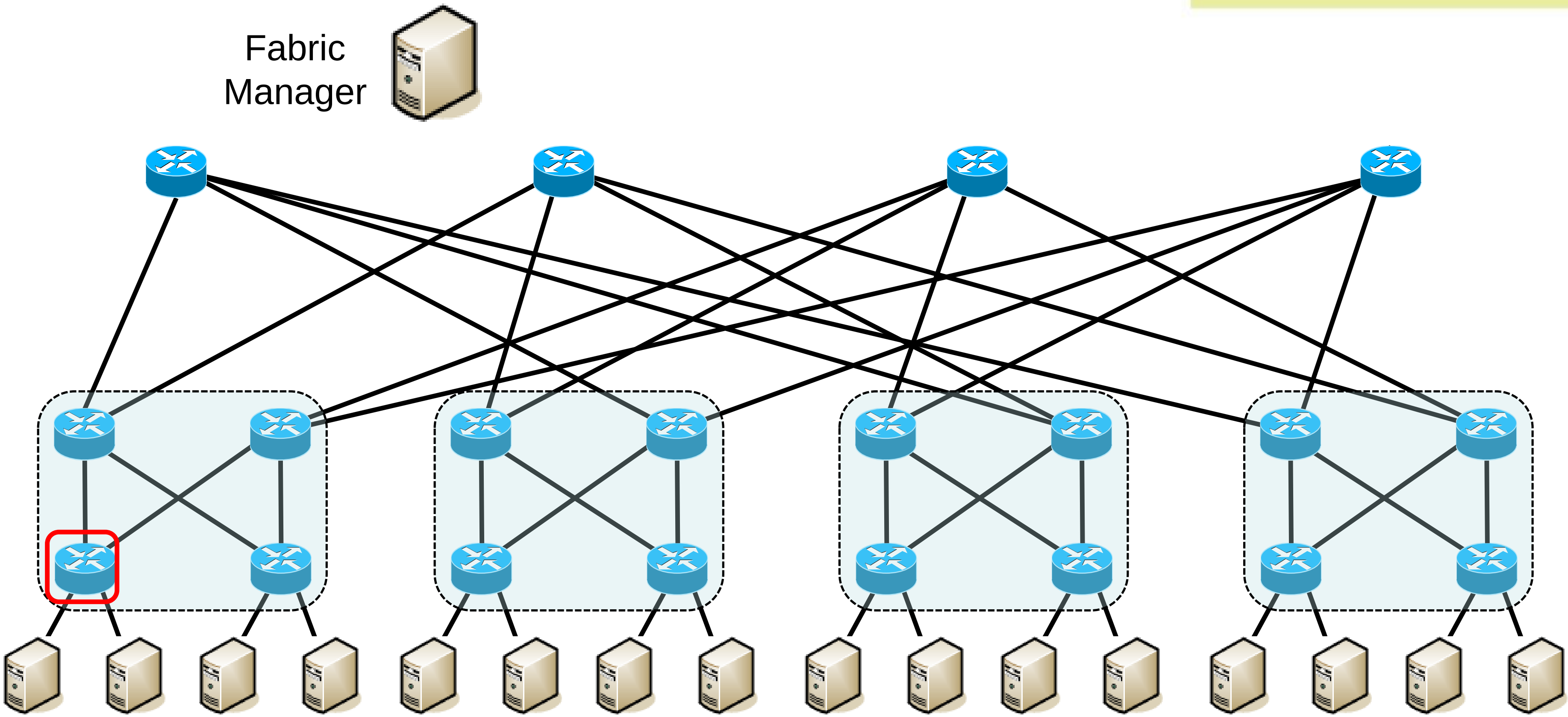
Switch ID	Pod Number	Position	Tree Level
A0-2B-FB-23-34-01			

Location Discovery Protocol



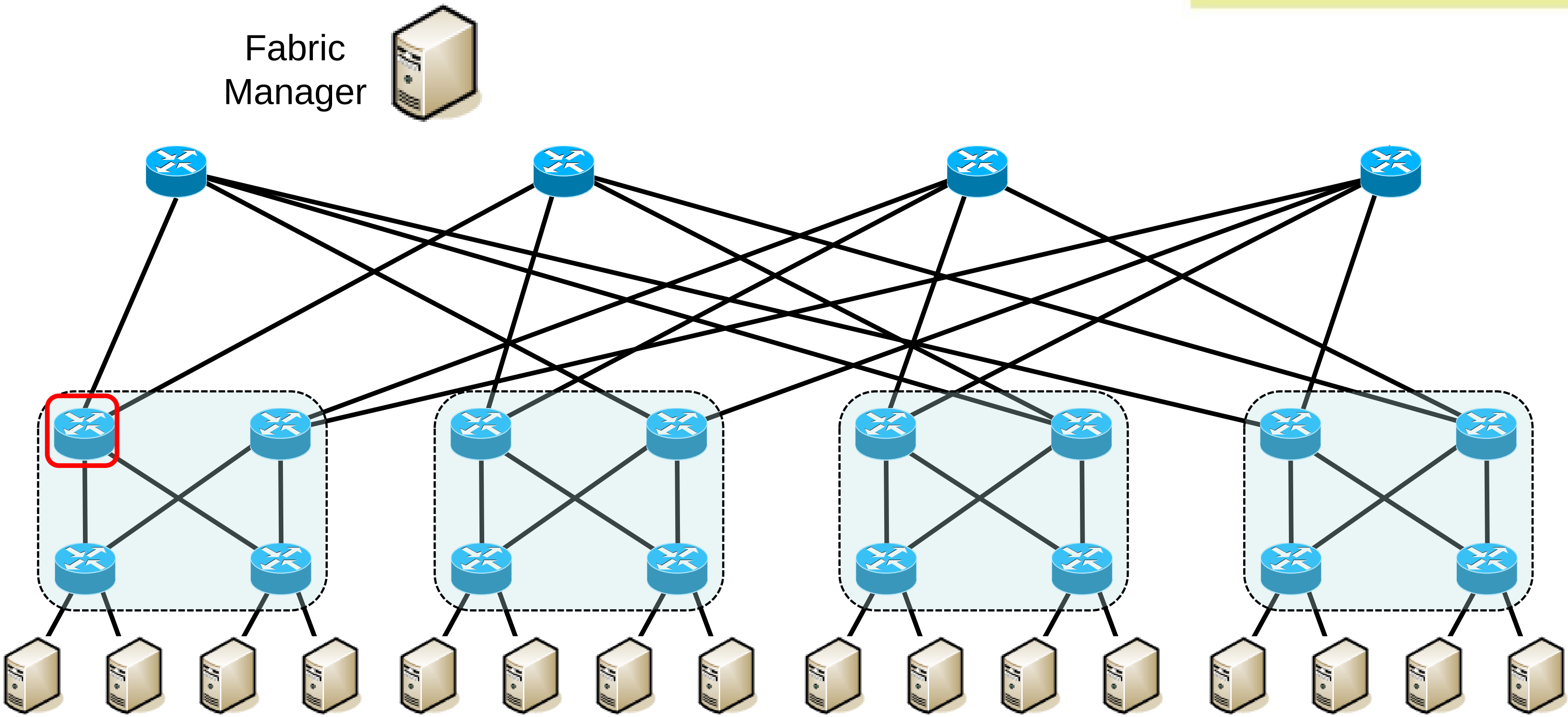
Switch ID	Pod Number	Position	Tree Level
A0-2B-FB-23-34-01			

Location Discovery Protocol



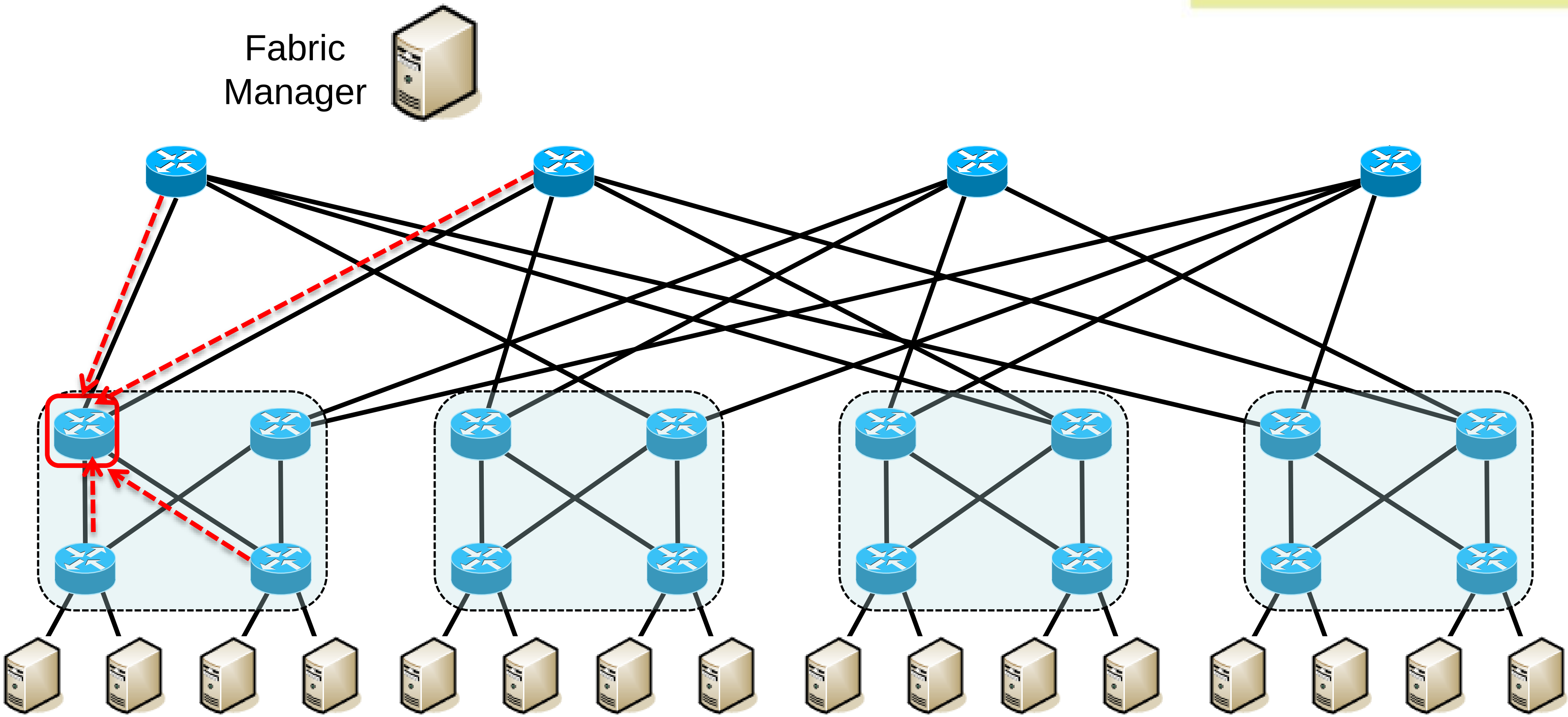
Switch ID	Pod Number	Position	Tree Level
A0-2B-FB-23-34-01			0

Location Discovery Protocol



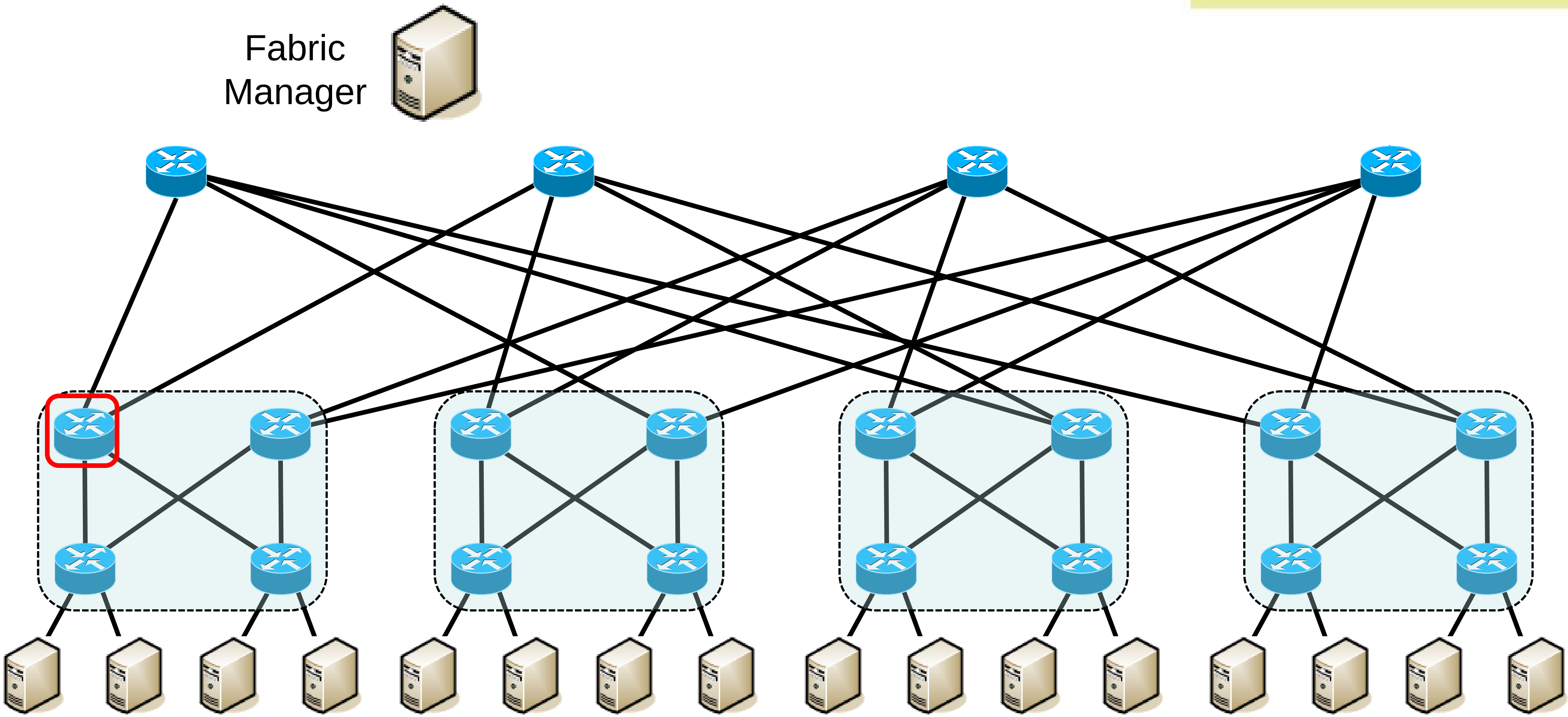
Switch ID	Pod Number	Position	Tree Level
A1-25-EB-23-2A-10			

Location Discovery Protocol



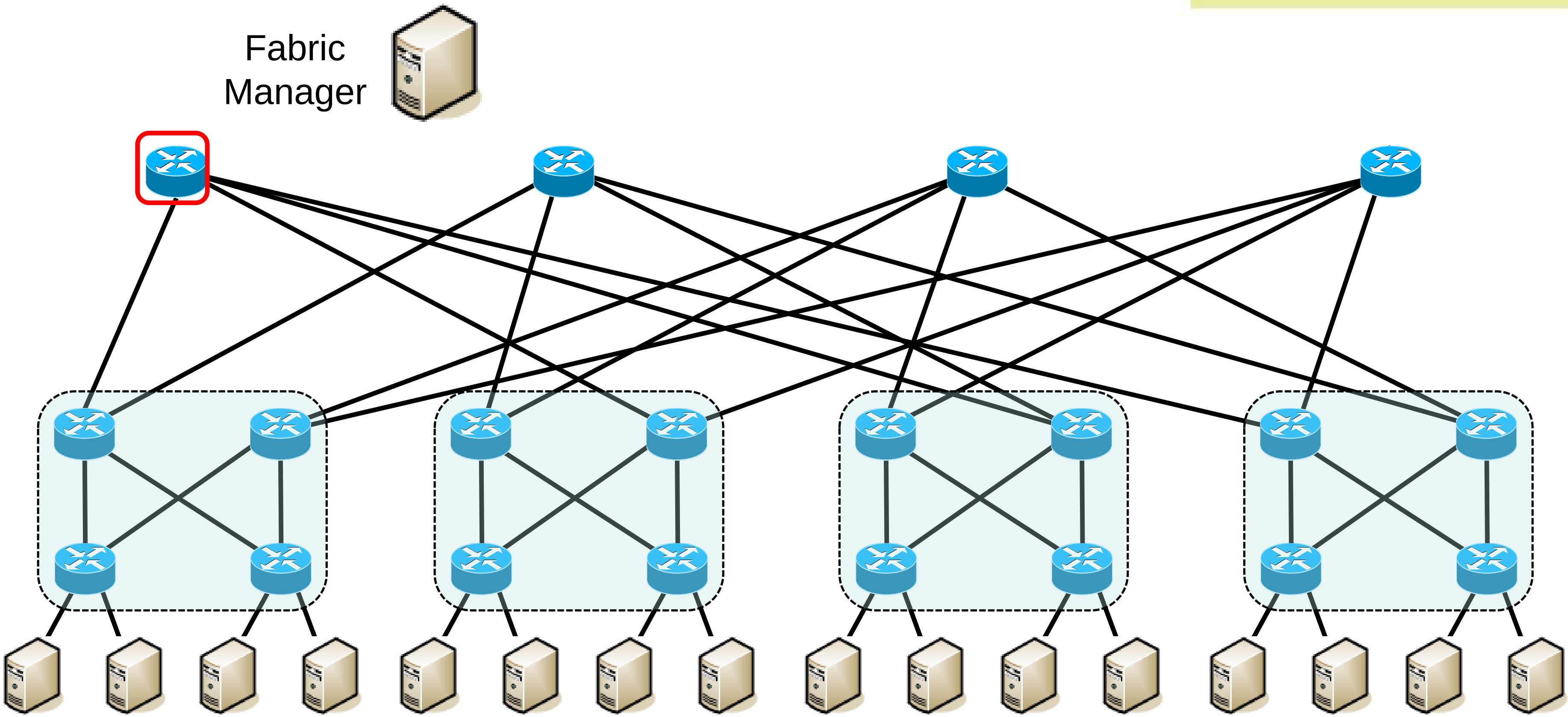
Switch ID	Pod Number	Position	Tree Level
A1-25-EB-23-2A-10			

Location Discovery Protocol



Switch ID	Pod Number	Position	Tree Level
A1-25-EB-23-2A-10			1

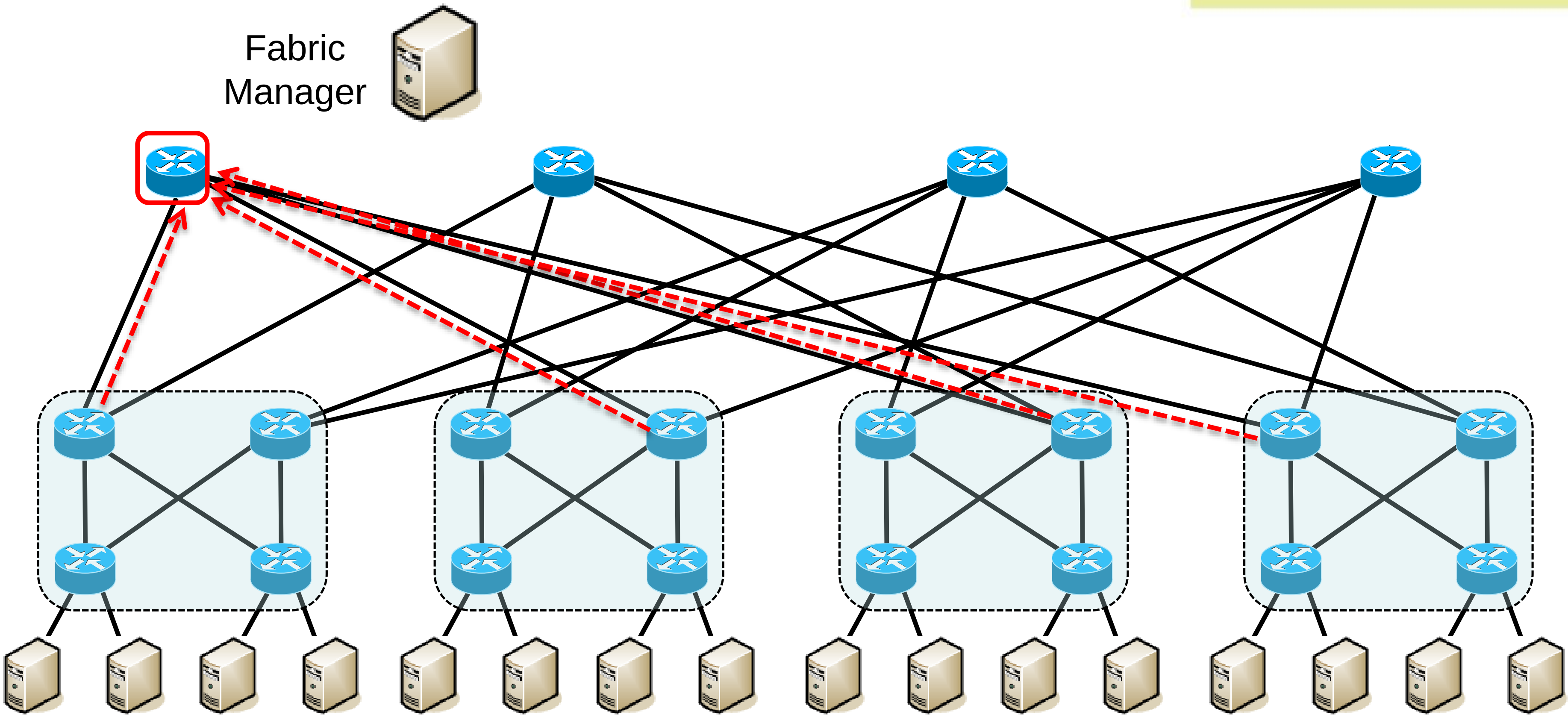
Location Discovery Protocol



Switch ID	Pod Number	Position	Tree Level
B1-25-34-13-2A-10			

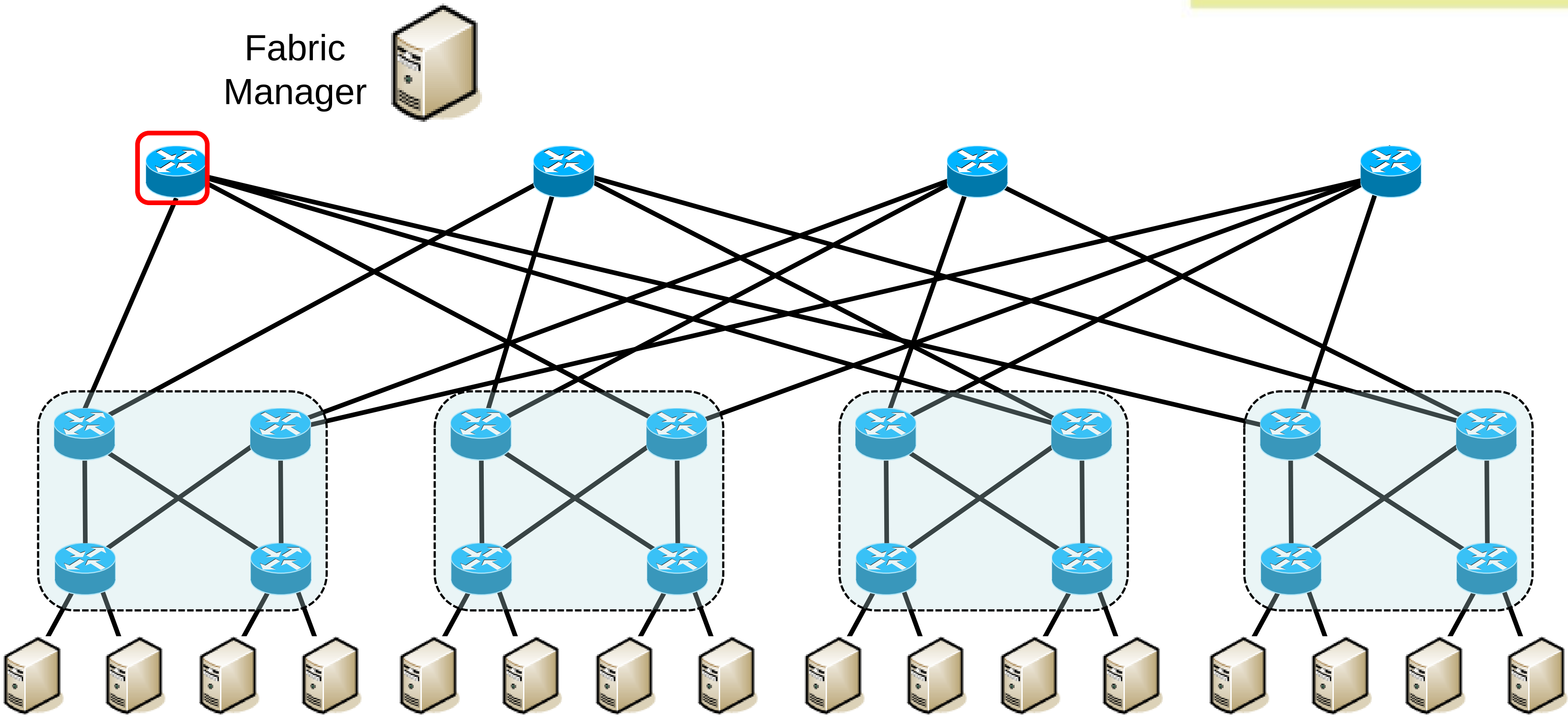
Location Discovery Protocol

Institut für Kommunikations-Technik



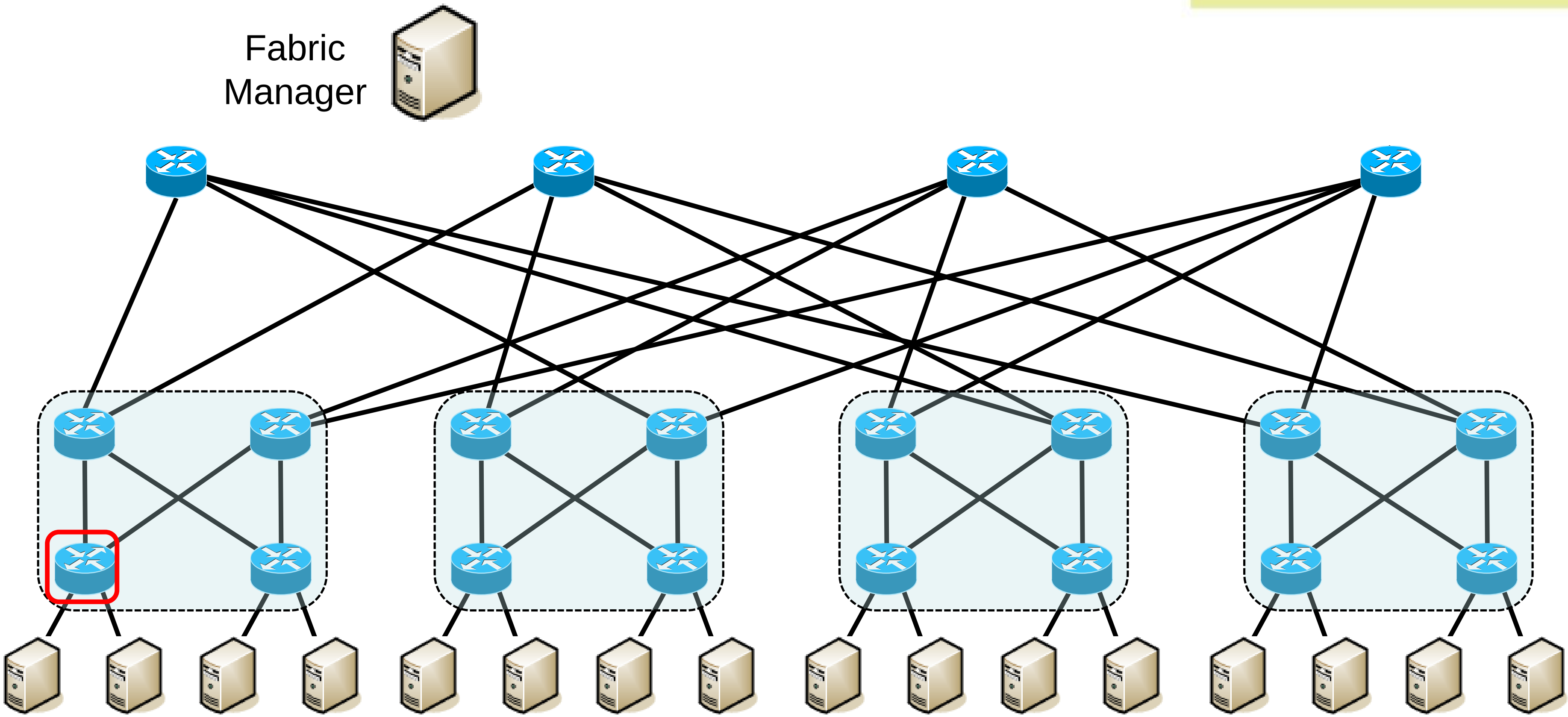
Switch ID	Pod Number	Position	Tree Level
B1-25-34-13-2A-10			

Location Discovery Protocol



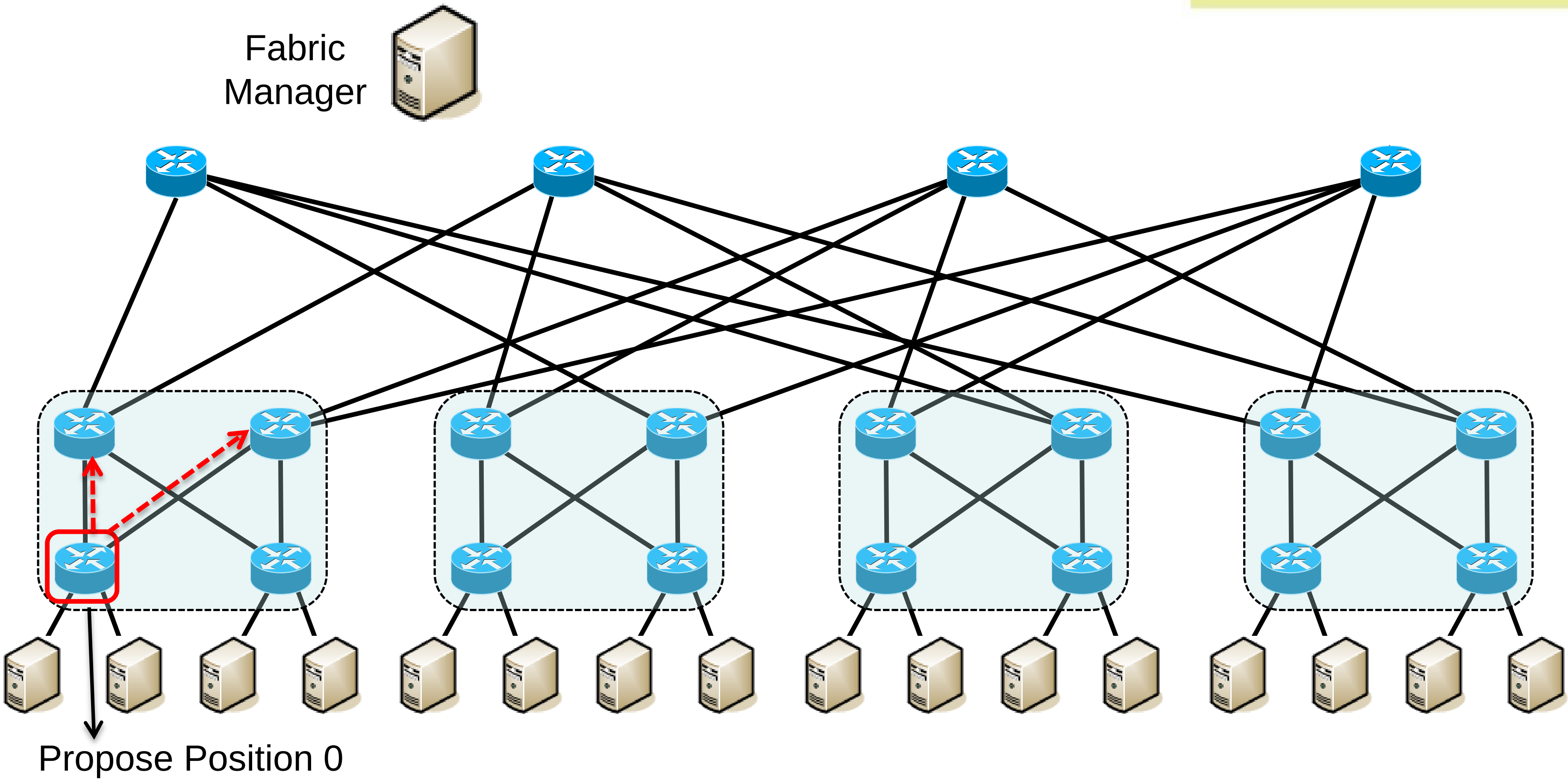
Switch ID	Pod Number	Position	Tree Level
B1-25-34-13-2A-10			2

Location Discovery Protocol



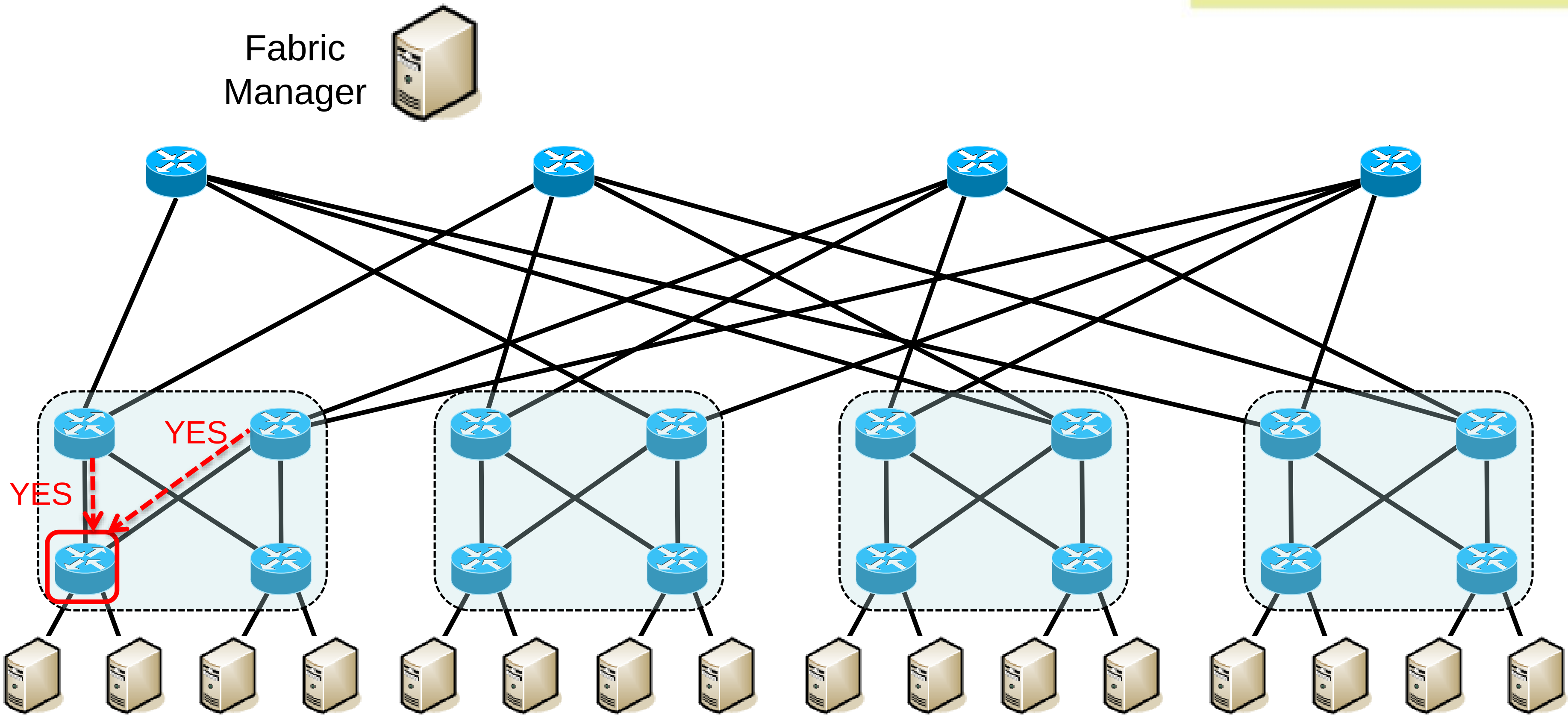
Switch ID	Pod Number	Position	Tree Level
A0-2B-FB-23-34-01			0

Location Discovery Protocol



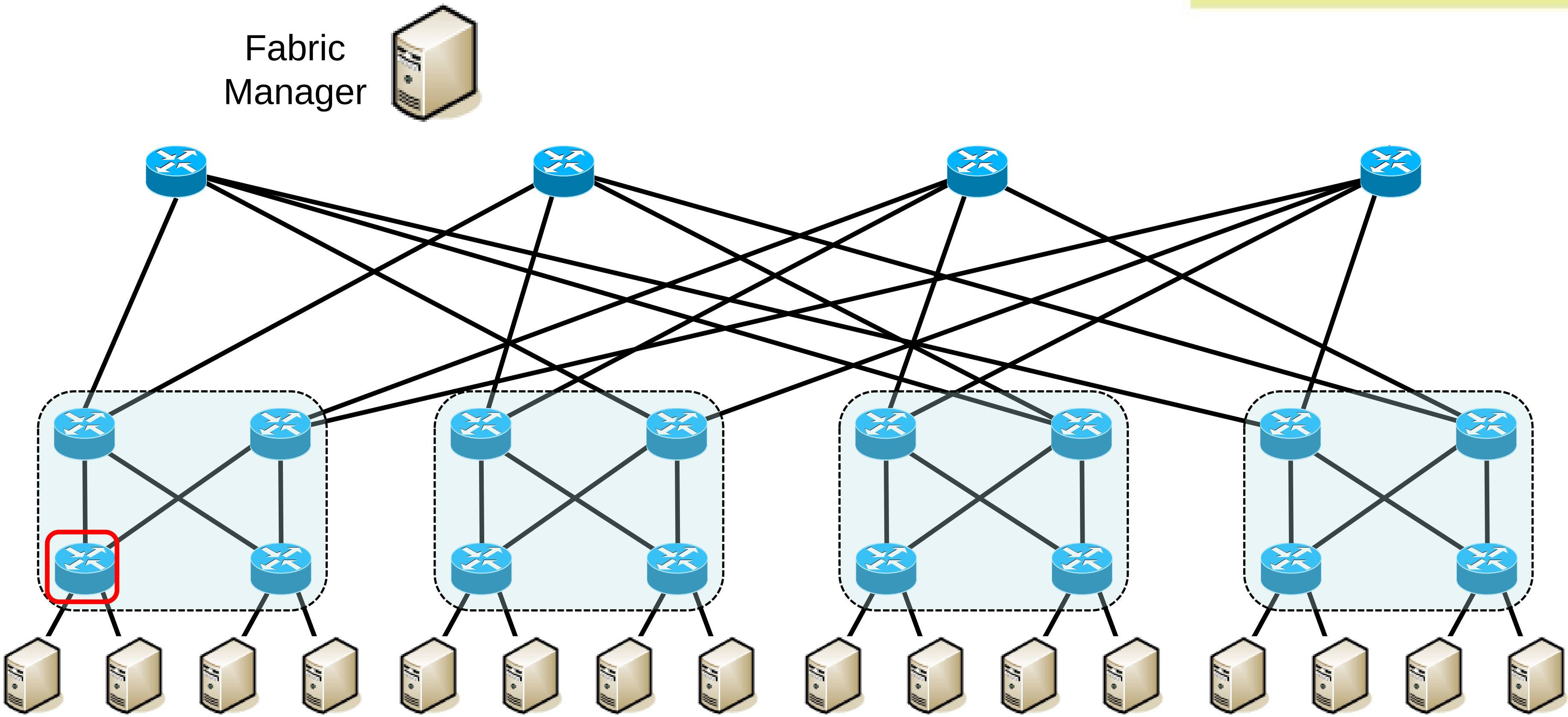
Switch ID	Pod Number	Position	Tree Level
A0-2B-FB-23-34-01			0

Location Discovery Protocol



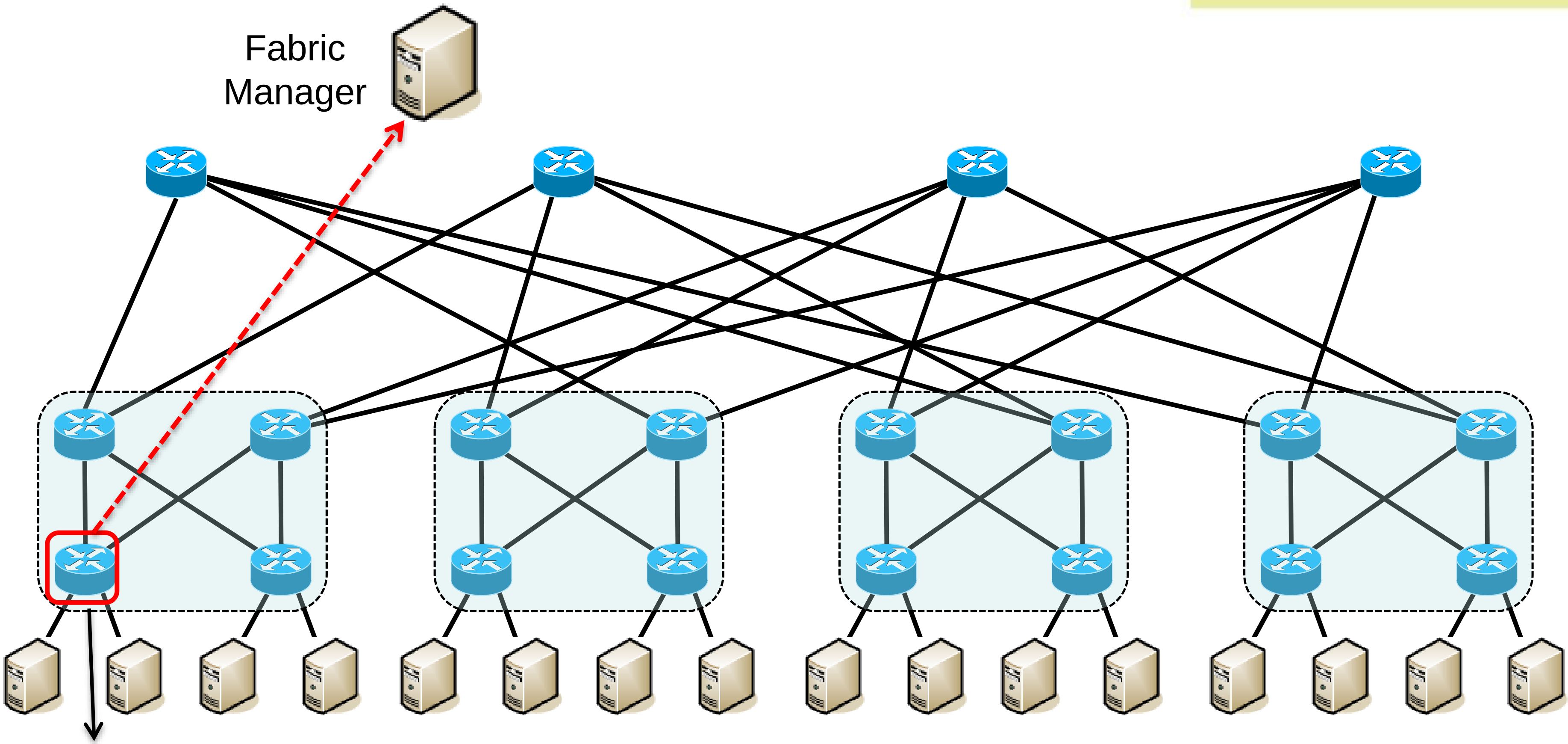
Switch ID	Pod Number	Position	Tree Level
A0-2B-FB-23-34-01			0

Location Discovery Protocol



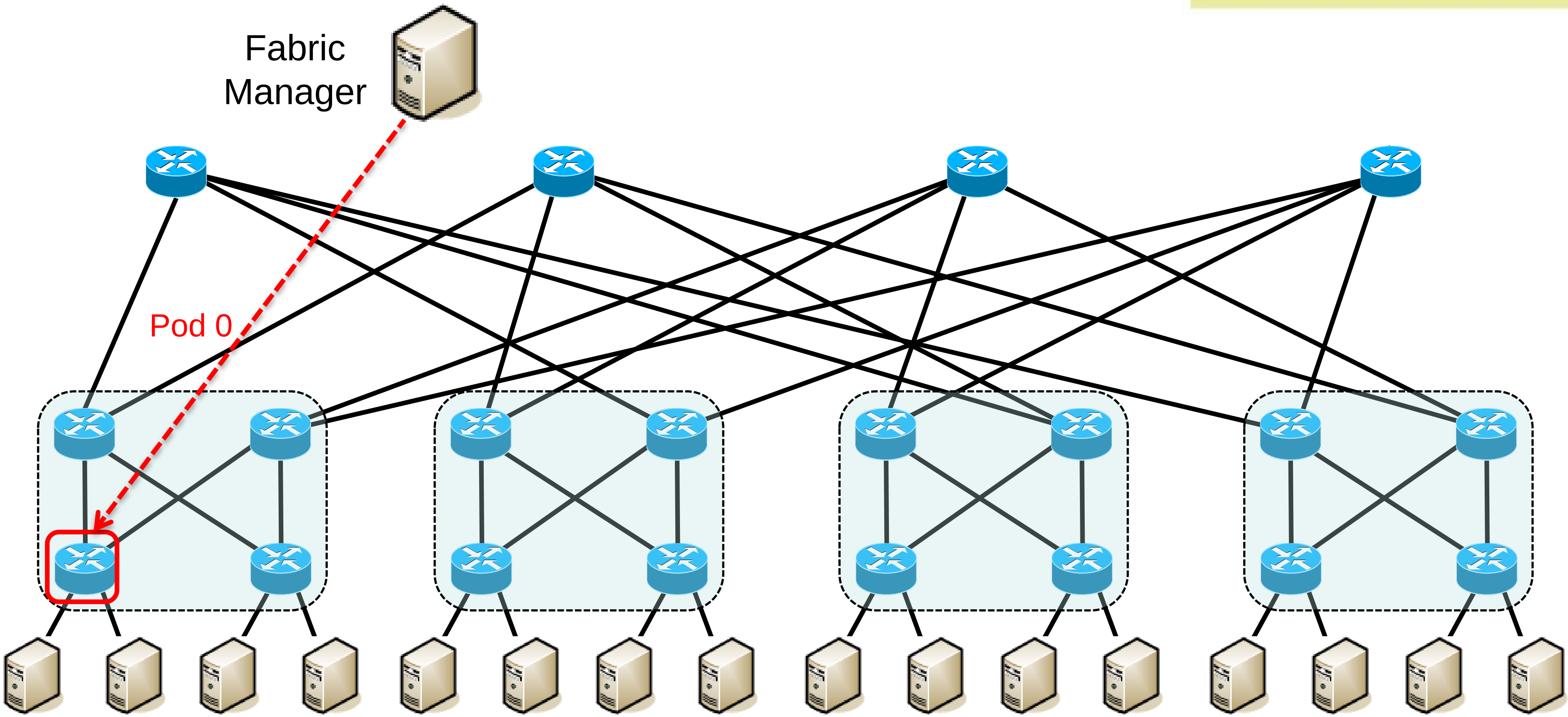
Switch ID	Pod Number	Position	Tree Level
A0-2B-FB-23-34-01		0	0

Location Discovery Protocol



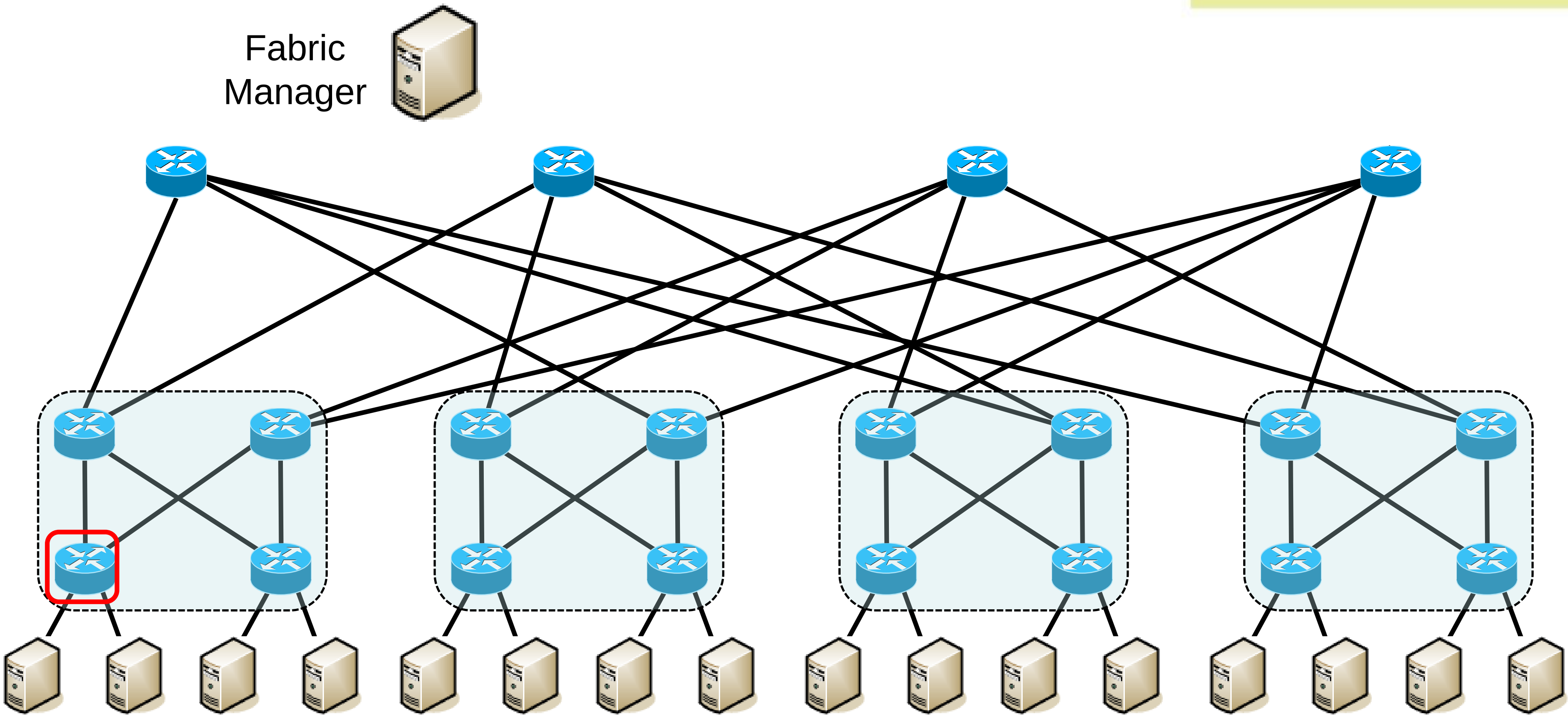
Switch ID	Pod Number	Position	Tree Level
A0-2B-FB-23-34-01		0	0

Location Discovery Protocol



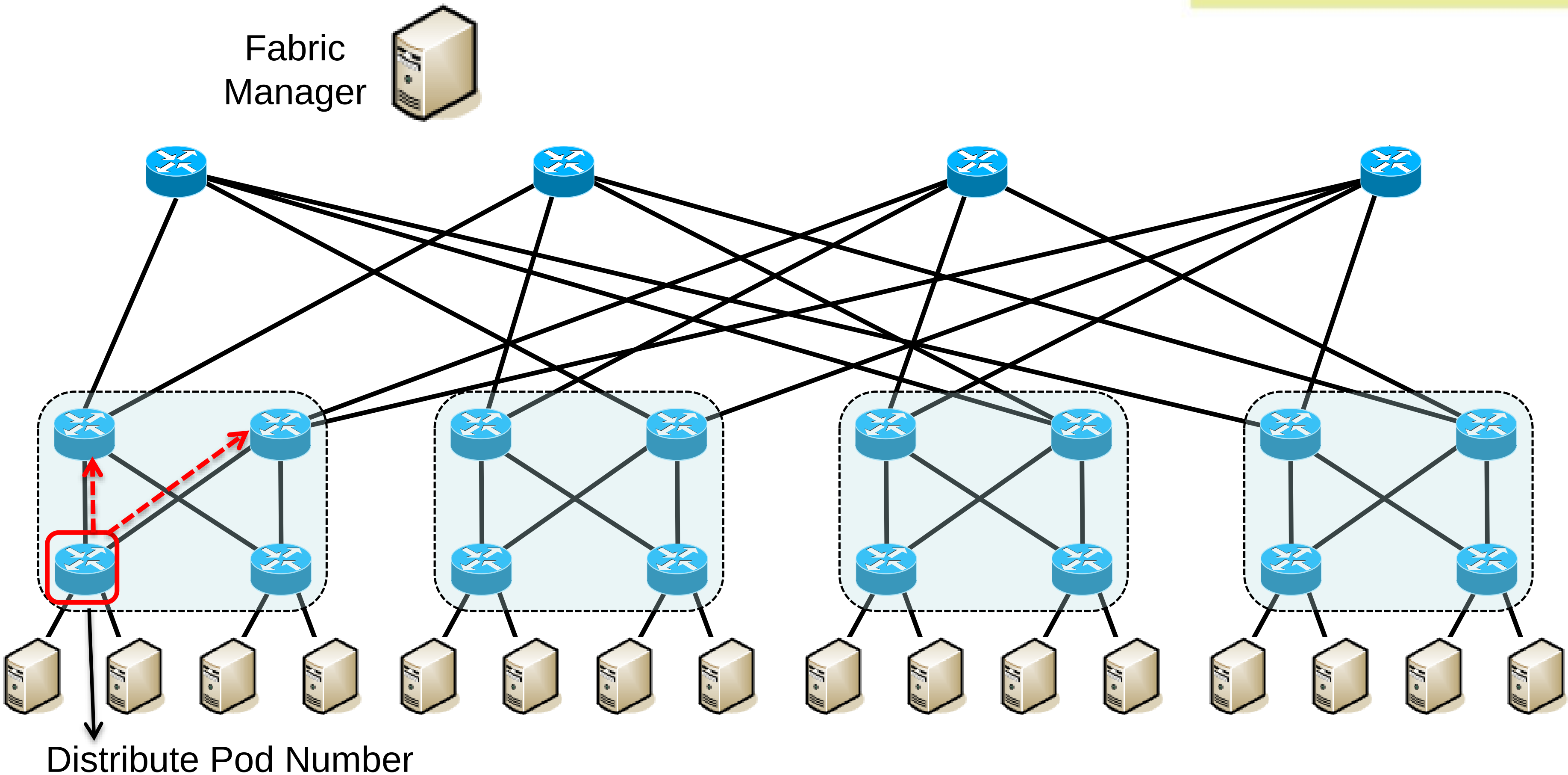
Switch ID	Pod Number	Position	Tree Level
A0-2B-FB-23-34-01		0	0

Location Discovery Protocol



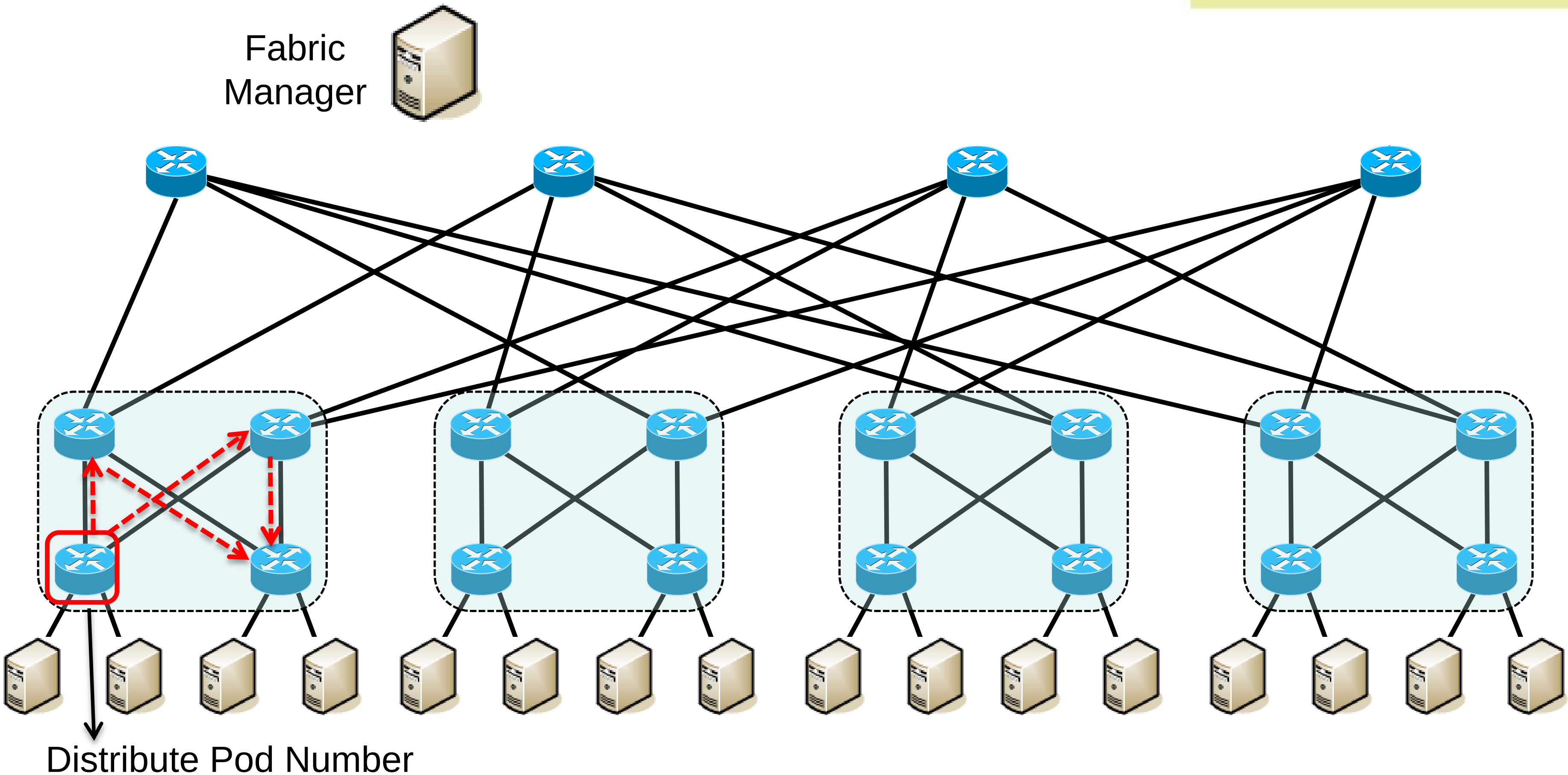
Switch ID	Pod Number	Position	Tree Level
A0-2B-FB-23-34-01	0	0	0

Location Discovery Protocol



Switch ID	Pod Number	Position	Tree Level
A0-2B-FB-23-34-01	0	0	0

Location Discovery Protocol



Switch ID	Pod Number	Position	Tree Level
A0-2B-FB-23-34-01	0	0	0

References

- A. Greenberg, et al., **VL2: A Scalable and Flexible Data Center Network,** ACM SIGCOMM 2009
- J. Mudigonda, et al., **SPAIN: COTS Data-Center Ethernet for Multipathing over Arbitrary Topologies,** USENIX NSDI 2010
- C. Guo, et al., **BCube: A High Performance, Server-centric Network Architecture for Modular Data Centers**, ACM SIGCOMM 2009
- R. Mysore, et al., **PortLand: A Scalable Fault-Tolerant Layer Data Center Network Fabric**, ACM SIGCOMM 2009
- A. Vahdat, et al., **Scale-Out Networking in the Data Center**, IEEE Micro 2010
- **Transparent Interconnection of Lots of Links (TRILL),** IETF, http://datatracker.ietf.org/wg/trill/charter/