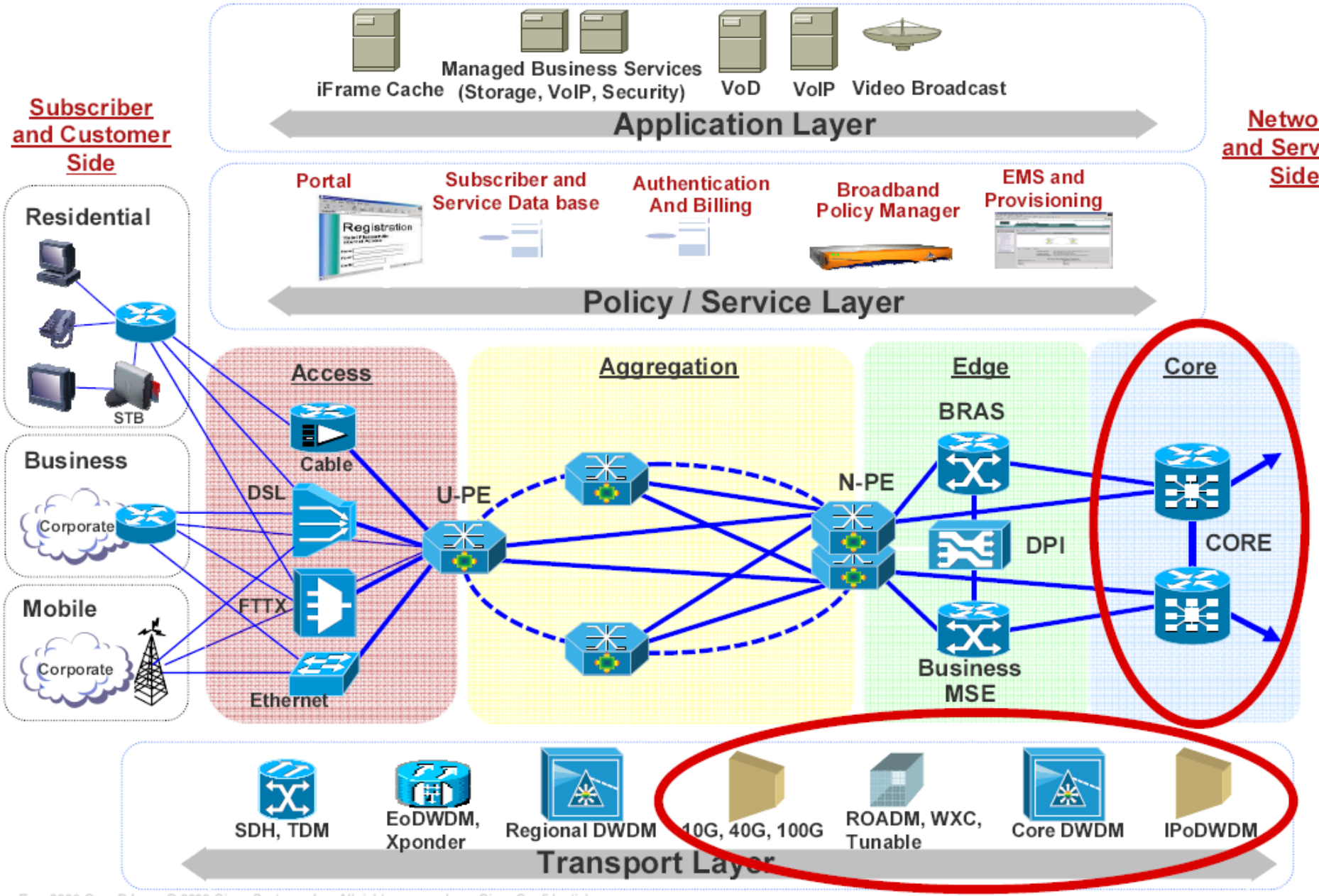
Design of High-Availability Resilient Converged Computer Networks

Focus of the Lecture

- Focus on network topology and protocol implementation considerations
- Mostly focused on enterprise campus and WAN design
- Security recommendations are not discussed
- QoS issues will be discussed in the next lecture version ;-)

General Global “next-generation network” Architecture Model



Design Areas

- Enterprise campus design
- WAN/MAN design
- High-performance carrier/ISP core network design
- Data center design
- SAN design
- Server Farm/E-Commerce Module design
- Intelligent WLAN design
- ...

Network Design Lifecycle

- Preparation
  - strategy, high-level conceptual architecture, financial justifications
- Plan
  - analyze initial requirements – goals, user needs, site characteristics, existing solution
- Design
- Implement
- Operate
- Optimize

Design and Implementation Metodologies

- Modularization & Layering
  - Decompose the network into more manageable pieces
  - well-defined interfaces
  - change in the module does not affect the other part
  - suboptimal, but benefits prevail
    - the general design is obviously further optimized little bit according to the real traffic characteristics
  - peer or hierarchical relationship between modules

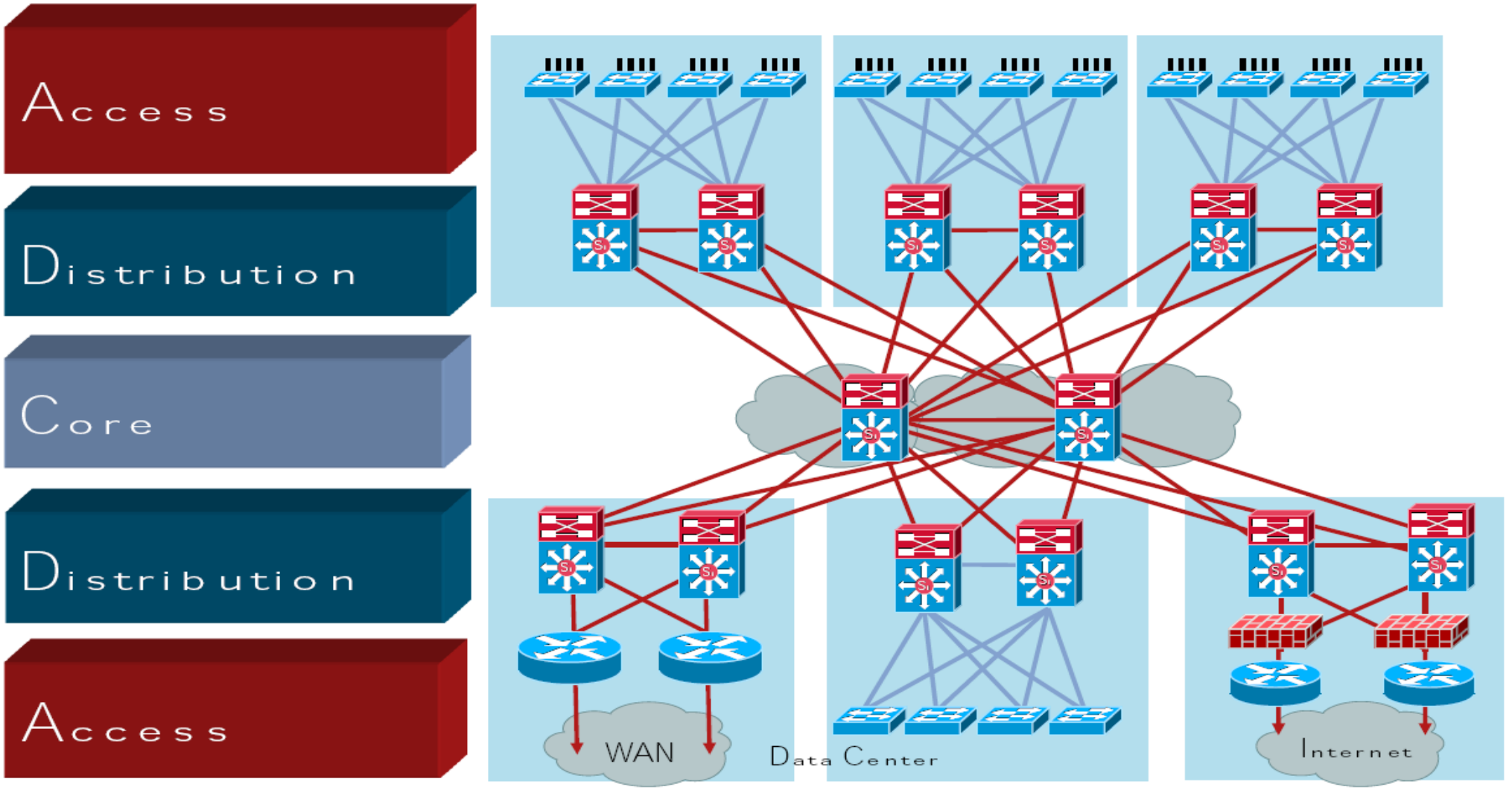
Hierarchical Network Model

- hierarchization divides network into multiple layers
  - from the topology/structure point of view
  - not to be confused with ISO RM layers
- provides modular view of a network
- helps to build scalable deterministic infrastructure
- ensures deterministic traffic patterns and predictable behavior in case of link/device failure
  - Simplifies troubleshooting
  - Helps to develop failure scenarios

3-tier Hierarchical Model

- Developed from the experience gained during Internet evolution
- Access Layer
  - aggregates workstations/IP phones/servers/APs/teleworkers and provides connectivity to distribution layer
  - L2 switches (LAN), aggregation devices (WAN)
  - access authentication (802.1x, MAC filters, NAC, ...)
- Distribution Layer
  - aggregates wiring closets, segments workgroup
  - typically L3 switches
  - policy, QoS
- Core Layer (backbone)
  - high-speed scalable packet switching (often MPLS)
    - no ACLs, no CPU-oriented processing
  - redundancy, fast convergence

Layered Design Example



Model Variations

Distribution and Core layer may be combined together in some simple cases (collapsed core architecture)

- reduces cost, but limits scalability

Reasons for Separation of Devices' Roles according to their Position in the Layered Model

- Individual device models are optimized for various tasks
  - very distinct HW/SW combinations in individual layers
- A designer should try to reach
  - Simple configuration
    - less risk of human error
  - Small OS images
    - less risk of software bugs, less expensive

Typical Characteristics and Responsibilities of Routers in Individual Layers

- Backbone routers
  - optimized for extremely fast packet switching
  - use limited set of WAN technologies and routing protocols
  - contain reachability information for all destinations in the network and in the outside world (large routing tables)
- Distribution routers
  - contain topological information for their region
  - for inter-region routing forward packets to the backbone
  - support various WAN technologies and routing protocols
- Access routers
  - connect customer/enterprise sites to distribution network
  - various WAN link technologies, including dial-on-demand customers
  - aggregate customers (hundreds, thousands)
  - authentication, ACLs, packet classification & marking, traffic policing, accounting

Recommended Link Oversubscription
(source: Cisco)

- 20:1 on access-to-distribution uplinks
- 4:1 on distribution-to-core uplinks
- Potential (infrequent) congestions have to be solved by implementation of QoS mechanisms

How to Reach High Availability (1)

- Optimal redundancy, avoidance of single point of failure
  - provide alternate paths
    - BUT: too much redundancy may cause unpredictable behavior (3+ alternate uplinks)
  - control plane redundancy
    - multiple control processors
    - control information exchanged between virtual interfaces (loopbacks) over the redundant physical infrastructure
- Recommended design:
  - fully-meshed core
  - redundant distribution layer switches + L3 link between them, redundant links to core layer
  - redundant uplinks of access switches

How to Reach High-Availability (2)

- Traffic-related methods
  - QoS
  - randomization
    - avoidance of the synchronization of network data or control traffic that can lead to cyclic congestion or instability
      - RED, random timers in routing/management protocols etc.
- Control plane related methods
  - hystheresis and dampening to avoid oscillations
    - rapid interface state changes, route flapping, ...
  - stabilizing routes improves TCP performance because of small RTT variance
    - retransmission timeout calculation

How to Reach High Availability (3)

- Localization of traffic
  - consider content caching as a natural part of network topology
- Analyze the network behavior during failure modes
  - consider failure of individual design components (and their combinations)
    - modular/hierarchical design approach simplifies this analysis considerably

Campus Network Design

Processes Involved in Recoveries from Failures

- multiple protocols have to converge before a failure is repaired
  - STP, FHRP, routing protocol
- ensure predictable and reasonable behavior even in transient states
  - fine-tune timers to ensure the proper order of convergence actions on L1/L2/L3 layers
- interface up/down pacing timers
  - quick reaction on interface failure event
  - be conservative after the interface goes up
    - the network operation was already re-established after failure, no need of quick changes

First-Hop Redundancy Protocols (FHRP)

- Virtual Router Redundancy Protocol (IETF)
- Hot Standby Redundancy Protocol (Cisco)
  - virtual IP/MAC address shared by multiple gateways
  - one active gateway, other(s) serve as backup
  - constant monitoring of active GW operation
  - no load balancing
- Gateway Load Balancing Protocol (Cisco)
  - provides first-hop load balancing

HSRP/VRRP optimization

- Router priorities for becoming HSRP/VRRP primary router
- Preemption
  - adjustable preemption timers
    - when the network is returning to the “default” state after the failure is repaired
  - needs to take into account STP and L3 protocol convergence times to avoid suboptimal multihop paths
- Object tracking
  - takes into account an operational state of uplinks, presence of specific route in routing table, ...
  - increase/decrease router priority based on tracked object state

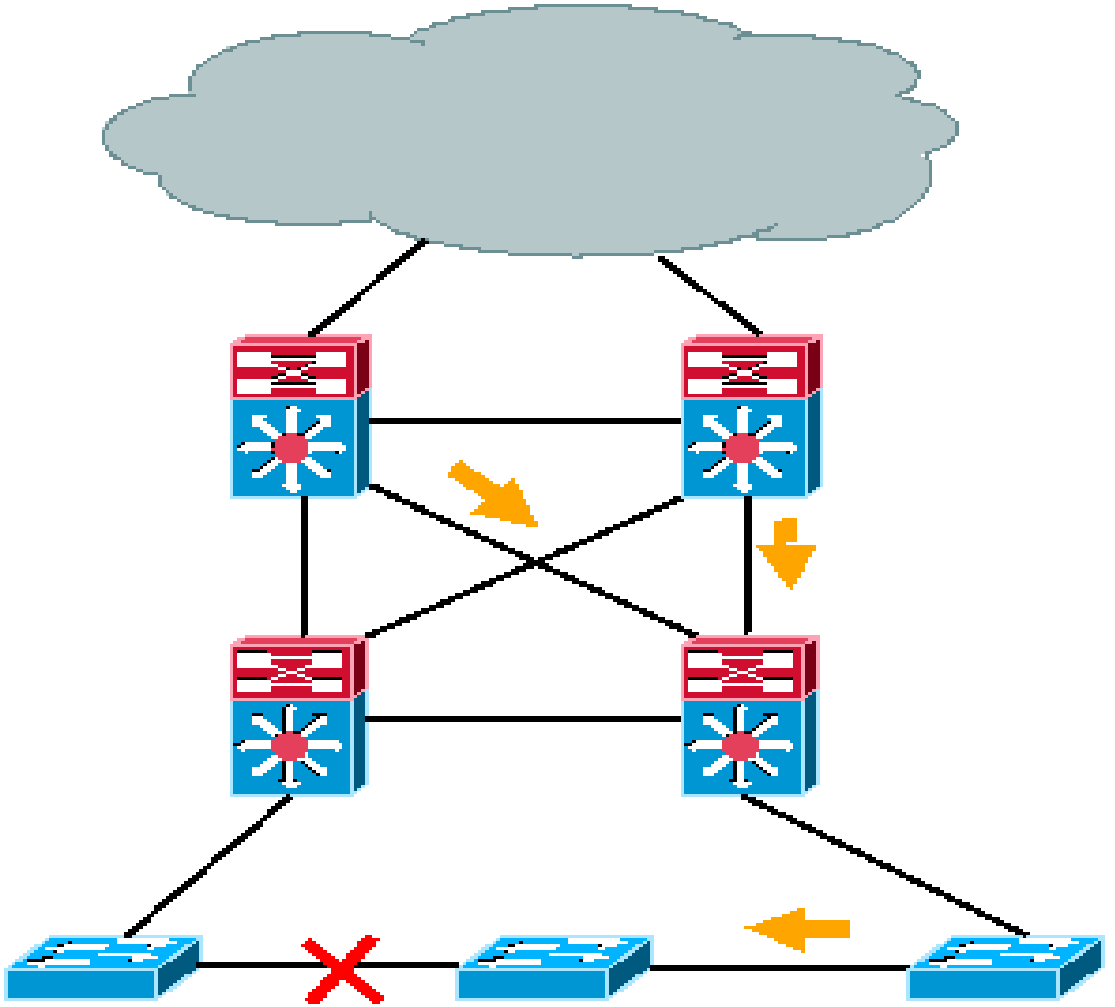
Gateway Load Balancing Protocol (GLBP)

- Load-balances between multiple gateways
- Active Virtual Gateway (AWG)
  - responds to ARP requests
    - uses multiple virtual MAC addresses for the single virtual GW IP address to distribute load among multiple GWs
    - response MAC addresses selected by round robin or takes current GWs'/uplinks' load into account
- Active+(multiple) Standby Virtual Forwarders for every virtual MAC address

L2 Topology Recommendations

- L2 core is problematic
  - failure of switch in the middle cannot be detected by router link state change
    - slow convergence
      - there is a need to wait until routing protocol notices the failure based on expired timers
- Avoid trains of switches connected to 2 routers on the sides
  - results to blackholing if the switch in the middle fails and core delivers the return traffic (or 50% of it in case of load balancing) to the router on the “wrong” side

Behaviour of Train of Switches



- 50% of the return traffic is dropped

STP Recommended Design Practices (1)

- Avoid (a slow) STP convergence as a mechanism of device/link failure repair
  - use STP just to protect against loops caused by miswiring or malicious users
  - STP works poorly with multicasts
    - after topology change, CAM table is flushed and the information learned from IGMP Snooping is lost
- Protect the root and preferred STP topology
  - RootGuard, BPDUGuard, unidirectional link detection

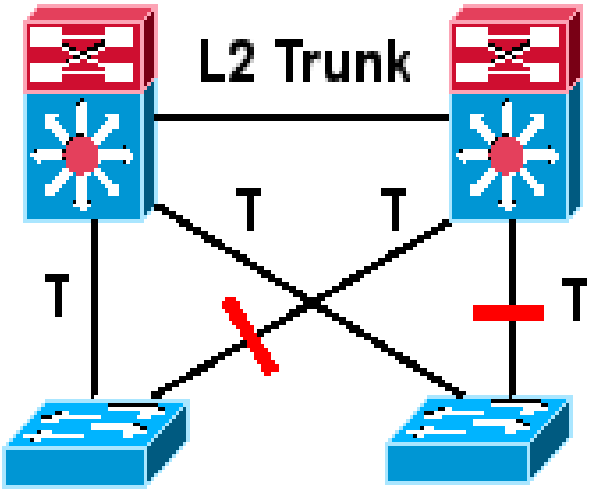
STP Recommended Design Practices (2)

- Implement mechanisms to accelerate convergence
  - PortFast, UplinkFast, BackboneFast
  - incorporated in RSTP and enabled by default
- Keep STP root and HSRP/VRRP primary active synchronized
  - to avoid transit traffic on link between distribution switches

Channel Bundling Best Practices

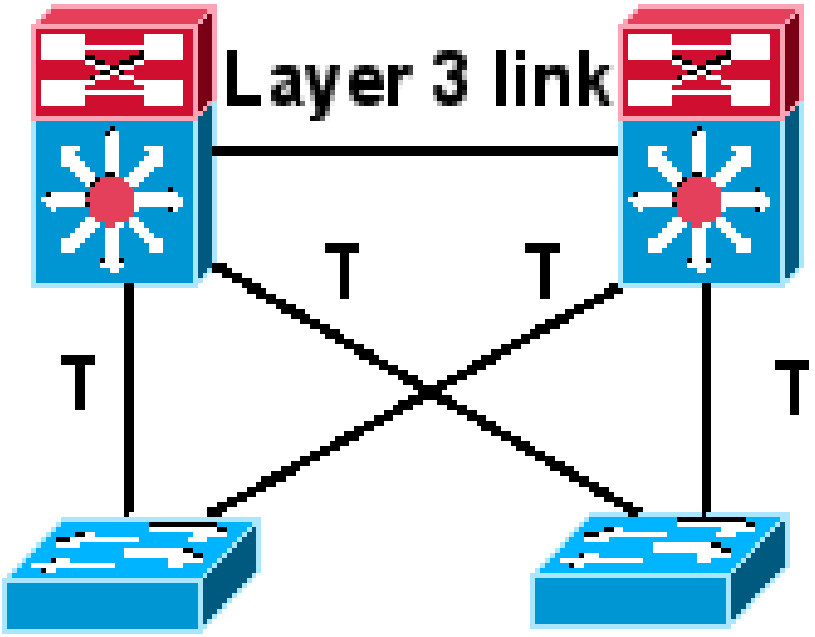
- L2 link bundles may increase uplink bandwidth without increasing number of L3 routing protocol adjacencies
- Routing protocol should be able to adapt (bundled) link cost according to the number of links currently in the operational state
- Selection of proper L2/L3 hash algorithm for load balancing
  - per-source, per-destination, combination

L2-to-L3 Boundary Design Options (1)



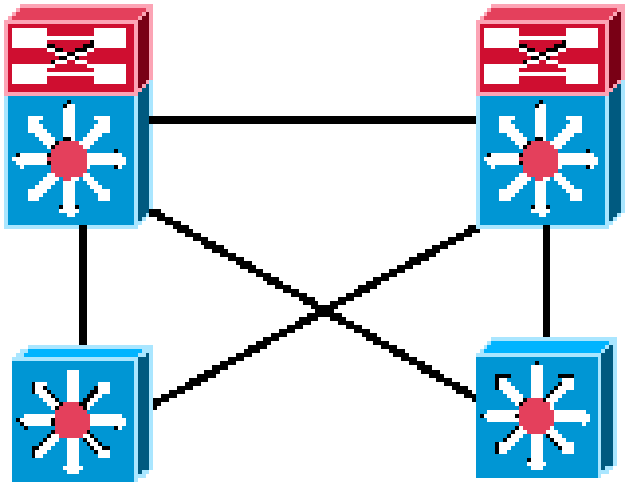
- L2 distribution switch interconnection
  - All links are L2
  - Not recommended – depends on STP convergence
  - HSRP and STP roots should be aligned to avoid multihop switching
  - Applicable if VLANs need to span multiple access layer switches

L2-to-L3 Boundary Design Options (2)



- L3 distribution switch interconnection
  - most recommended
  - VLAN = subnet, no VLAN spans across access-layer switches
  - STP: both uplinks are forwarding

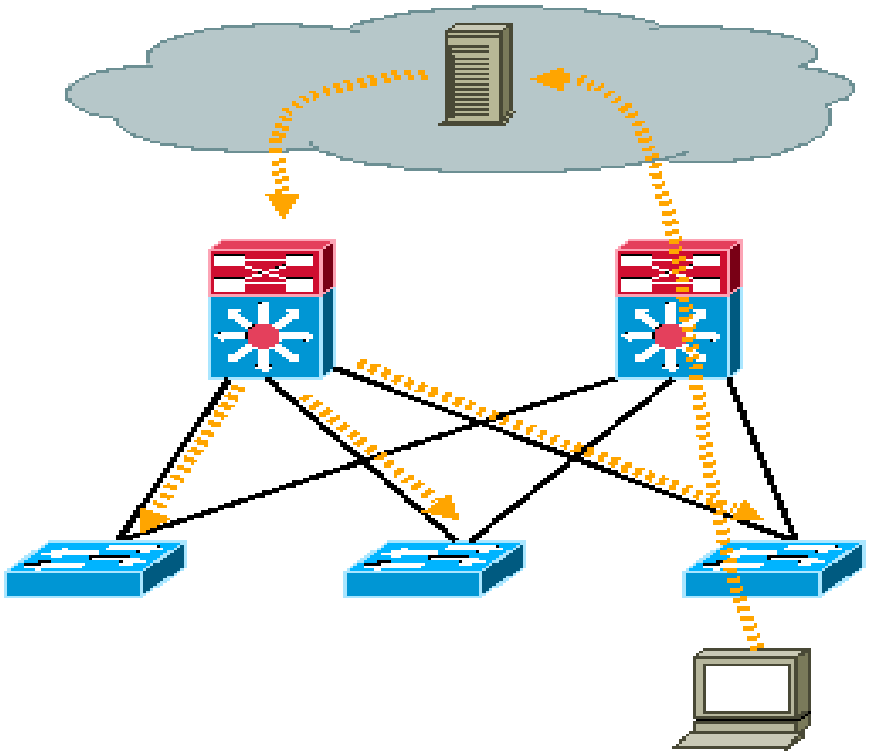
L2-to-L3 Boundary Design Options (3)



- L3 access-to-distribution layer uplinks (routed model)
  - **all** links are L3
  - no STP, sub 200-ms convergence (900 ms in previous cases)
  - load-balancing (**equal-cost** L3 uplinks)
  - OSPF timers may be tuned to subsecond convergence as CPU resources are as scarce as in WAN
  - expensive

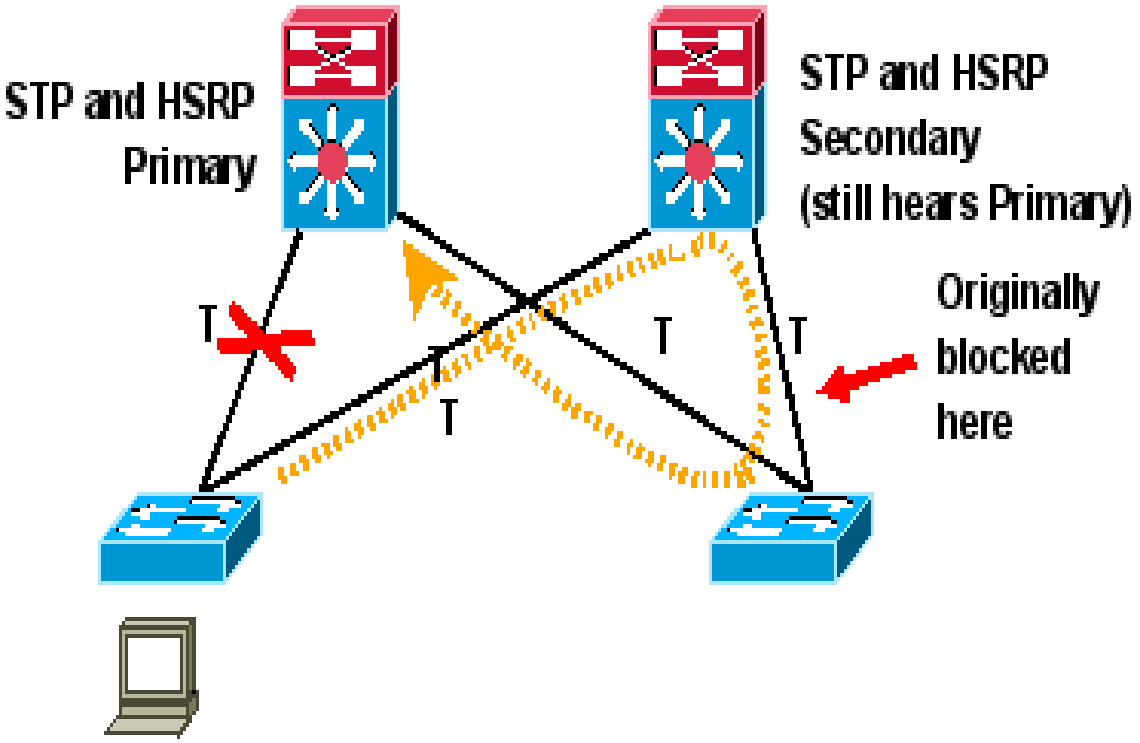
Why Not to Span VLANs Across Multiple Access-Layer Switches (1)

- Asymetric routing + unicast flooding
  - A switch that receives return traffic has no chance to learn the port of the source machine



Why Not to Span VLANs Across Multiple Access-Layer Switches (2)

- Multihop switching (looped figure-8 shape)



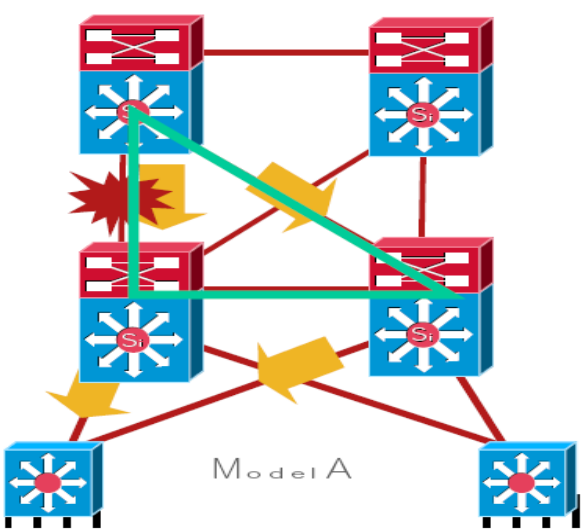
Routing Protocol Design Considerations (1)

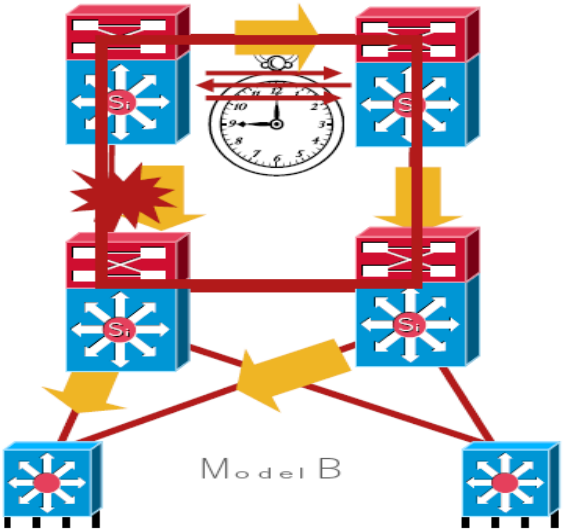
- Routing protocol runs on distribution-to-core and core-to-core links
    - advantageous also for access layer, but not widely implemented because of the high cost
- Need of fast detection of link failures
  - OSPF hellos are NOT primarily designed as a mechanism of fast link failure detection
    - as they all have to be processed by control plane
  - use something like Cisco BFD and routing process notification instead (50ms reaction)
    - processing may be offloaded to (distributed) data plane

Routing Protocol Design Considerations (2)

- Limit number of adjacencies
  - memory, CPU cycle and bandwidth consumption
    - reliable LSA propagation requires CPU
- Peers only on transit links
  - avoid bandwidth/memory/CPU consumption (hellos on multiple VLANs)
  - avoid transit transit via access-layer links
    - alternative distribution-to-core link should be used
  - configure passive interfaces on access layer (trunk) uplinks
- Summarize routes propagated to the core
  - Speeds up the routing convergence process as less LSAs has to be processed
  - Allocation of a summarizable address range in a building block is necessary

Optimization of Distribution-to-Core-layer Convergence





- Alternative equal-cost paths exists on triangle topology
  - link failure is quickly detected by HW
  - no IGP topology recalculation is needed
- Routing protocol must converge on the square topology

OSPF Recommendations (1)

- 1 distribution block = 1 totally stubby area
  - link flaps not propagated beyond distribution switch pairs
- area 0 = core/distribution layer
  - do not extend area 0 to access layer
  - access layer not used for transit
- area definition considerations
  - area placement according to geographic and functional grouping
  - be conservative when adding routers to area 0
    - design to avoid partitioning by single link failure
    - small backbone increases stability
  - make nonbackbone areas stub/totally stubby
  - summarize IP address ranges

OSPF Recommendations (2)

- Recommended area size
  - number of adjacent neighbors proved to have more impact than total number of routers
  - consider amount of information that has to be flooded within the area
  - link quality/stability has an important impact
  - keep LSAs size under MTU (to avoid CPU-demanding fragmentation)
  - no more than 50 routers in any area

OSPF Fast Convergence

- Fast hellos
  - or use BFD to detect link failure and notify OSPF process
- Incremental SFP

IBGP Scalability

- Poor scalability of IBGP full mesh configuration
- Use route reflectors insteads
  - cluster
  - RR clients
  - nonclients
- Confederations are an alternative solution
  - not so much popular nor elegant

Load Sharing Considerations

- routing protocol has to support multiple paths
- per-packet or per-flow
- per-flow is recommended to avoid packet reordering
  - reordering and alternate paths lead to varying round-trip times, which makes TCP operation less optimal

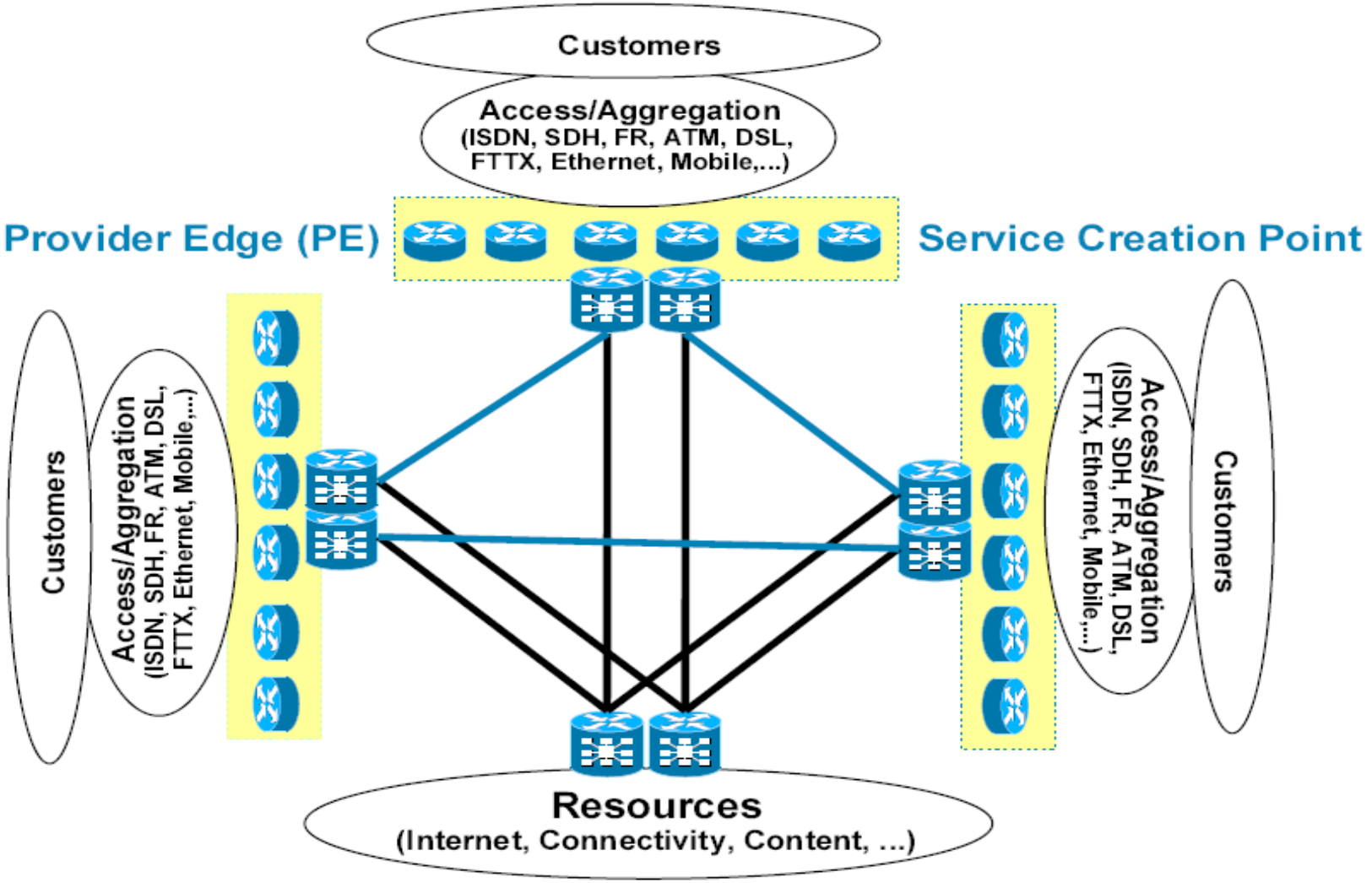
Core Network Design

The Purpose of the Backbone

- interconnects regional distribution networks
- provides connectivity to other peer networks
- must be reliable and scalable

Role of the Core Network (WAN)

Core = Interconnection of PoPs



Distribution/Regional Network Design

- routes intra-regional or inter-regional traffic
- often hub-and-spoke topology
  - distribution center (DC) as hub
    - placement chosen according to geographical proximity to other sites
  - points-of-presence (POPs) at spokes
  - transit POP routers may be also utilized
  - Usual DC implementation
  - dual aggregation LANs
  - dual backbone routers
  - dual backbone WAN connections
- DC may provide services
  - DNS, e-mail and Web hosting, ...
  - services may also be provided in major POPs

Core Topology Design Considerations

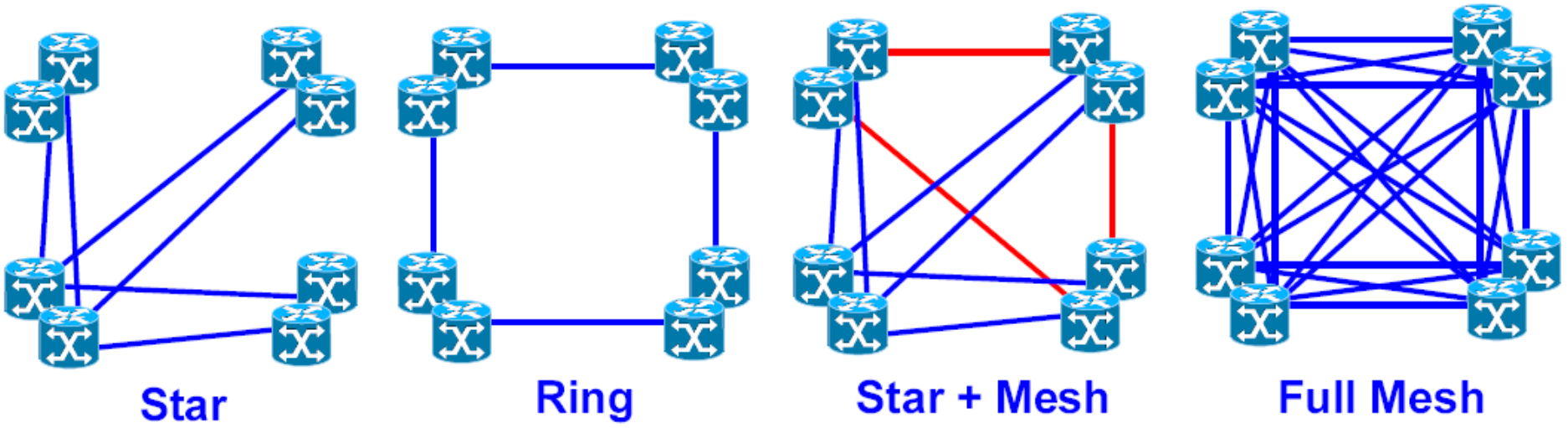
- Both client-server and peer-to-peer traffic patterns
  - In TCP/IP environment, it is extremely hard to predict resource consumption for individual sessions
- General hierarchical design
  - the currently investigated traffic pattern may change in the future
- Design is initially based on financial constraints, population density and application needs
  - may be refined in the future by statistical analysis of traffic
- Full mesh implies routing complexity and consumes a lot of routers' resources

Typical Core Topologies

- Economical approach: implement ring and then add links on as-needed basis
  - bandwidth allocation should consider failure modes
  - problem with traffic analysis that is based just on interface counters
    - Netflow-like techniques are necessary
- Typical topology of bigger cores: full mesh inner core + dual homed outer routers
- Other favorite topologies
  - star/ring/mesh combination
  - big double-star (Nx PE + 2x P)

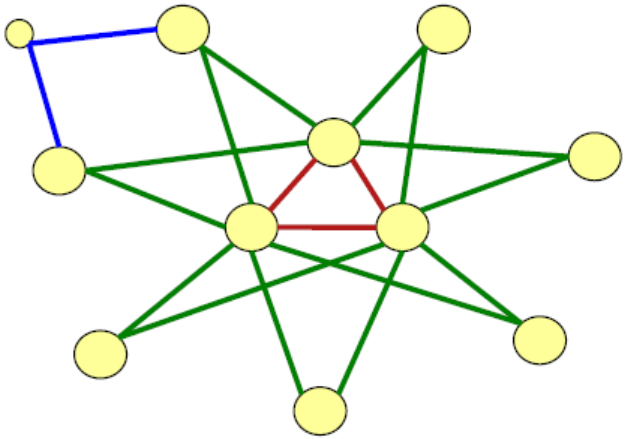
Favorite core topologies

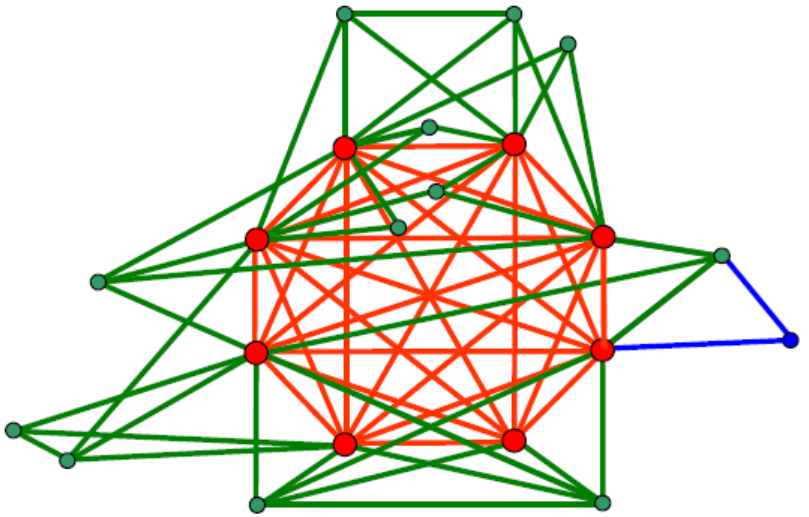
Two redundant mutually interconnected site border routers

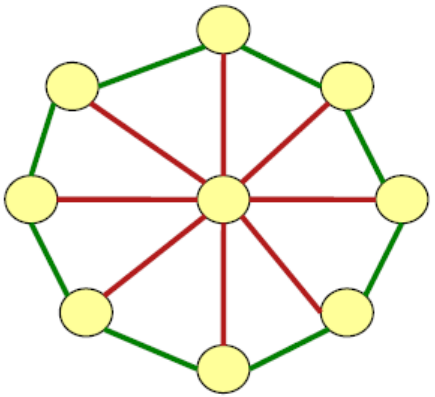


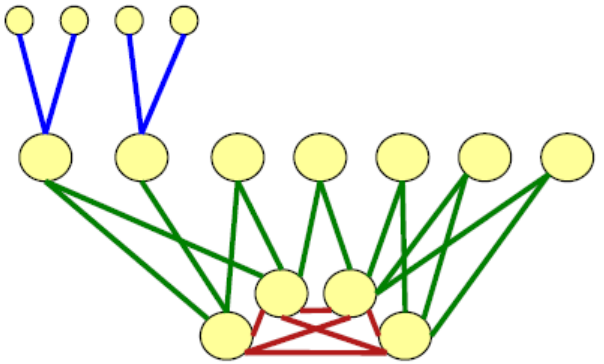
Core Architectures

real-life examples









Current Core Network Technologies

- IP over DWDM
- MPLS (MPLS-TE, FRR, ...)
  - Strict separation of P and PE routers is recommended to minimize configuration changes on backbone routers
- QoS-capable (DiffServ)