LAN Strategies for Performance and High Availability on HP-UX

Rick Petlin

System Support Engineer
Hewlett-Packard
rick.petlin@hp.com

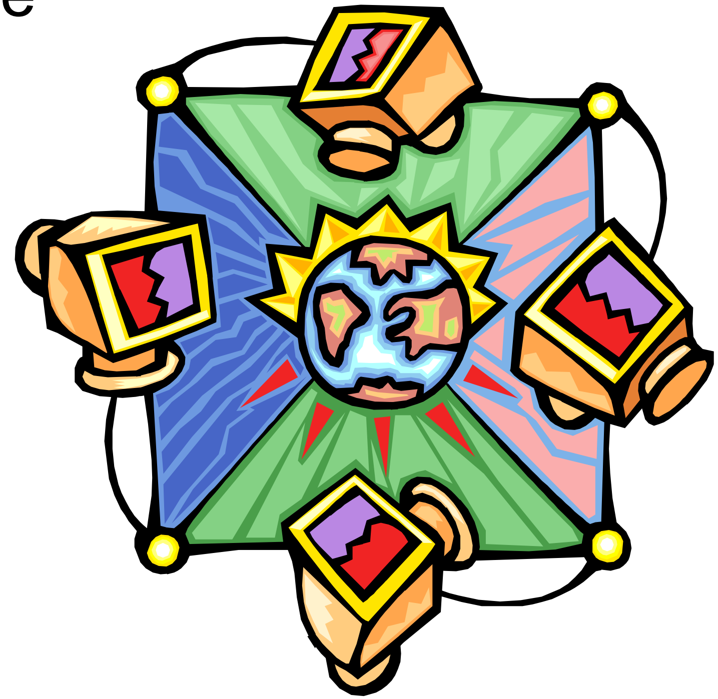
HP WORLD 2003
Solutions and Technology Conference & Expo

Purpose

- Provide an overview of technologies and implementations.
- Review factors that affect LAN network performance and availability.
- Examine network strategies and technologies that can improve HP-UX LAN performance and availability

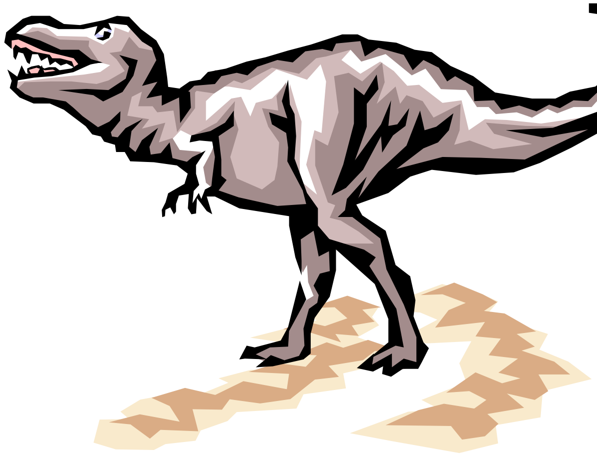
Agenda

- Ethernet and Legacy LAN Technologies
- LAN link throughput expectations
- Fast-Ethernet and Giga-bit Ethernet
- Jumbo Frames for Performance
- Trunking of LAN Links for Performance and HA
- Virtual LANs
- HyperFabric
- Futures…



Darwinism of LAN Technologies

- Survival of the fittest
  - Ethernet
  - Token-Ring
  - FDDI
  - ATM
  - 100VG



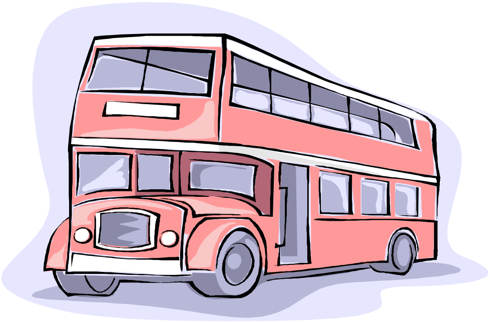
Link Speed & Throughput

Max throughput expectations:

- XXXX Base-X is not promise of XXXX Mbit/sec
  - System CPU speeds, I/O bus architecture and DMA rates impact throughput
  - Application/Transport driving the connection Competing network activity from other nodes
  - Switched versus shared connections and topologies
  - Network Link Trunking and Load Balancing
    - Auto Port Aggregation
- Wide throughput variance in specific tests

System Architecture

- I/O busses used in HP-UX systems
  - NIO/HPPB
  - EISA
  - HSC
  - PCI 1x, 2x, 4x
  - PCI-X
- I/O bus bridges
  - HCS-to-PCI
  - PCI-to-PCI



Gigabit Ethernet

- Two primary implementations
  - 1000Base-SX, fiber based
  - 1000Base-T, using common UTP cable
- Why pick one over the other?
- Cabling specifications

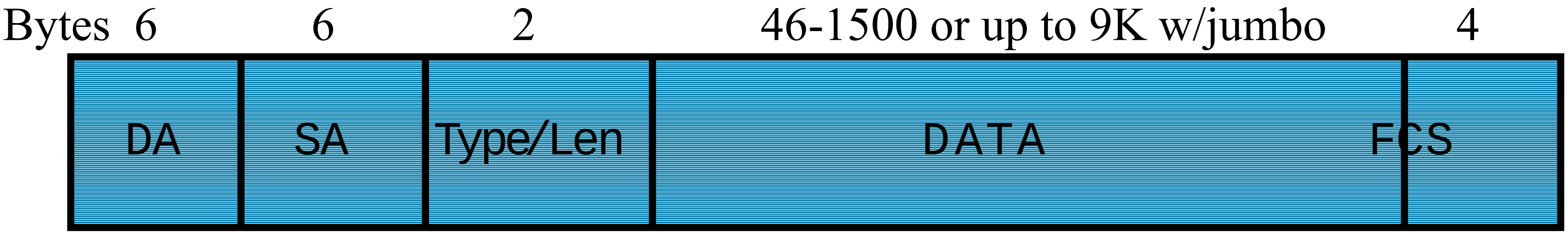
Gigabit Ethernet Performance

HPUX Systems

- gelan GbE driver and adapter
- igelan GbE driver and adapter
- Core I/O
- I/O Bus Considerations

Jumbo Frames on GbE

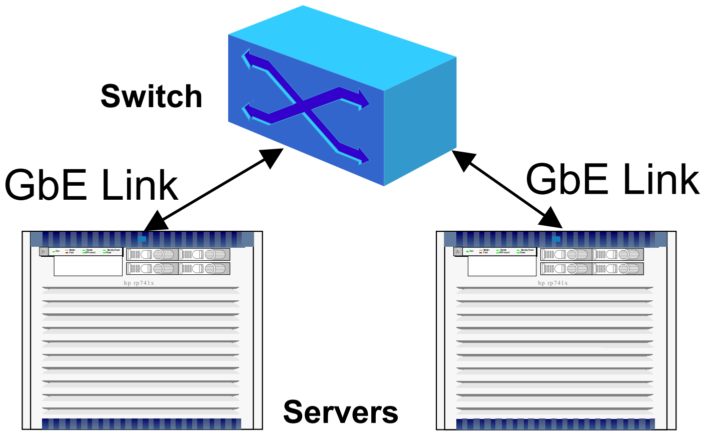
- Jumbo Frames up the Ethernet MTU from 1500 to 9000 Bytes
- Reduced CPU overhead
- Increase NIC throughput

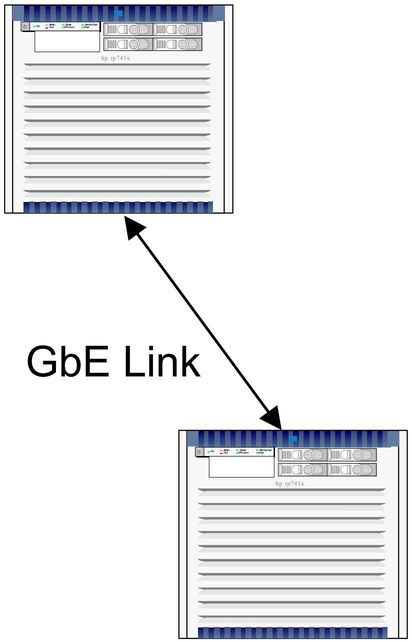


Bytes	6	6	2	46-1500 or up to 9K w/jumbo	4
	DA	SA	Type/Len	DATA	FCS

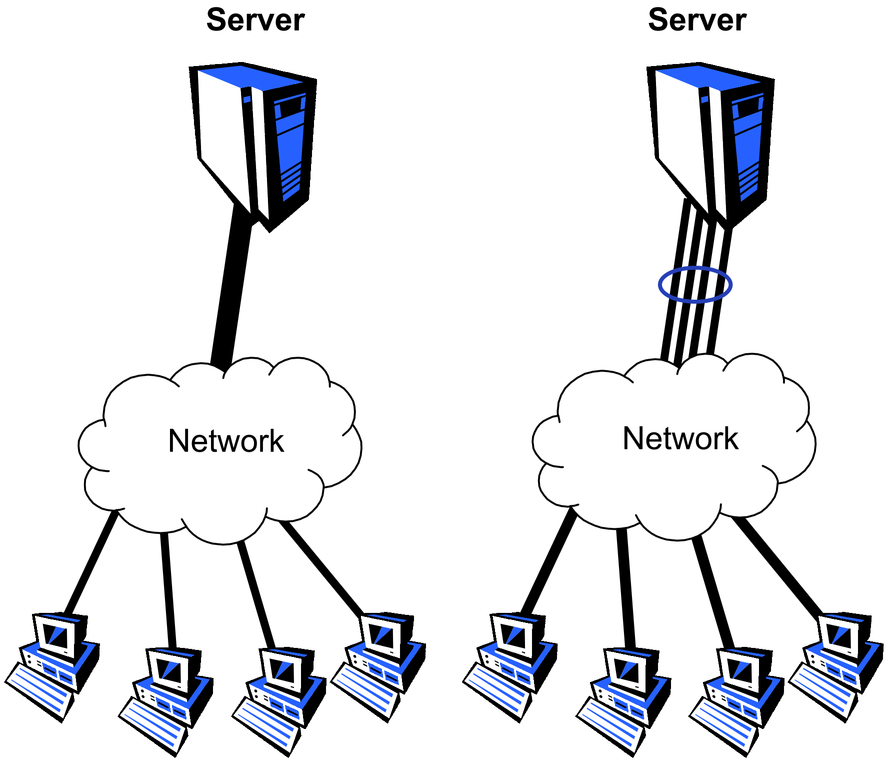
Deploying Jumbo Frames

- Point-to-point and switched configuration
- All devices in network need to support Jumbo Frames





Boosting Network and Server Access

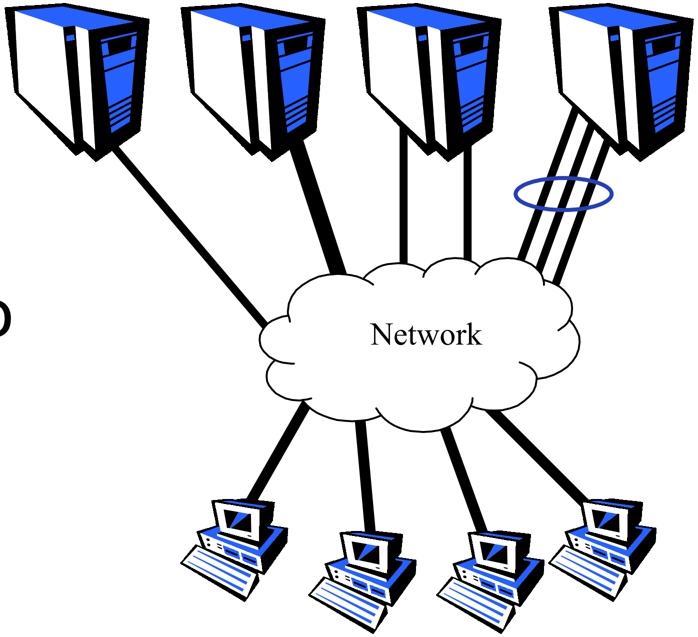


- Do I upsize to the next faster link? 10Mb->100Mb->1Gb
- Or, use multiple slower links?
- These are the same questions for end nodes as for the network infrastructure
- Design requirements

Adding Links to System

Common Methods for Adding Links

- Add multiple links w/ multiple IP addresses
- Add standby links and manually configure if needed
- Implement MC/Service Guard to manage standby links
- Deploy Link Aggregation technologies



Higher Throughput

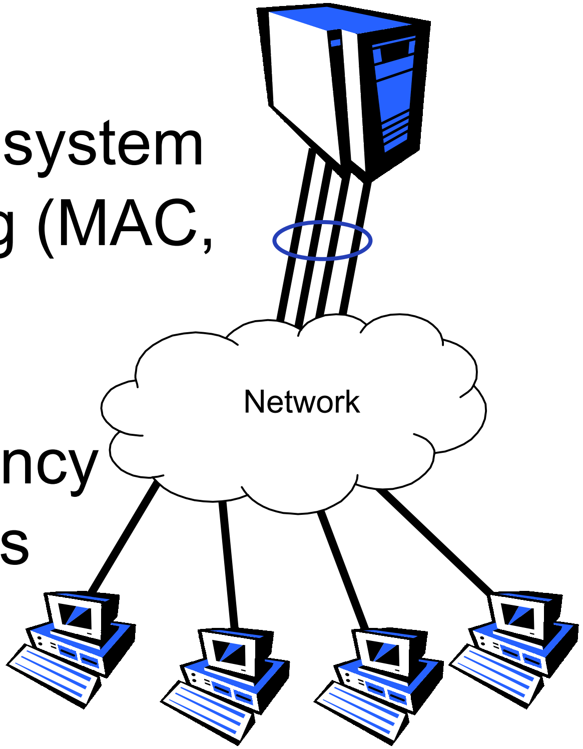
Why not just a higher speed link?

- Cost effective to trunk multiple lower speed links
- End-system may not be able to utilize the higher speed
  - System may not need a 10X boost in network speed
  - Available copper links are more pervasive for lower speed links
  - 10/100Mbps (and even GbE) NICs and switch ports are generally very cheap
- Protect investment in existing infrastructure
- Multiple links may provide higher availability and resiliency

Link Aggregation

Desirable Features

- Transparent, Available and Fast
- A single network presence
  - minimize impact of multiple links in a system
  - Provide transparent address mapping (MAC, IP)
- Automatic link fail-over
  - keep link up and running via redundancy
  - provide transparency from link failures
- Active load balancing
  - utilize invested resources
  - maximize available performance



Link Aggregation

Link Layer

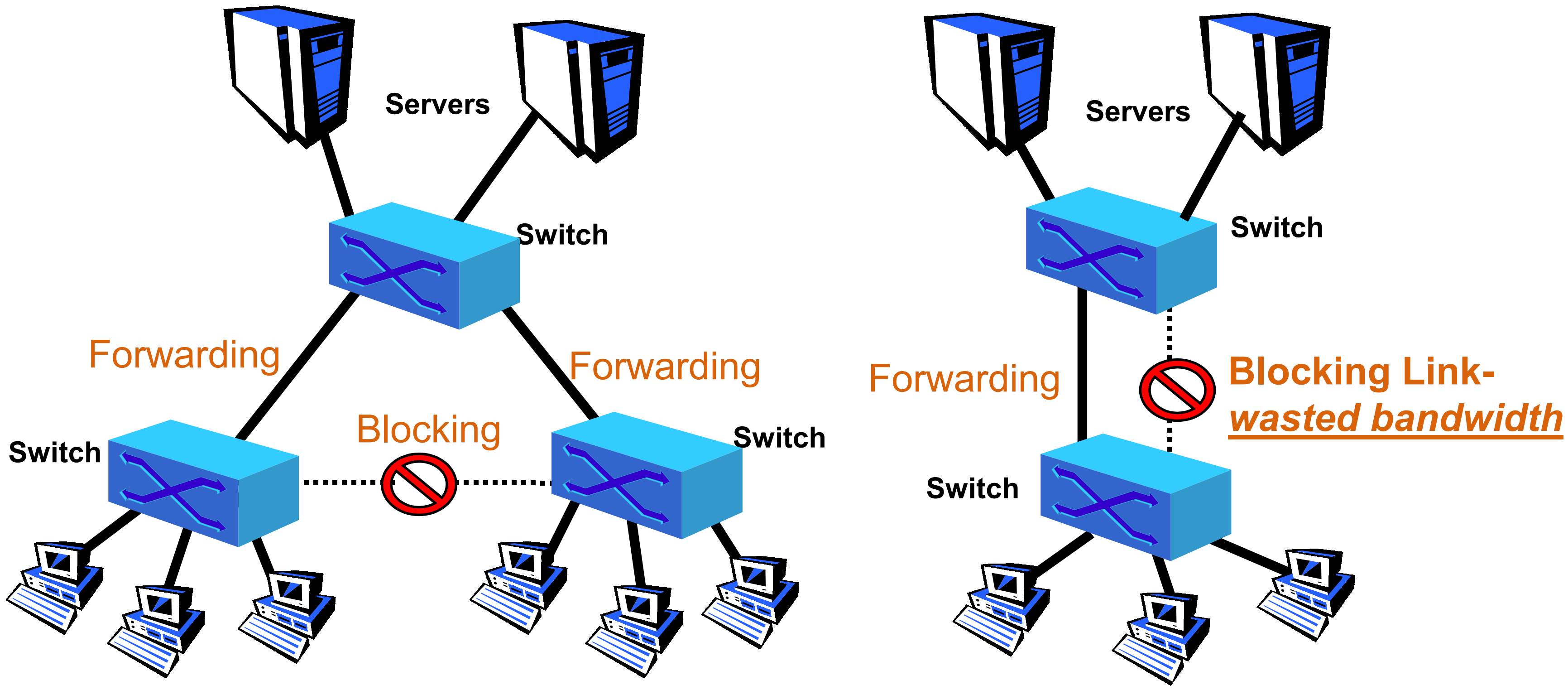
- Layer 1 solution requires new PHY and MAC chips
- Layer 3 solution are not a transparent to end-stations and switches/router
- Layer 4 solutions require even more complexity then layer 3

➢ A Layer 2 implementations maintain MAC and IP addressing and requires no new hardware*.

A single network presence

Non-Aggregate Link Fail-over

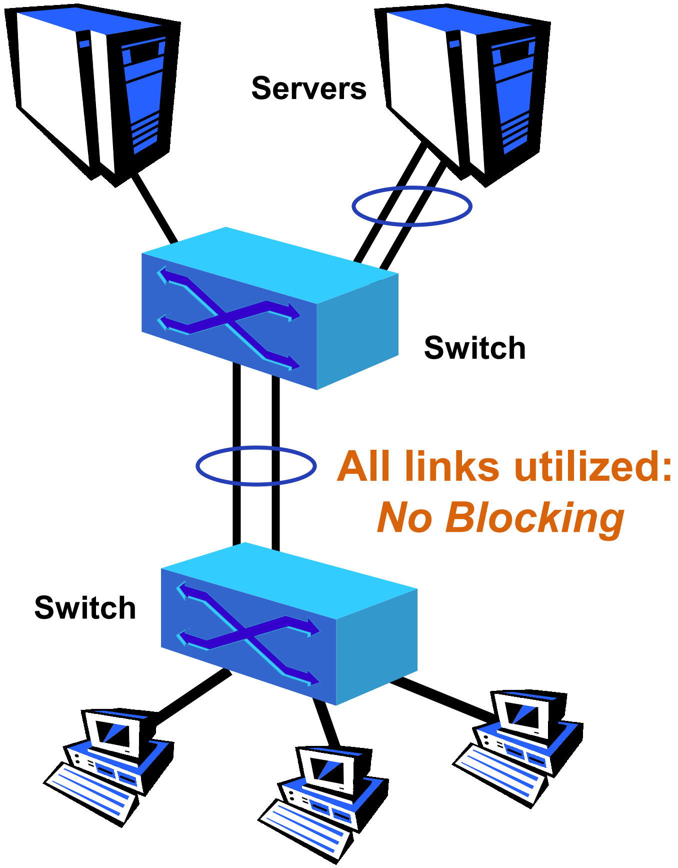
- Review of Spanning Tree



Using 802.1d Spanning Tree Protocol provides fail-over protection and prevent loops but may waste available bandwidth

Automatic Link Fail-over

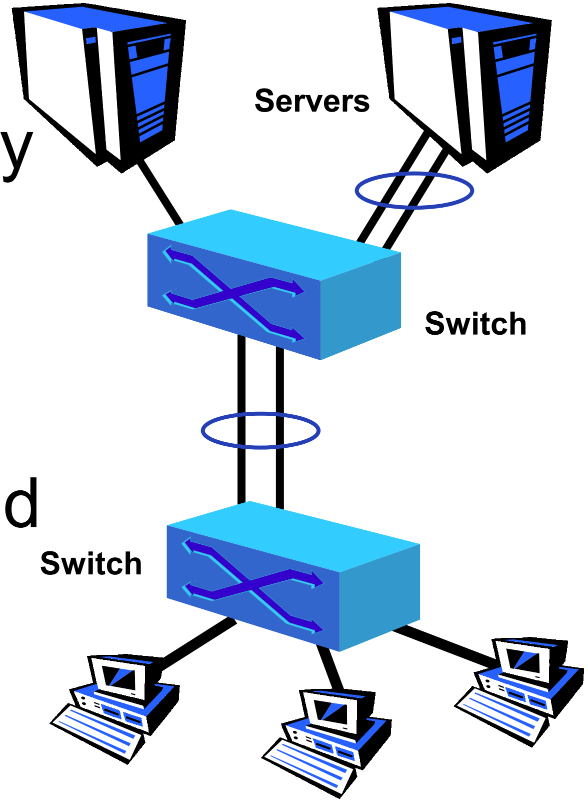
- Link Aggregation does *not* use Spanning Tree Protocol
- Link failure/recovery faster then Spanning Tree Protocol
- Multiple links can be utilized
- Aggregated links appear to be one physical link
- An individual link failure can be transparent
- Link Aggregate operational until all links in aggregate fail



All links utilized:
No Blocking

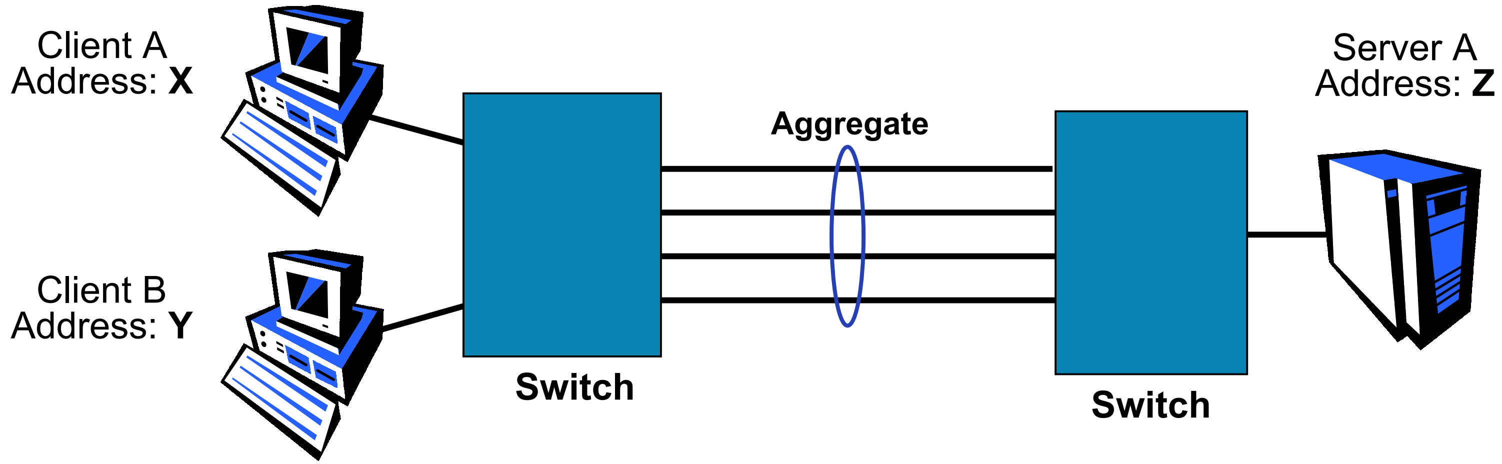
Load Balancing the Aggregate

- Layer 2 implementation requires distribution of complete frames
- Load distribution algorithm generally based on MAC addresses
- Other distribution algorithms useful depending on configurations
- Distribution attempts to balance load
- Must not mis-order frames
- Must not send duplicate frames



Load Balancing

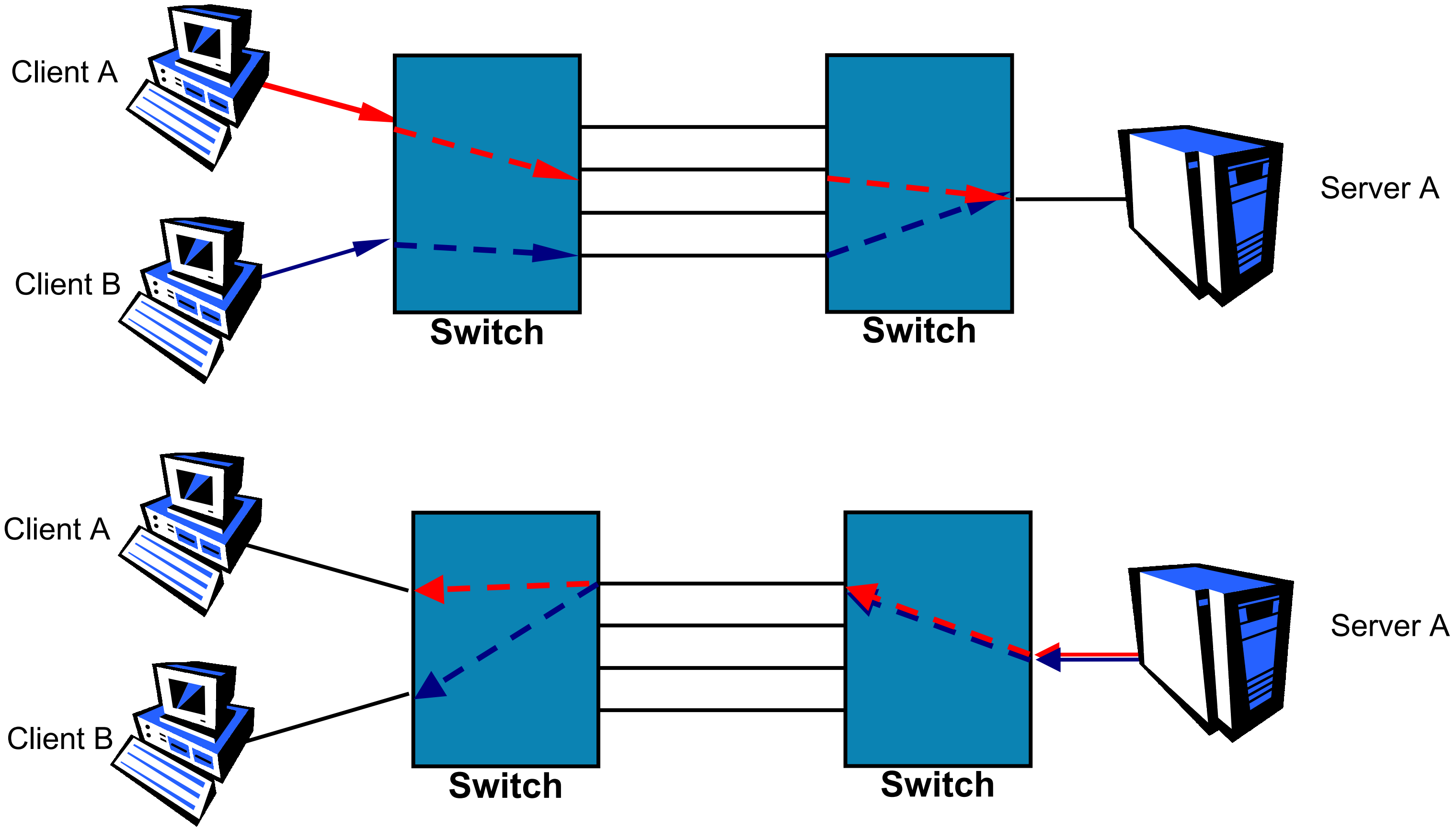
- Switch-to-Switch



- **Source Address (SA Only),** which means that the conversation is assigned using only the source address
- **Destination Address (DA Only),** which means that the conversation is assigned using only the destination address
- **Source Address/Destination Address (SA/DA),** which means that the conversation is assigned using the combination of the two addresses

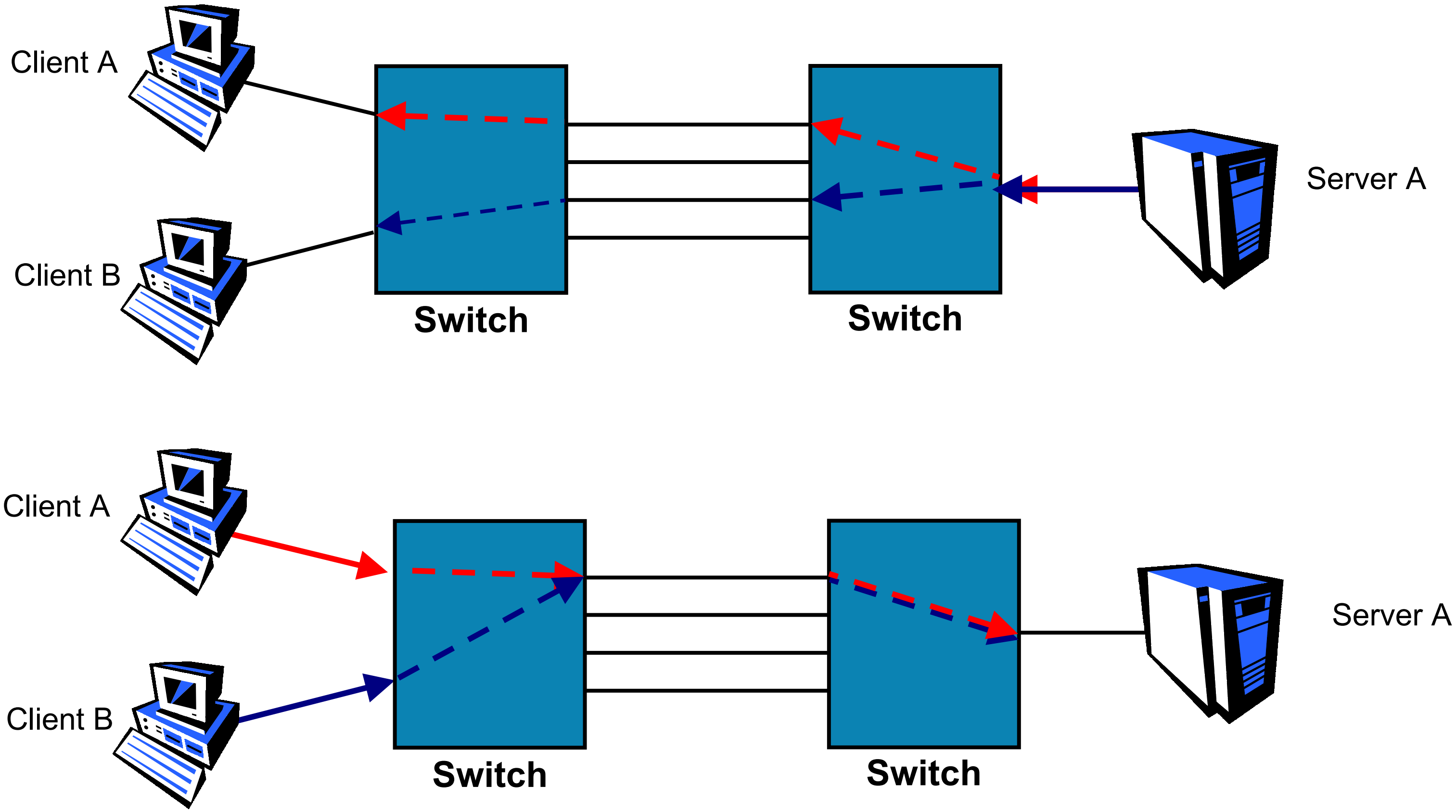
Source Address (SA) Only

- SA Only provides “one-way” load balancing



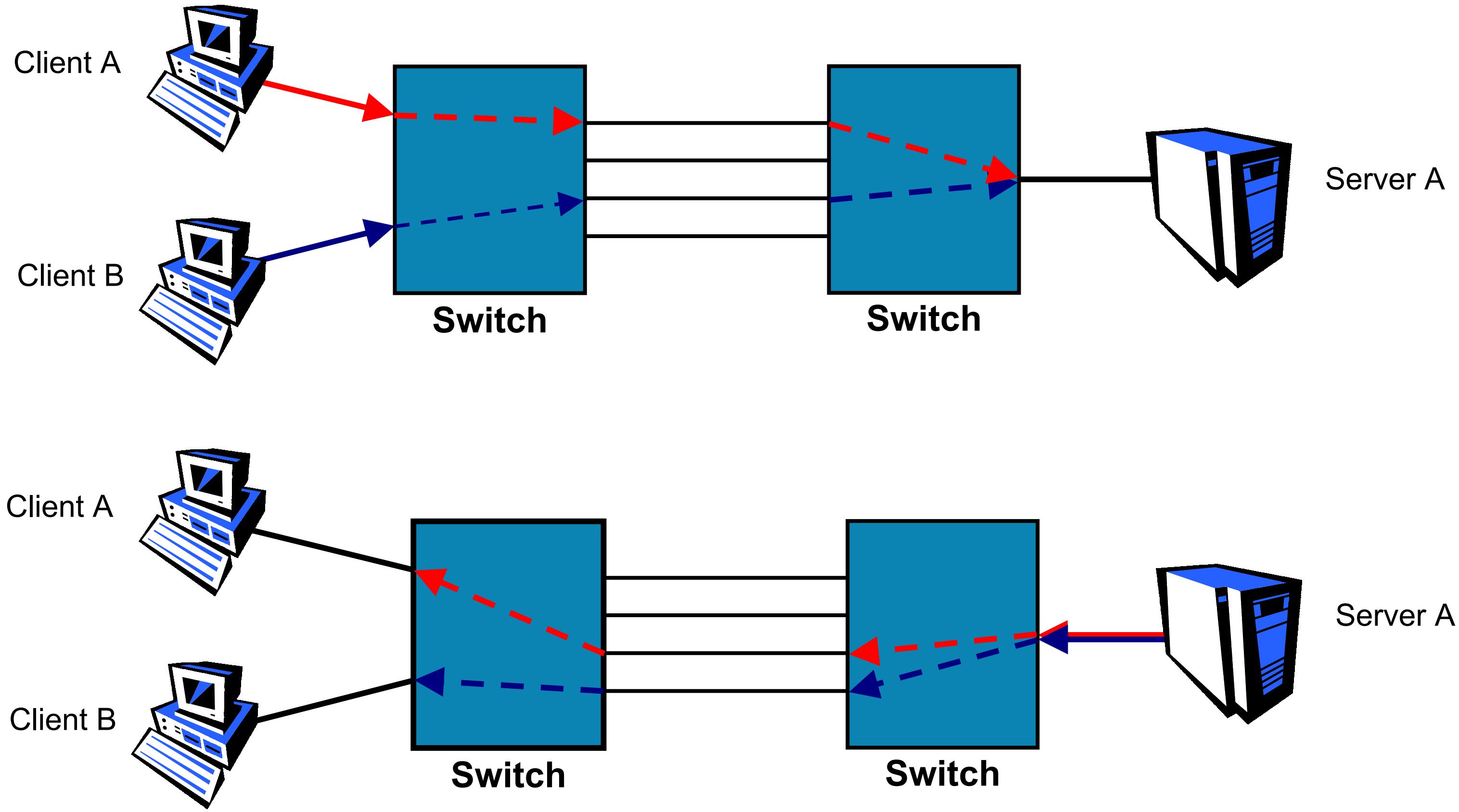
Destination Address (DA) Only

- DA Only provides "one-way" load balancing



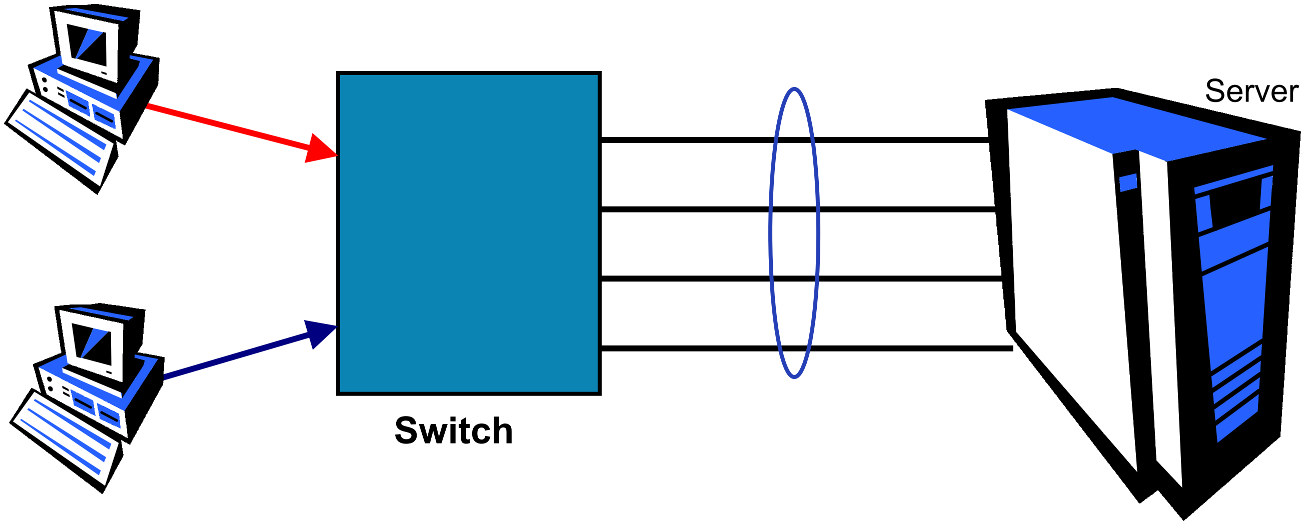
SA/DA Load Balancing Algorithm

- SA/DA Bi-directional load balancing



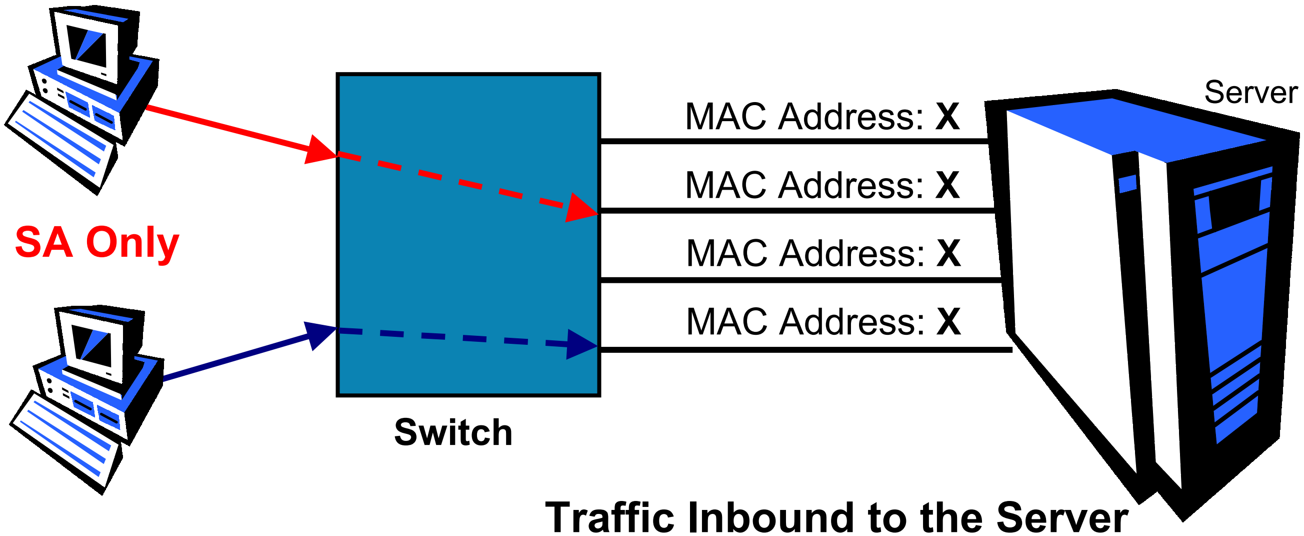
Load Balancing: Switch-to-Server

- Additional considerations



Load Balancing: Switch-to-Server

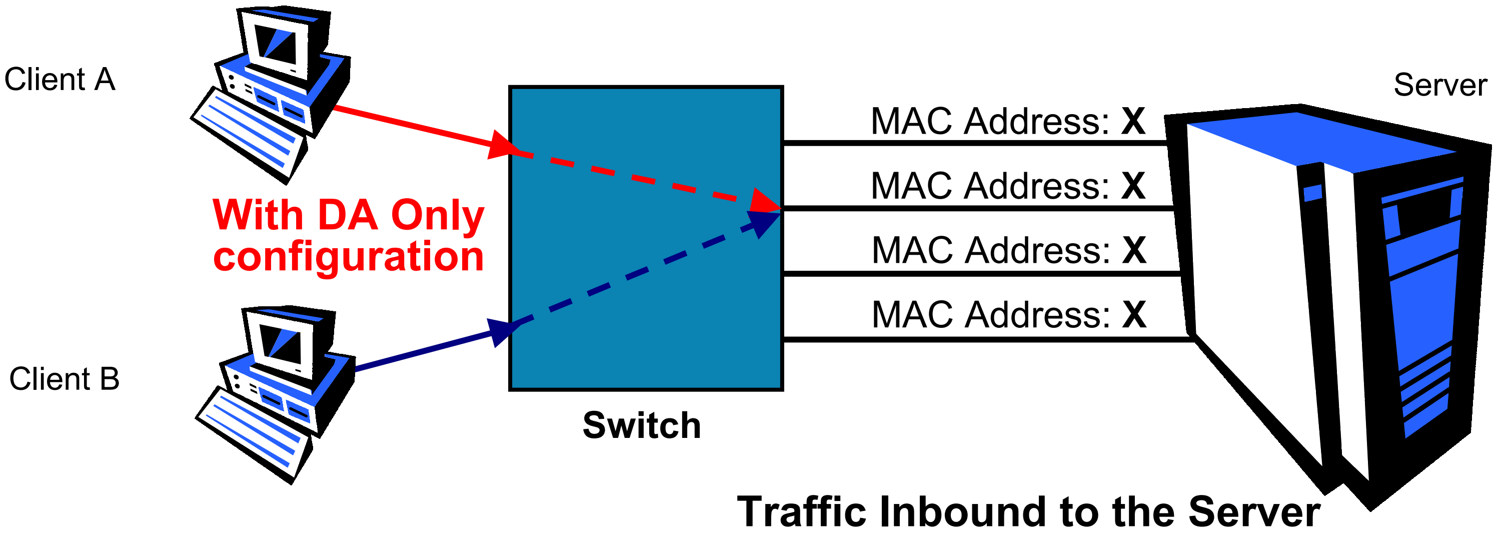
- Port Aggregation Links with Single MAC Address
  - *SA only distribution algorithm*



Traffic Inbound to the Server
SA Only: Balanced

Load Balancing: Switch-to-Server

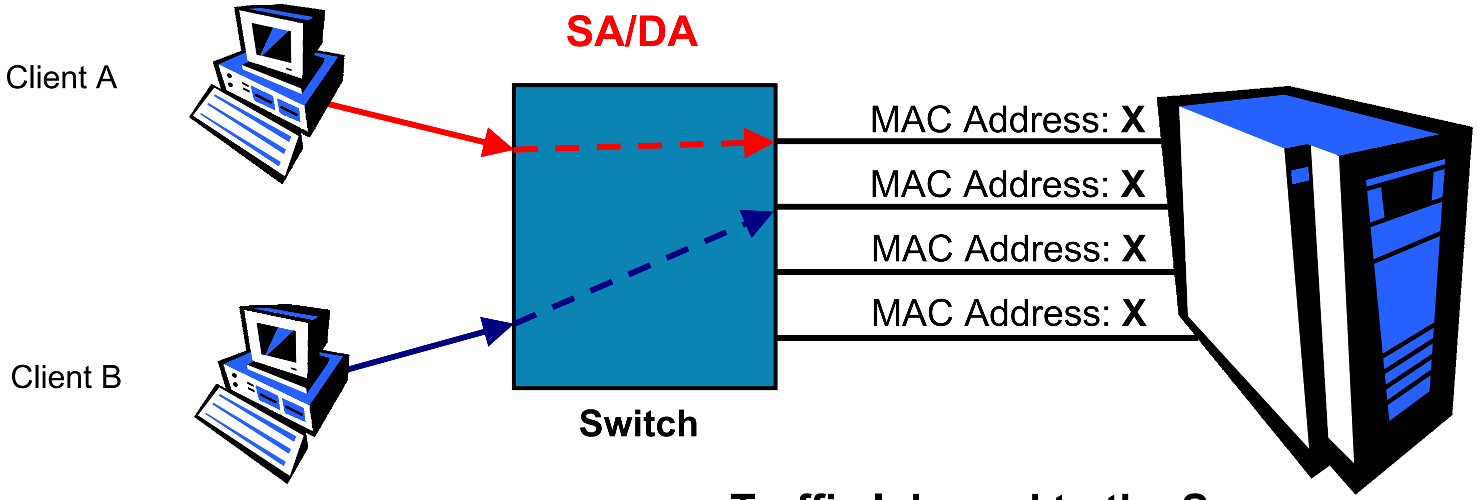
- Port Aggregation Links with Single MAC Address
  - *DA only distribution algorithm*



SA Only: Balanced
DA Only: ***Unbalanced***

Load Balancing: Switch-to-Server

- Port Aggregation Links with Single MAC Address
  - *SA/DA distribution algorithm*



Client A

Client B

SA/DA

Switch

MAC Address: **X**

MAC Address: **X**

MAC Address: **X**

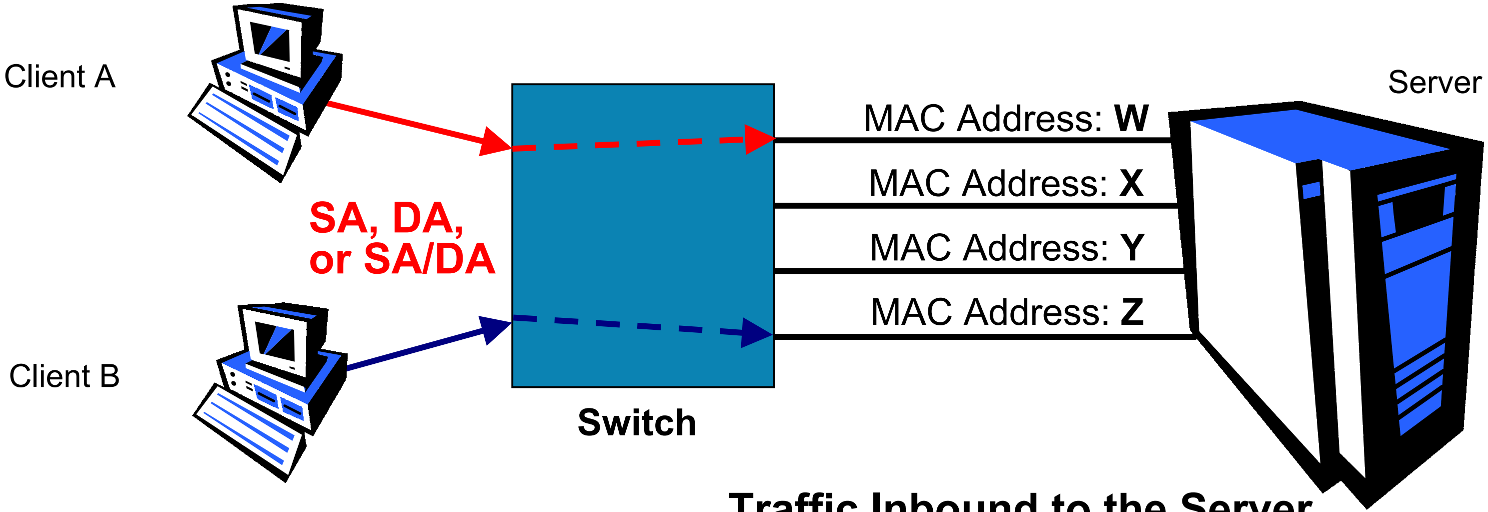
MAC Address: **X**

Traffic Inbound to the Server

SA Only:	Balanced
DA Only:	Unbalanced
SA/DA:	Balanced

Load Balancing: Switch-to-Server

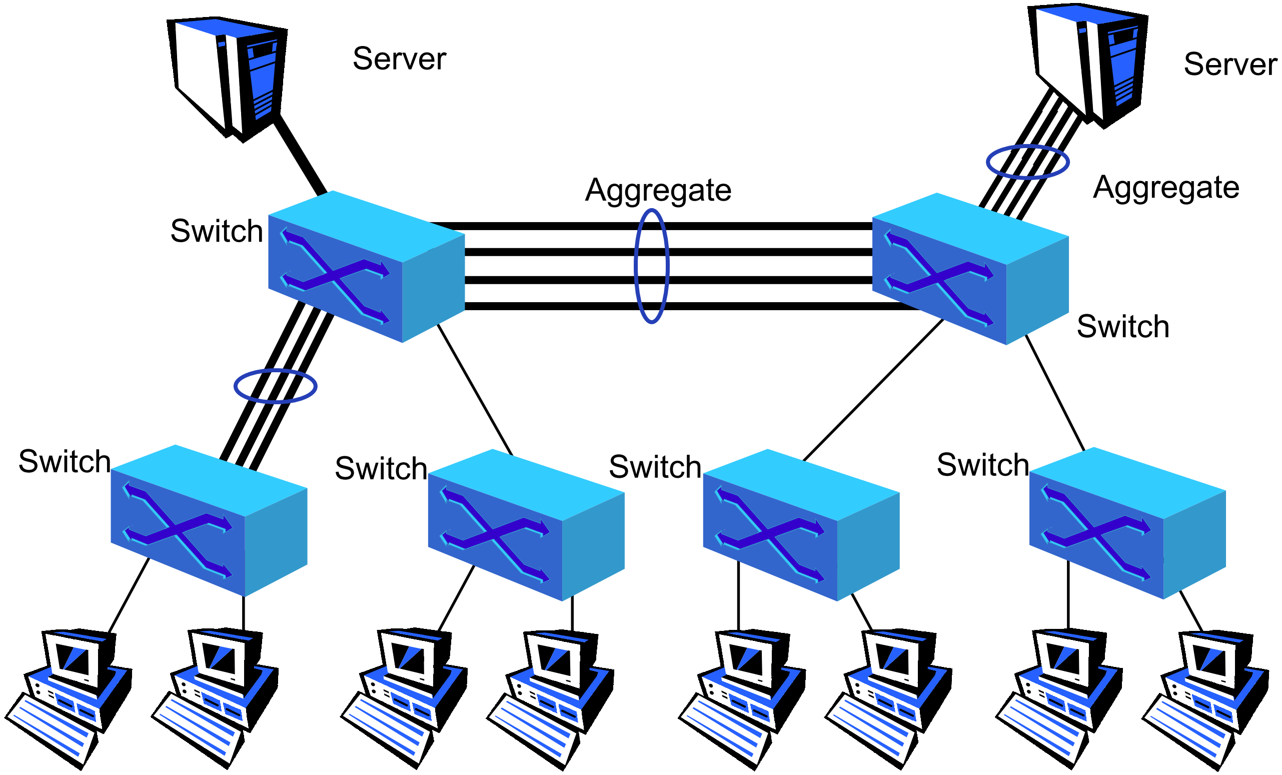
- Port Aggregation Links: Multiple MAC Addresses
  - Unique hybrid configuration with multiple MACs



Traffic Inbound to the Server

SA Only: Balanced
DA Only: Balanced
SA/DA: Balanced

Link Aggregation: Big Picture

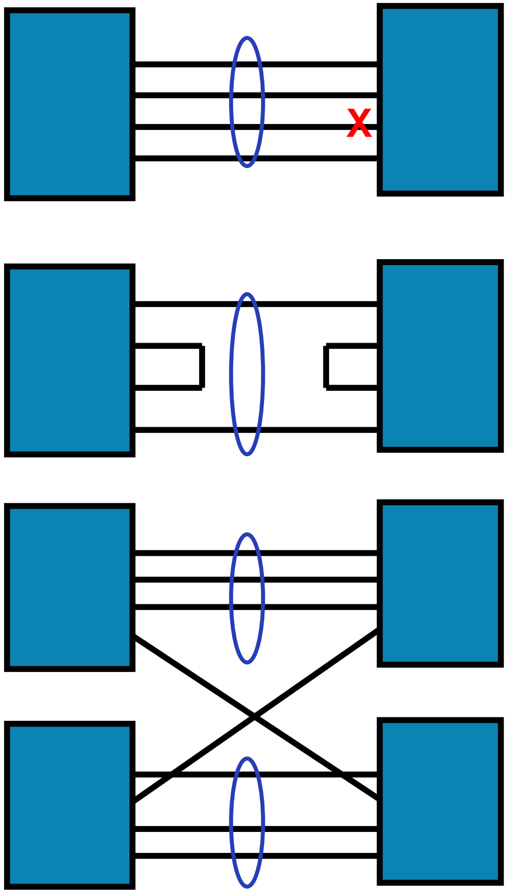


Automatic Link Aggregation

Protocol for Reliability

- Link down is not enough
- Loop-backed connection error
- Split trunk configuration error

- Automatic Configuration
  - Key cost savings



Port Aggregation Protocol / Cisco Fast EtherChannel®

- EtherChannel name first used by Kalpana to describe their 10Mbit trunk product.
- PAgP, proprietary protocol developed by Cisco
- Provide automatic trunk configuration
- Typically limited to 4 links per aggregate
- Implemented in various Cisco product families
- Implemented in non-Cisco switches and link products
- Implemented on HP-UX and HP ProCurve Switches

Link Aggregation Control Protocol

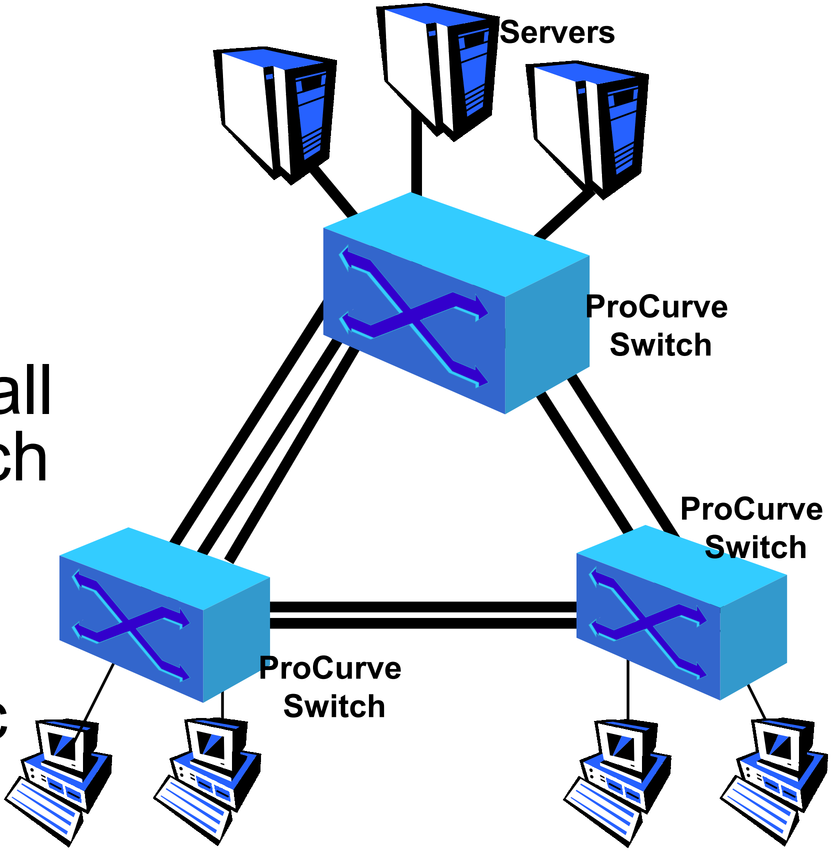
Feature set very much like Cisco PAgP/FEC

- Ratified 802.3ad standard in 2000
- Number of links in aggregate not limited by standard
- Switch vendors committed to support standard
- Supported on HP-UX 11.0 & 11i (11.11)

HP WORLD 2003
Solutions and Technology Conference & Expo

HP Switch-to-Switch Meshing

- Proprietary- HP ProCurve Switches
- HP Switch Meshing is alternative to other Link Aggregation techniques
- Switch Meshing aggregates all link and switches in the Switch Mesh
- Spanning tree is not used
- Switch selects the best traffic path
- ProCurve Switches also support FEC/PAgP aggregates links



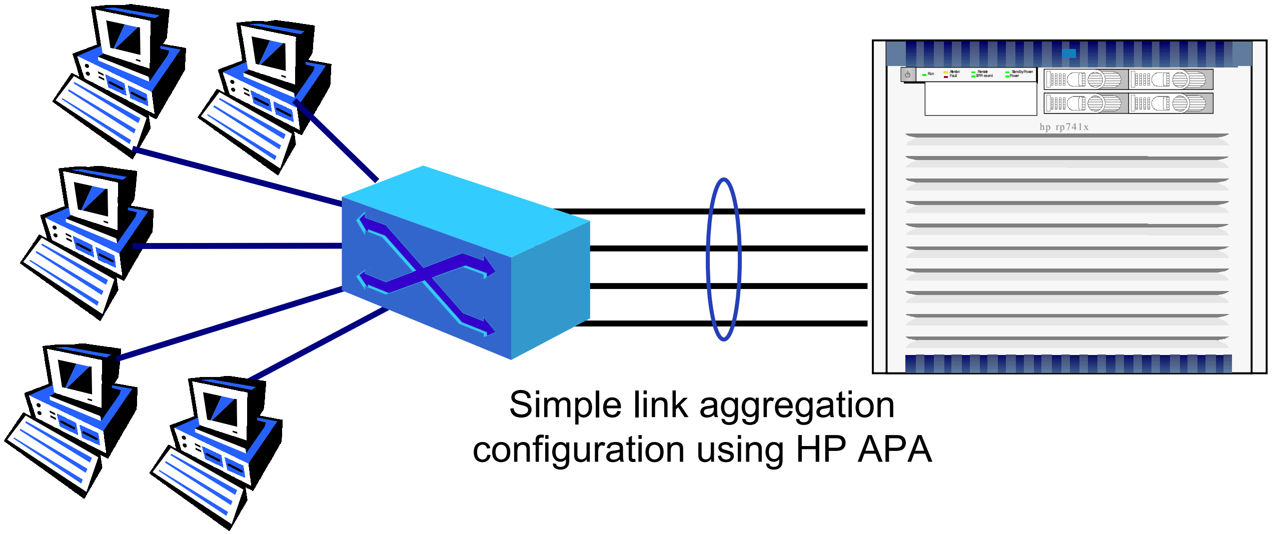
Switch to Server Vendors and Products

3Com	EtherLink Server
Adaptec	Duralink ®
Intel	Adaptive Load Balancing
Sun	Sun Trunking
IBM	EtherPipe
Hewlett-Packard	HP Auto Port Aggregation

Auto Port Aggregation

HP-UX's Link Aggregate Implementation

- Aggregates multiple physical LAN links into one logical link
- APA includes both load sharing link aggregates as well as fail-over aggregates.



Simple link aggregation configuration using HP APA

Benefits of HP APA

- Bandwidth Scalability
- High Availability: A link aggregate will continue to operate as long as there is at least one port operating.
- Load Balancing: MAC-based, IP-based, CPU-based, and TCP/UDP port-based distribution
- Single MAC address: HP APA link aggregate share a single, logical MAC address
- Flexibility: ports can be aggregated to achieve higher performance
- Investment Protection: leveraging existing end-stations, management tools and training.

APA Features

Provides Flexible Configuration Options

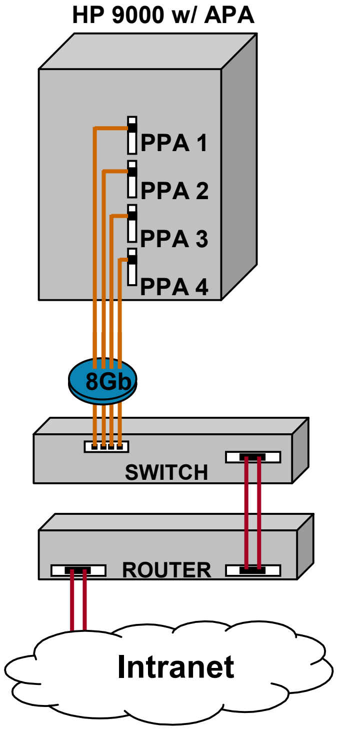
- HP Auto-Port Aggregation fully inter-operates with Cisco and HP switches and routers, while maintaining compatibility with other vendors' devices
- HP Auto-Port Aggregation provides the right load balancing algorithm for server's environment
- Automatic discovery and configuration of Aggregates
- Higher availability in conjunction with MC/Service Guard
- Improves manageability through
  - automatic detection of LAN failures
  - automatic traffic redirection in case of failed channel

APA Link Aggregation Modes

- Four link aggregation configuration modes:
- PAgP (Port Aggregation Protocol) is Cisco’s developed protocol that supports automatic configuration of aggregates of Ethernet or Gigabit Ethernet links with up to 4 links per aggregate
- LACP (Link Aggregation Control Protocol) is the IEEE 802.3ad standard for automatic configuration with up to 32 links per aggregate
- Manual configuration of link aggregates for other vendors’ switches that do not support PAgP or LACP
- LAN Monitor for MC/Service Guard like fail-over of aggregates, or individual links, including Ethernet, Token-ring, and FDDI

APA Manual Mode

- User manually configures the Server and Switch Ports to be aggregated.
- Caution MUST be used when configuring manual Link Aggregates as there are limited diagnostic checks to verify the configuration

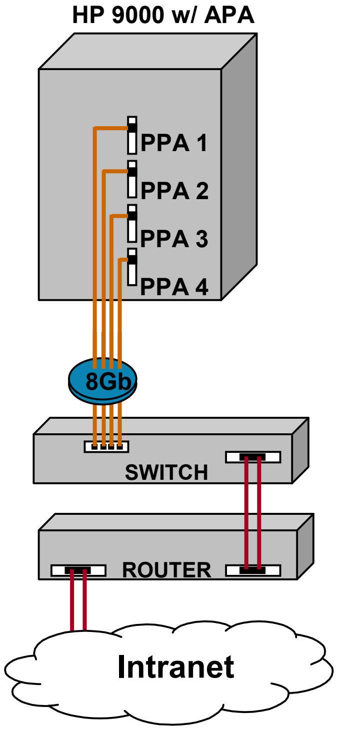


1. Select the ports on the Server to be aggregated
2. Use APA SAM/CLI interface to aggregate ports. For Example: lanadmin -X -a 1 2 3 4 100.
3. Select the ports on the Switch to be aggregated
4. Use the Switch GUI/CLI to aggregate the selected ports (See appropriate switch documentation).

APA Automatic Modes

PAgP & LACP

- Use automatic protocols to configure the Server and Switch Ports.
- The protocols prevents illegal configuration of invalid Link Aggregations.



1. Select the ports on the Server to be aggregated

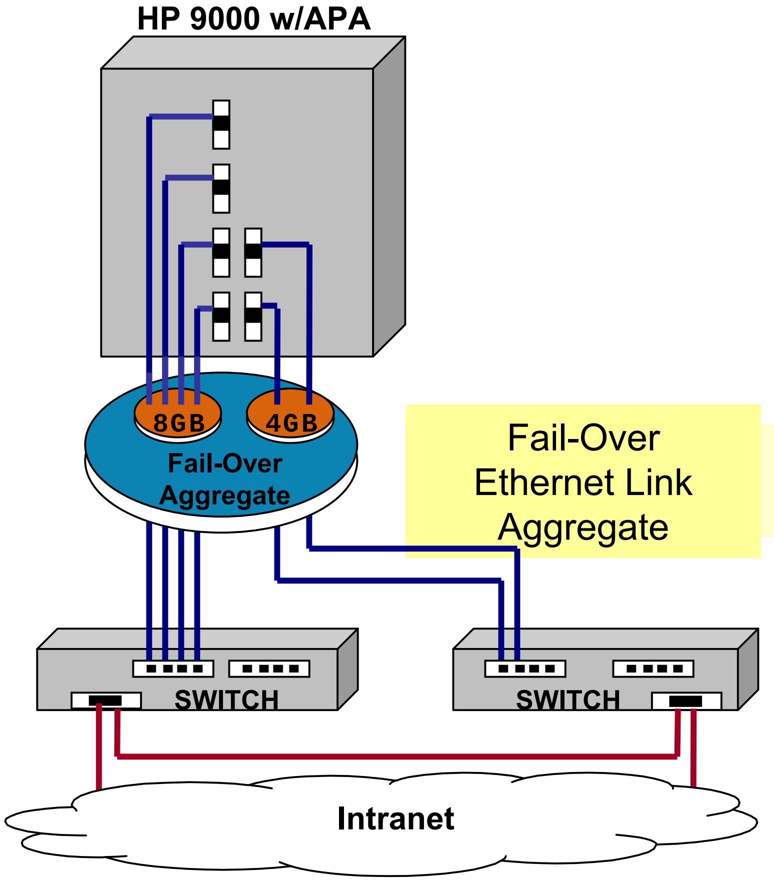
2. Use APA SAM/CLI to turn on the appropriate protocol (PAgP is the default).

3. Select the ports on the Switch to be aggregated

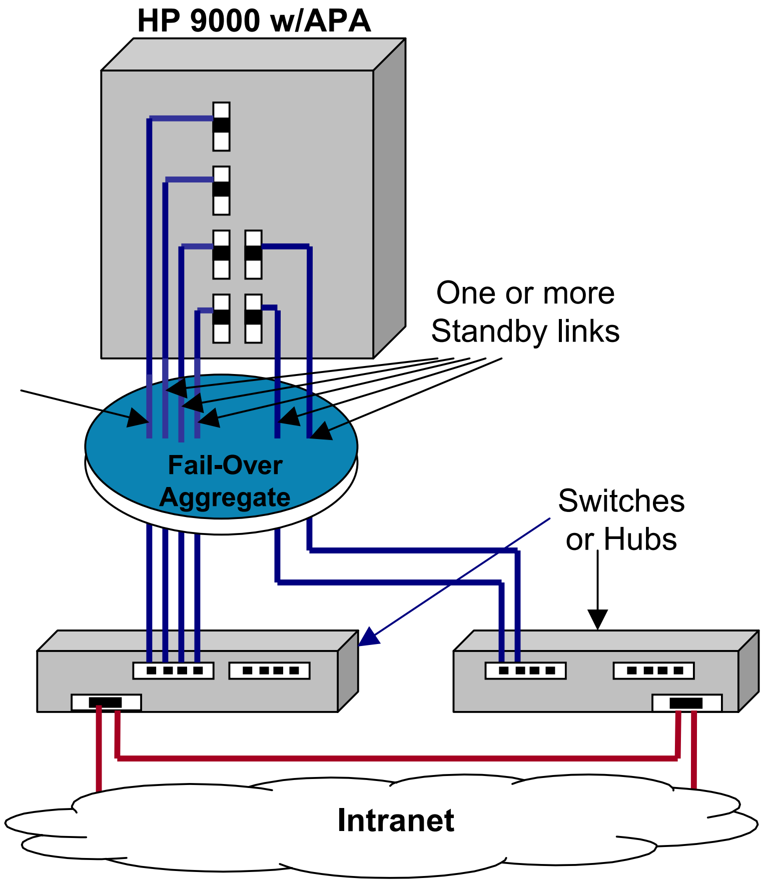
4. Use the Switch GUI/CLI to turn on the appropriate protocol. The Switch and Server protocols must be the same in order for automatic aggregation to occur.

APA LAN Monitor for 11.0/11i

- LAN Monitor introduced June 2000
- Simple, low cost single system HA link fail-over solution w/o MC/Service Guard complexity and expense
- LAN Monitor Aggregates primary and standby links can be made of individual links or logical link aggregates*



APA for 10.20



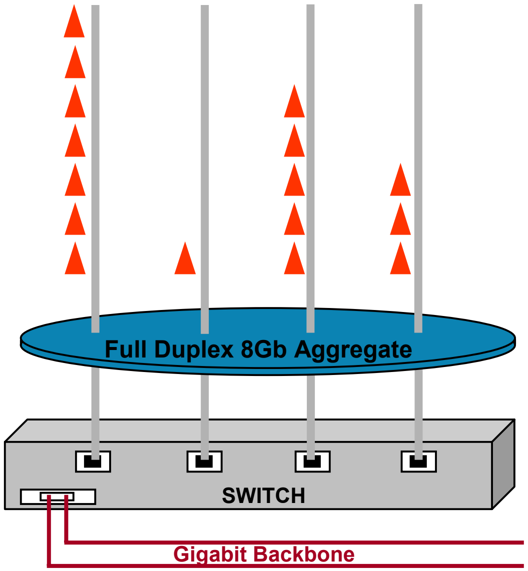
- LAN Monitor mode only
- Primary and standby are single links only (not link aggregates)
- Simple, low cost Single System HA network link solution
- Link fail-over w/o MC/Service Guard complexity and expense
- Support for 100Mb, 1Gb, FDDI, Token-Ring
- Supports all hubs and switches

APA Load Balancing

Provides the Right Load Balancing for the Environment

- MAC Based - Uses the least significant 8 bits of the destination MAC address
- IP Based - Uses the least significant 8 bits of the destination IP address
- CPU Based - Uses the processor index to determined which link to transmit frame out
- TCP/UDP Port Based - Uses destination and source port to distribute frames.

MAC Based Load Balancing

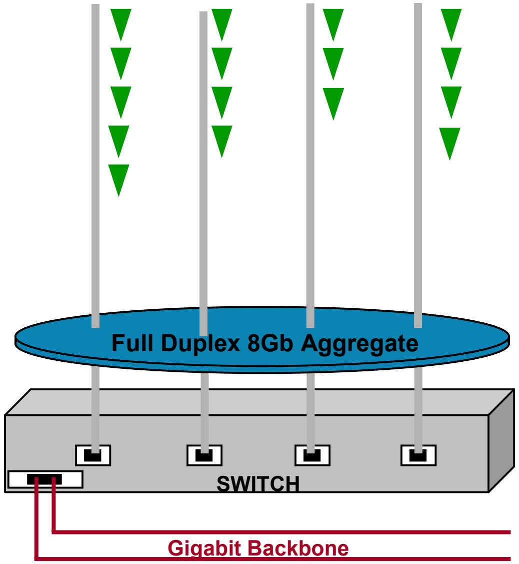


FROM SWITCH
Sample MAC Load Balancing

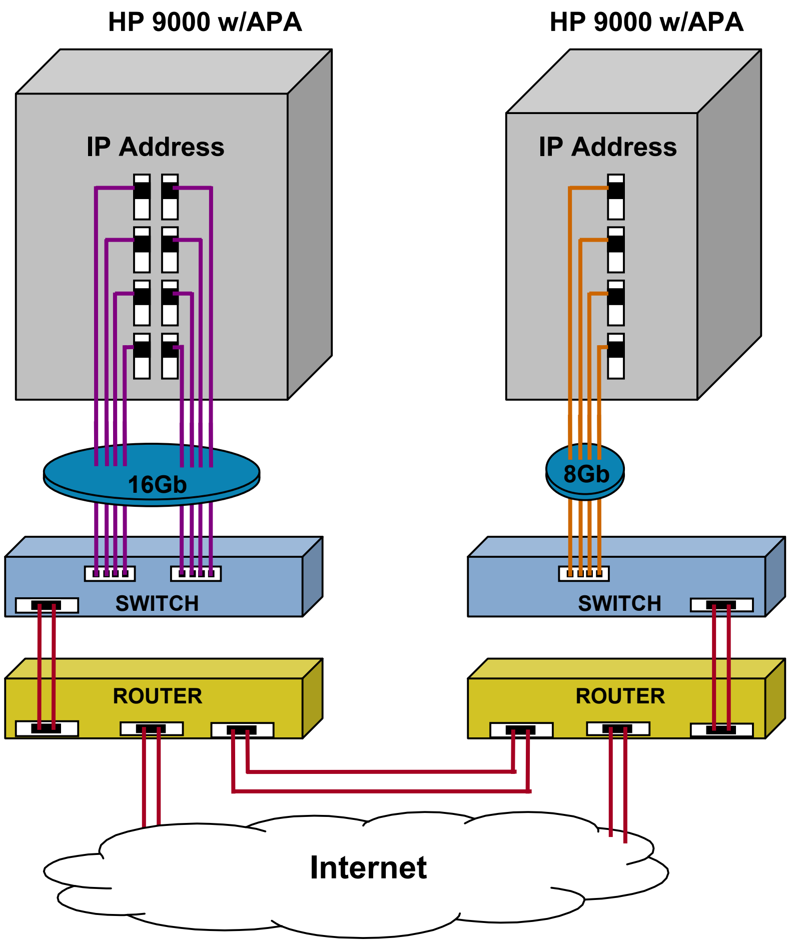
A typical switch hashes the 2 least significant bits of each address for load balancing with limited results.

TO SWITCH
Sample MAC Load Balancing

HP APA software hashes the 8 least significant bits of each address, for switcher-style load balancing, using a data table with 256 options.



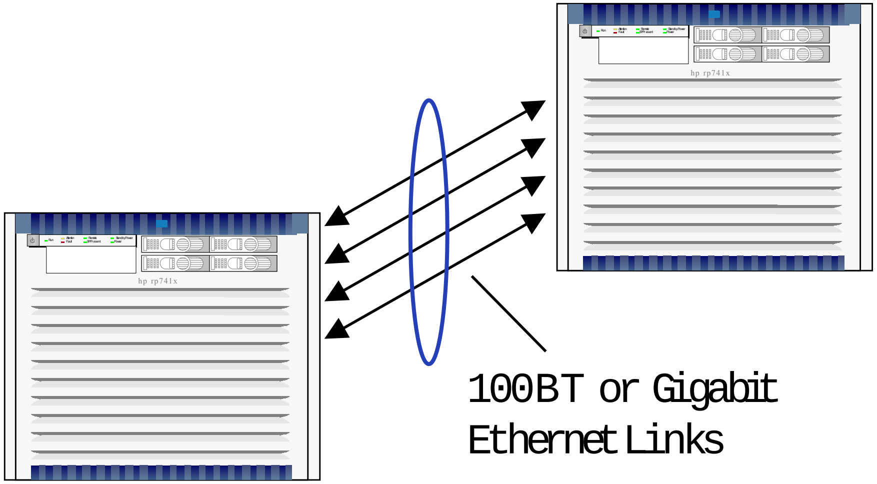
IP Based Load Balancing

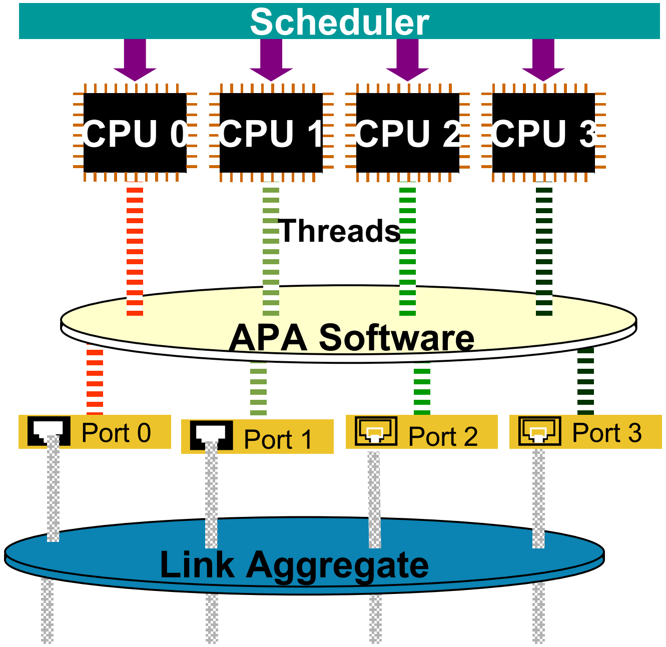


- IP address mechanism
- All links are active and load balanced
- Useful when connections are to non-local network clients

CPU Base Load Balancing

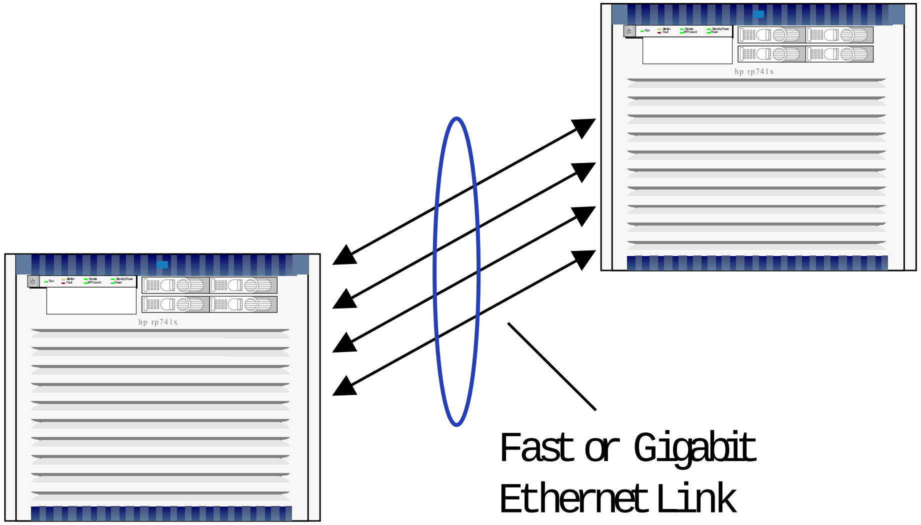
- Enables direct Server to Server connections for Backups, Data Warehousing, etc.
- Requires the use of Processor Affinity





TCP/UDP Port Based LB

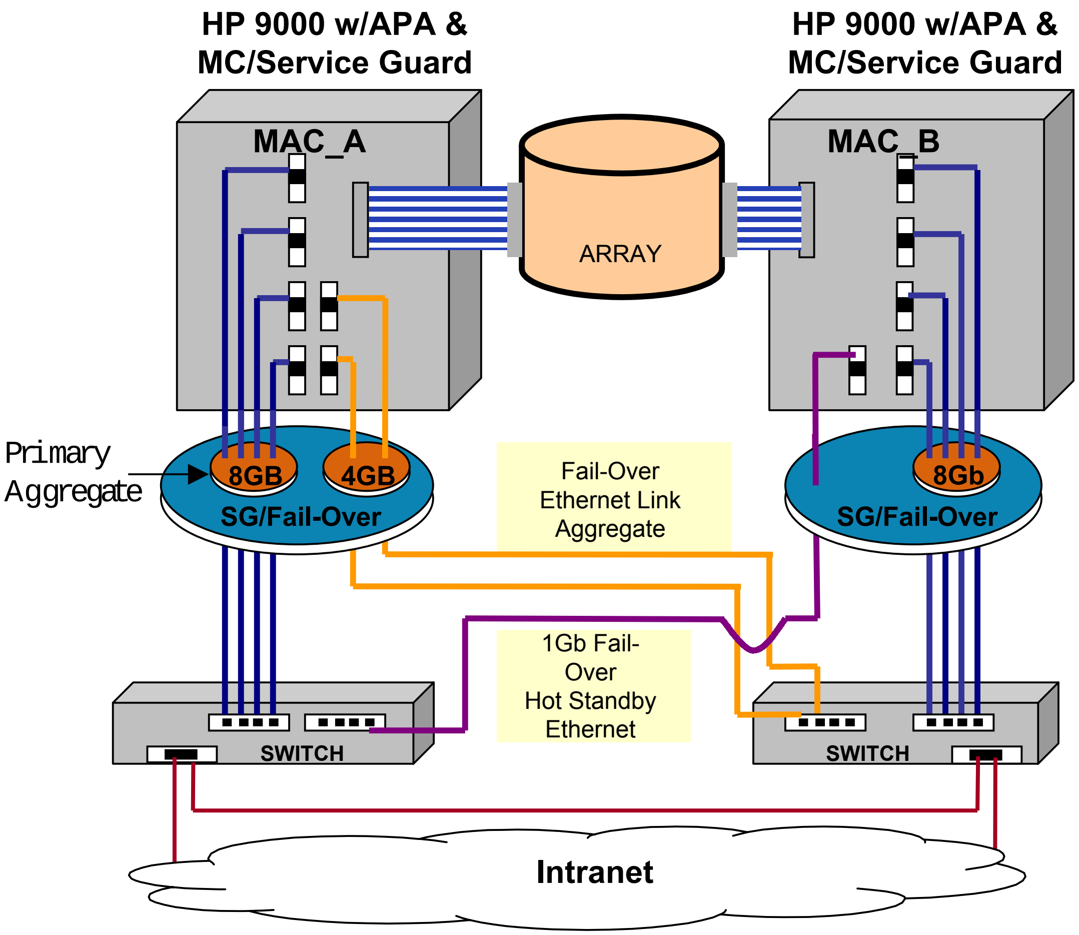
- Better mechanism then CPU base load distribution
- Can be used other then back-to-back configurations



APA Hot-Standby Mode

- Hot-standby Mode provides fail-over protection
- Hot Standby Mode switch configures link aggregate to only sends data out one link.
  - Therefore, no load balancing with Hot-Standby
- Fail-over configuration provided by Hot-Standby Mode *or* LAN Monitor

APA with ServiceGuard Fail-over



HP 9000 w/APA & MC/Service Guard
HP 9000 w/APA & MC/Service Guard
MAC_A
MAC_B
ARRAY
Primary Aggregate
8GB
4GB
SG/Fail-Over
8Gb
SG/Fail-Over
Fail-Over Ethernet Link Aggregate
1Gb Fail-Over Hot Standby Ethernet
SWITCH
SWITCH
Intranet

- APA integration with ServiceGuard
- Support of link aggregates
- LAN Monitor not currently supported

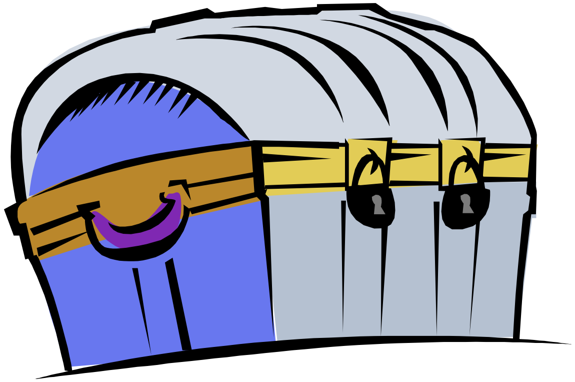
APA Summary Features

- Very good solution for one-to-many or many-to-many connections
- Incremental bandwidth
- Protect investment
- LAN/Monitor is excellent for link failover

Trunking Issues…

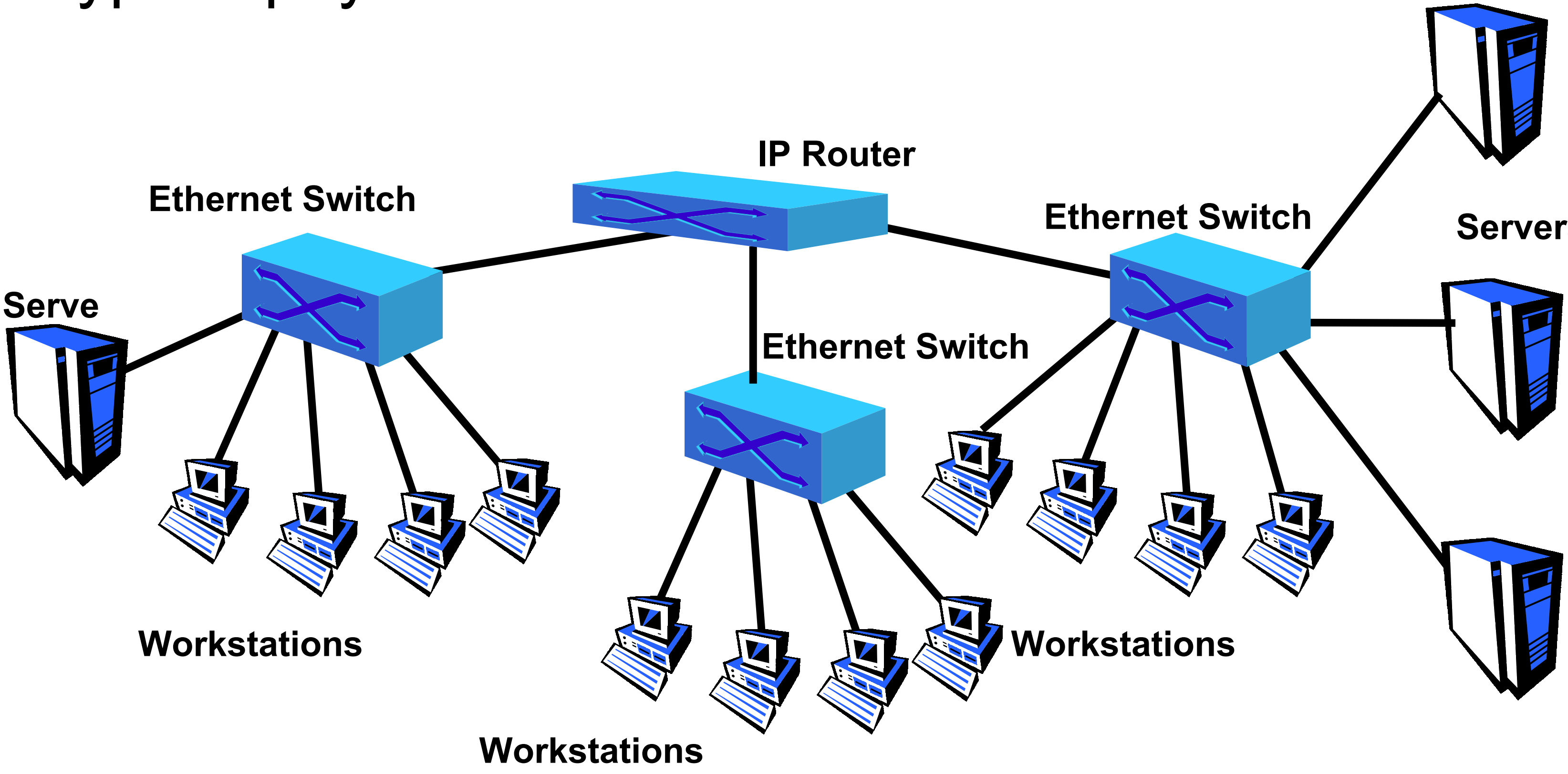
Potential Limiting Features:

- Frame distribution limits available bandwidth per connection to speed of a single link.
- Additional cabling to install and manage.
- Troubleshooting multiple LAN links on both system and switch.
- Some added management cost to implement.
- Not currently available on IPF.



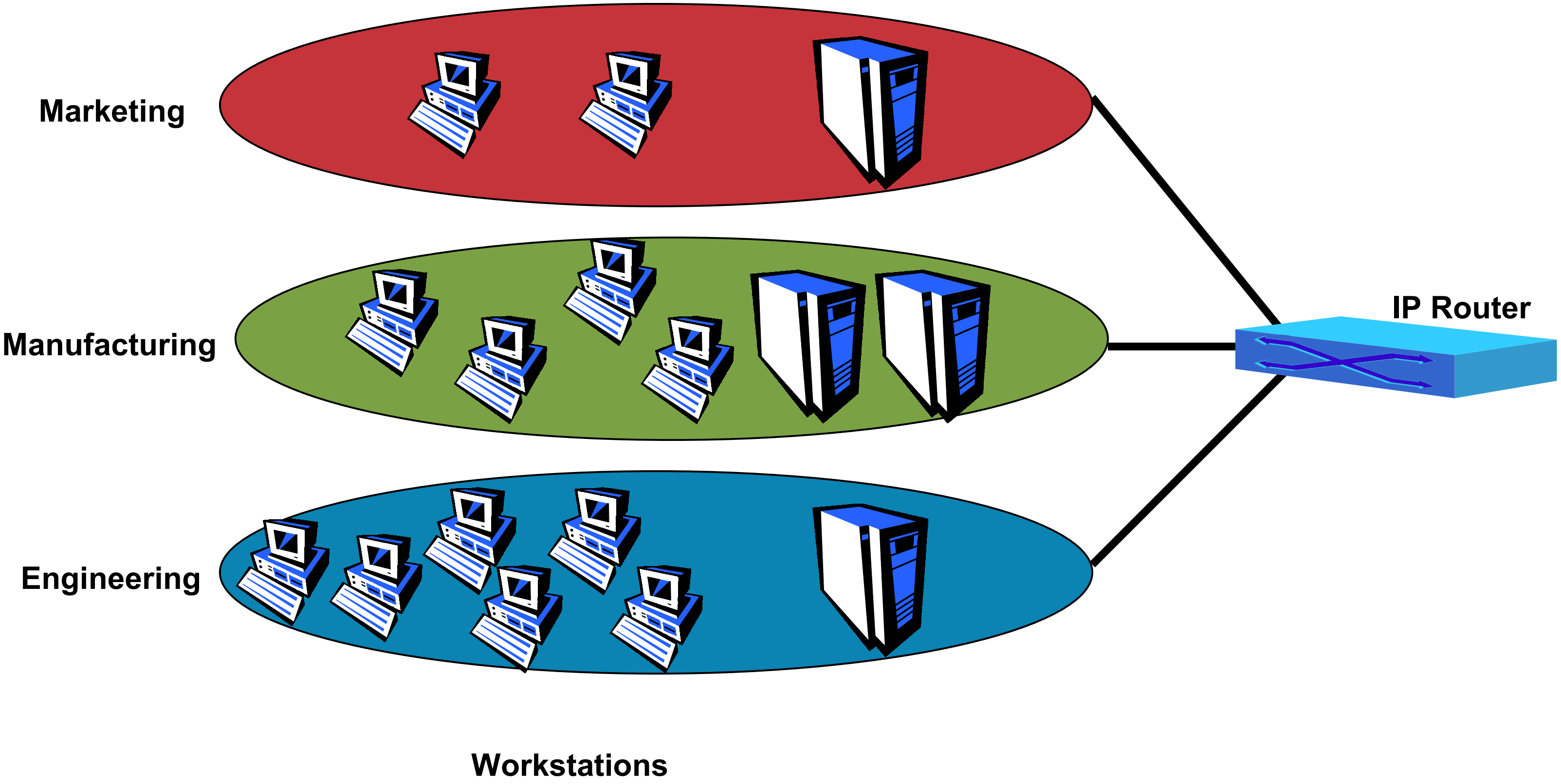
Introduction to Virtual LANs

- Typical physical network *without* VLANs



Logical Network by Department

- Logical Network by Department

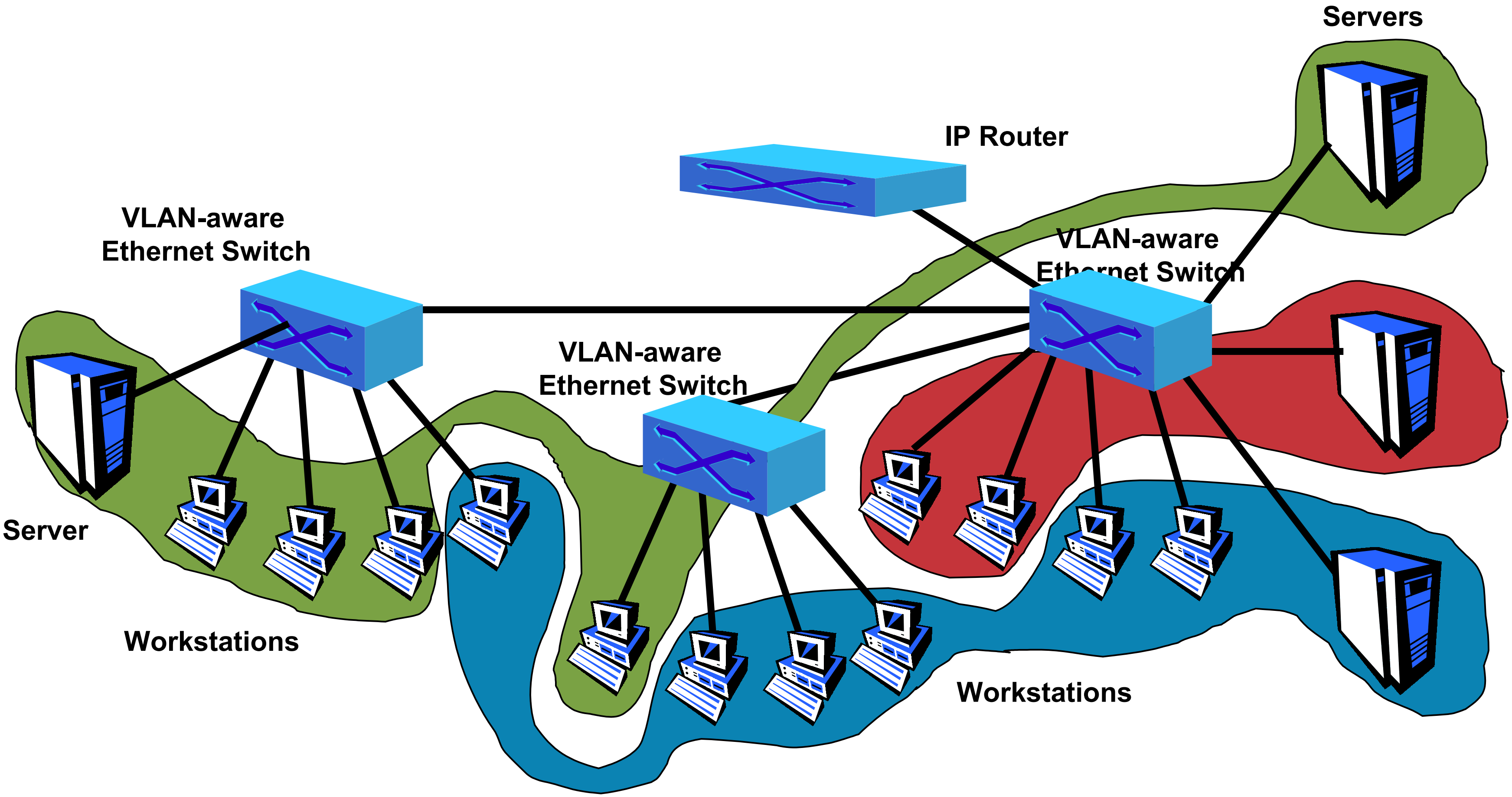


Logical Network Partitioning

- High level network design
- Implemented with VLAN aware switches
- Explicit and implicate VLAN association
- Each VLAN identified by VLAN ID
- Individual switch ports configured to belong to one or more VLAN

Typical VLAN implementation

- Typical physical network with VLANs

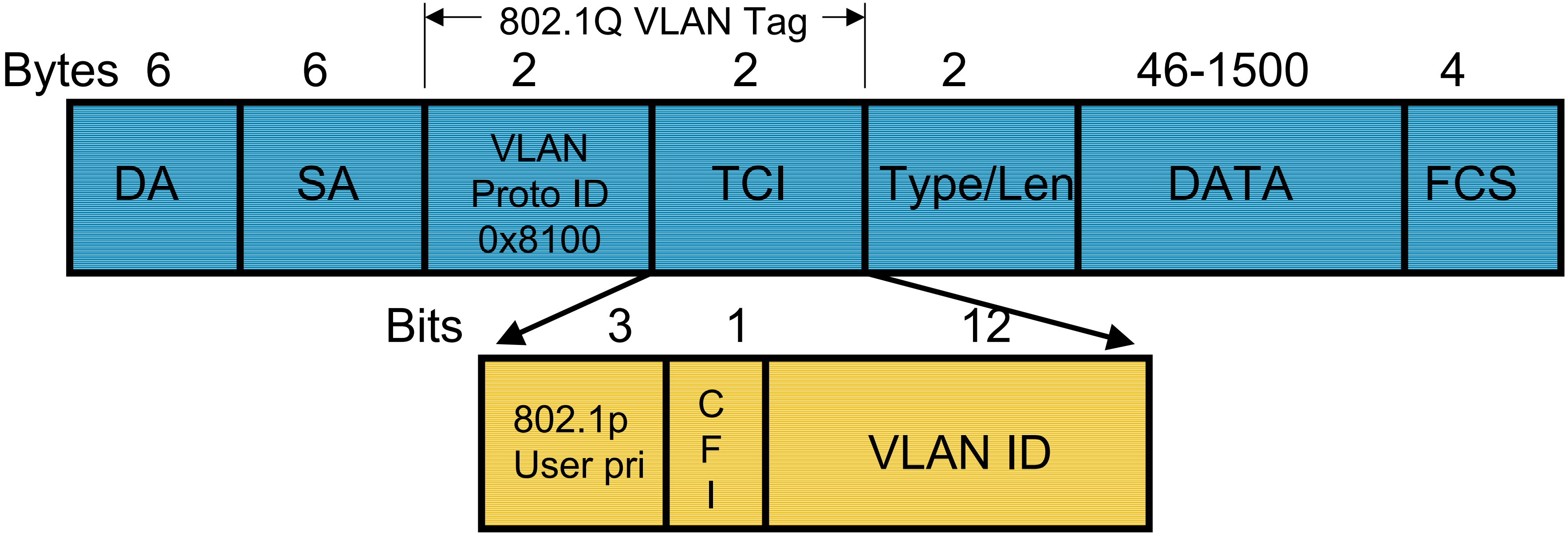


Benefits of VLANs

- Manageability (console or closet)
- Enhanced Security
- Bandwidth Preservation
- Better use of server resources
- Link consolidation

VLAN Tagging

- Diagram of frame with 802.1 Q/p tag



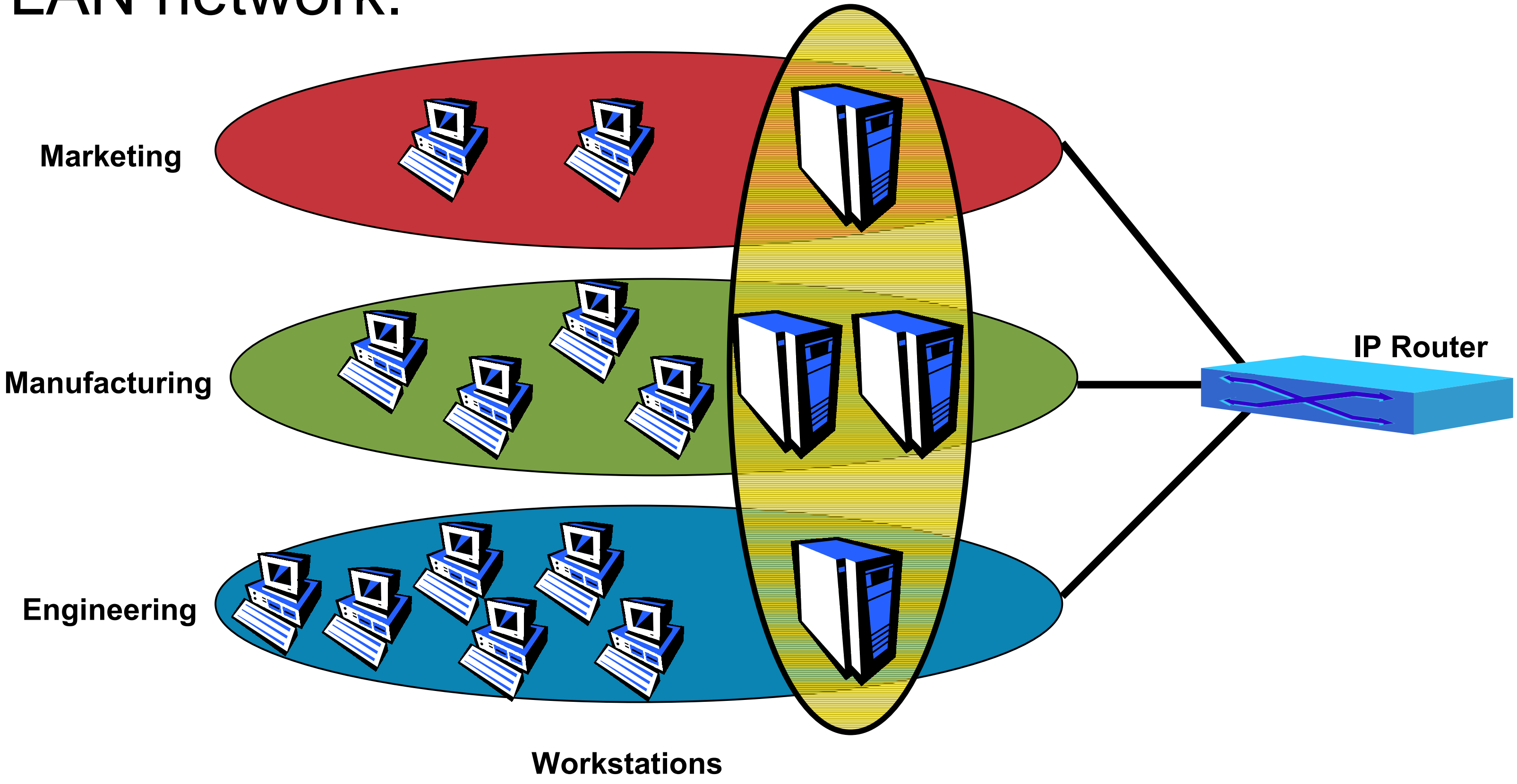
HP-UX VLAN Features

- Host-base 802.1Q/p tagging supported on 11i March 2002 w/patches.
- Supported on HP's HSC & PCI Fast Ethernet and Gigabit Ethernet NIC cards
- Up to 1024 VLANs per NIC port
- MC/Serviceguard Support
- Tagged Virtual LANs configured via SAM or directly via lanadmin

Adding New VLAN Subnet

- New VLAN configured for new server only

VLAN network.



VLAN Summary

- Allows flexible network configuration
- Potential improvements in overall network and system throughput.
- No cost enhancement to 11i (11.11)
- Future evolution of HP VLANs

HyperFabric

- HyperFabric is a very high speed, low latency system interconnect.
  - Proprietary cluster connections
  - Uses switched fabric technology
  - Designed to provide large cluster solutions

HyperFabric Benefits

- Improved throughput by increasing Bandwidth and reducing Latency
- Hyper Fabric provides complete End to End High Availability by implementing
  - Dynamic routing
  - Active-Active High Availability
  - Transparent fail over at link level
  - Makes it ideal platform to run mission critical applications such as ERP, DSS. Eg: SAP
- Increased Scalability

Points to consider…

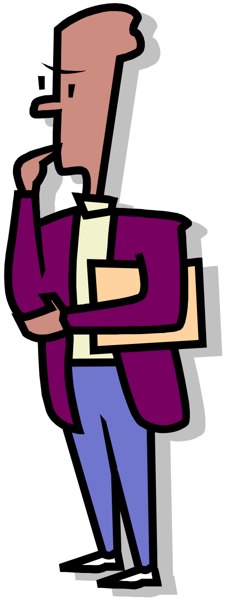
- Requirements should drive design.
- Check product features and compatibility.
- Read product release notes.
- Install with latest product version and check install patches and patch dependencies.
- Get familiar with new features before rolling into production.

Industry Futures

- 10Gigabit Ethernet
- TOE
- iSCSI
- RDMA
- PCI-X 2.0
- Infiniband

more information…

- www.docs.hp.com/hpux/netcom/index.html
  - Check HP-UX network performance white papers
- www.hp.com/products1/unixserverconnectivity/adapters/index.html
  - HP-UX network connectivity products
- www.hp.com/go/network_city
  - Switch technologies and case studies
- www.cisco.com/warp/public/473/4.html
  - Cisco's write-up on their various families of switches and routers distribution methods
- http://www.10gea.org/
  - Information about current 1Gigabit Ethernet as well a future technologies inc; 10GBE, TCO, iSCSI



HP WORLD 2003

Solutions and Technology Conference & Expo

Interex, Encompass and HP bring you a powerful new HP World.

interex
shared knowledge • shared power

encompass
AN HP USER GROUP

hp
invent