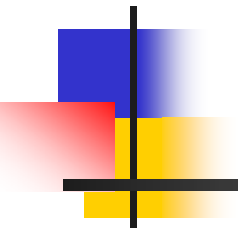


# Link Aggregation in Server End stations

Ilango Ganga, Intel

July 16, 2007

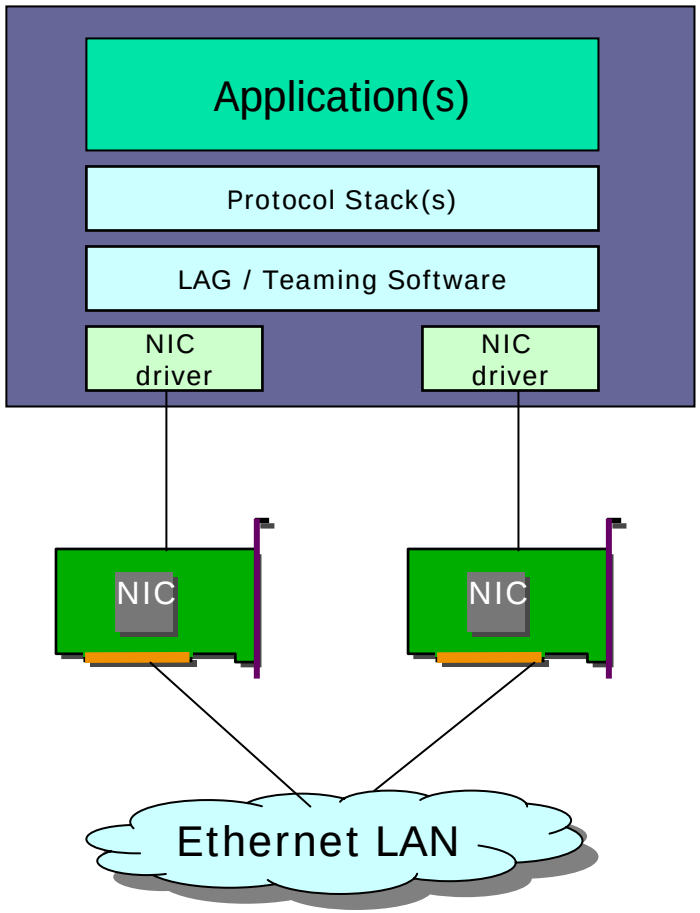
# LAG in servers overview

- Link aggregation allows multiple Ethernet NICs to appear as single network interface for the applications
- A variation of this technology referred as “Teaming” is used to bond multiple Ethernet links together
- Link Aggregation provides high availability in Servers, avoids single point of failure
- A distribution algorithm directs traffic over to one of the NICs in the host
- Increases overall throughput for aggregated flows
- Does not increase the bandwidth for a single flow or source/destination pair
- Does not improve latency

# LAG limitations explained

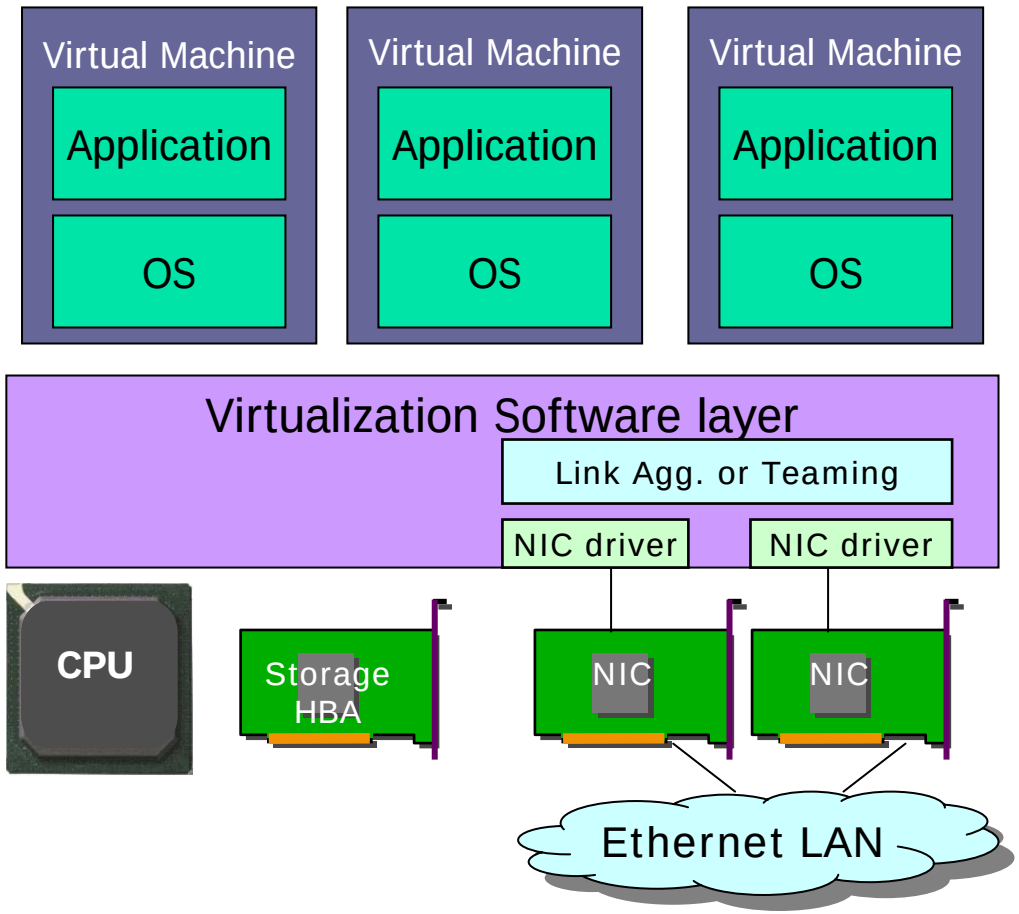
- Uneven distribution of traffic over a team
  - Inefficiency based on traffic types
  - E.g.  IPSec, video, multicast, storage, etc.,
  - Does not scale throughput per flow or source/target pair (e.g. storage traffic)
    - Helps to scale throughput in targets where multiple flows are aggregated
  - Security encapsulations adds complexity to implementing load balancing algorithms
  - Adds to CPU utilization, higher power/cost
- Inefficiencies explained in bennet_01_0906 and bechtel_01_0307
- 802.3ad LAG facts explained in frazier_01_0407
- Good and the bad from a server perspective explained in muller_01_0507
- Cost of LAG explained in vandoorn_01_0507

# LAN controller LAG or Teaming



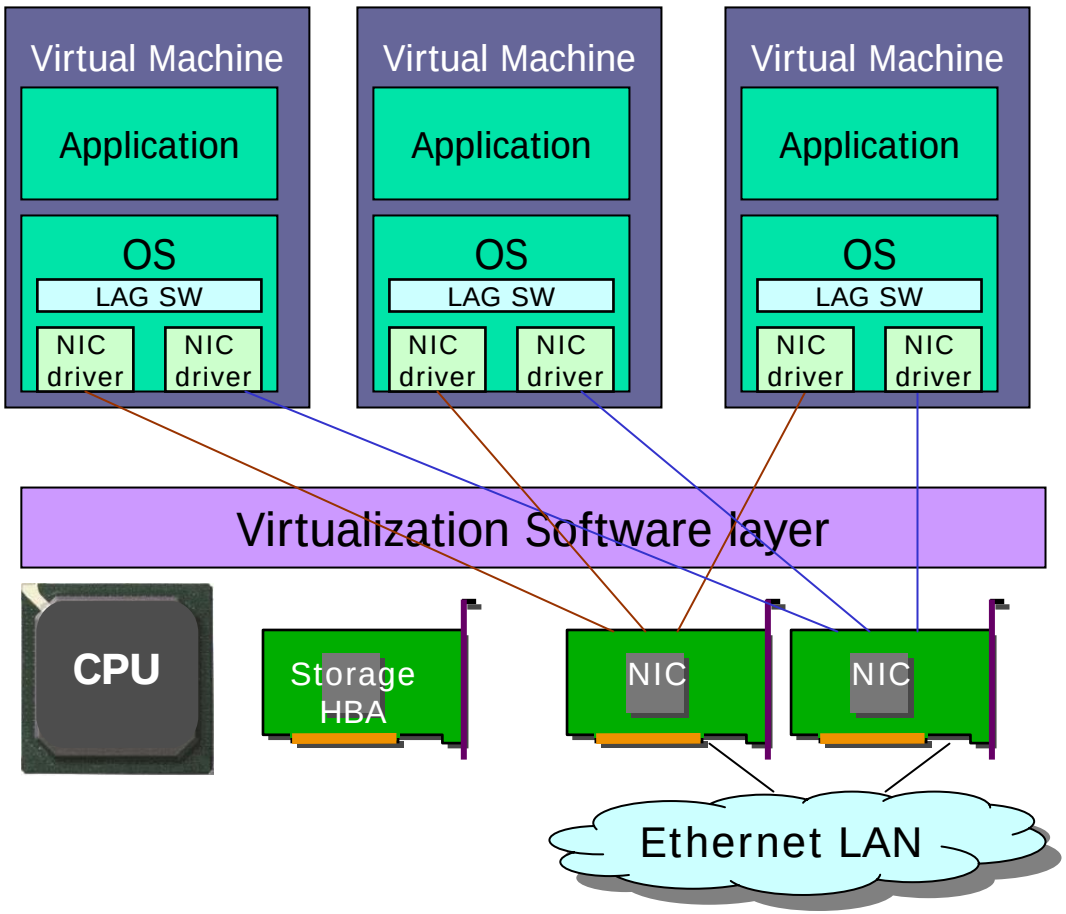

- LAG performed in teaming driver/software
- Load distribution based on hashing
  - Packets classified L2/3/4 headers
  - E.g. MAC, IP, TCP ports
- Throughput increases as more ports added to a team
- Increases CPU utilization
- Reduces CPU performance for application processing

# LAG with I/O Virtualization



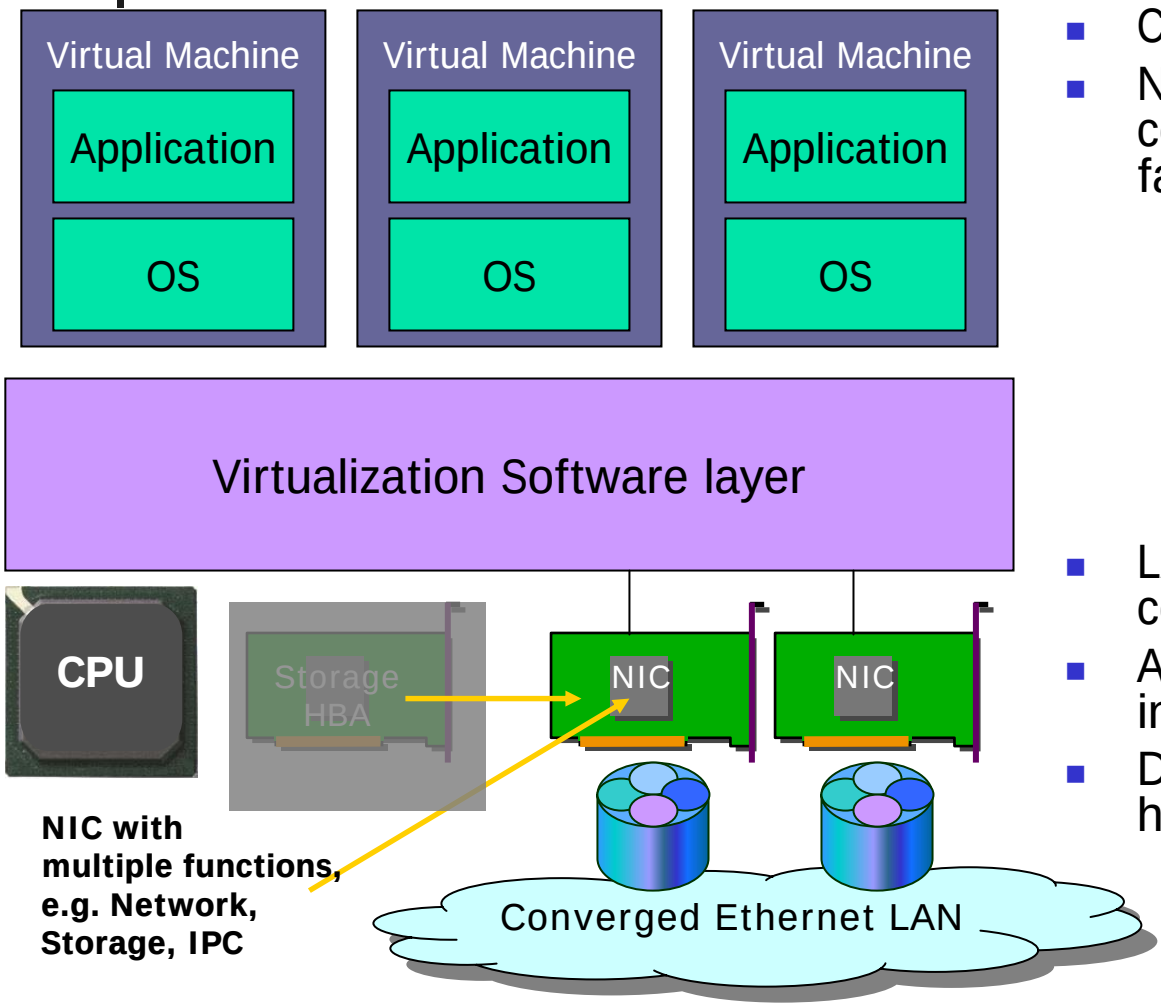

- LAG performed in virtualization software layer
- Distribution algorithm for load distribution managed in Virtualization layer
- May have more flows due to multiple VMs
- Adds CPU overhead for packet processing
- Lowers CPU performance for virtual machines

# I/O Virtualization – an alternate architecture



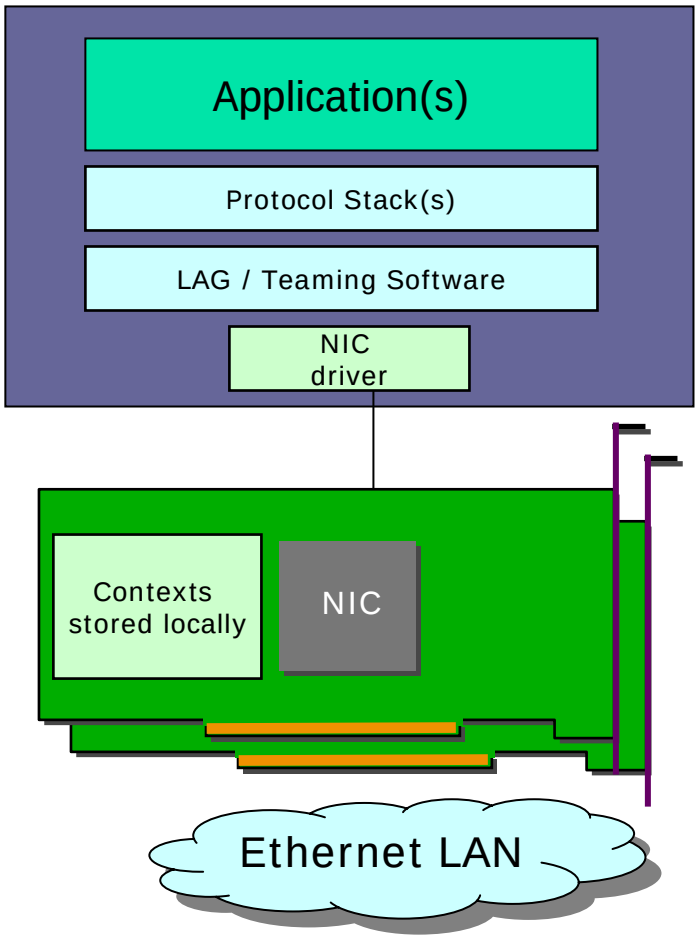

- Each VM has a direct path to virtualized NIC
- LAG performed in VMs by teaming software
- VMs do not have global view of flows in adjacent VMs
- Inefficiency in managing load distribution
- Adds CPU overhead
- Lowers CPU performance for virtual machines

# End station Virtualization – with converged network fabric



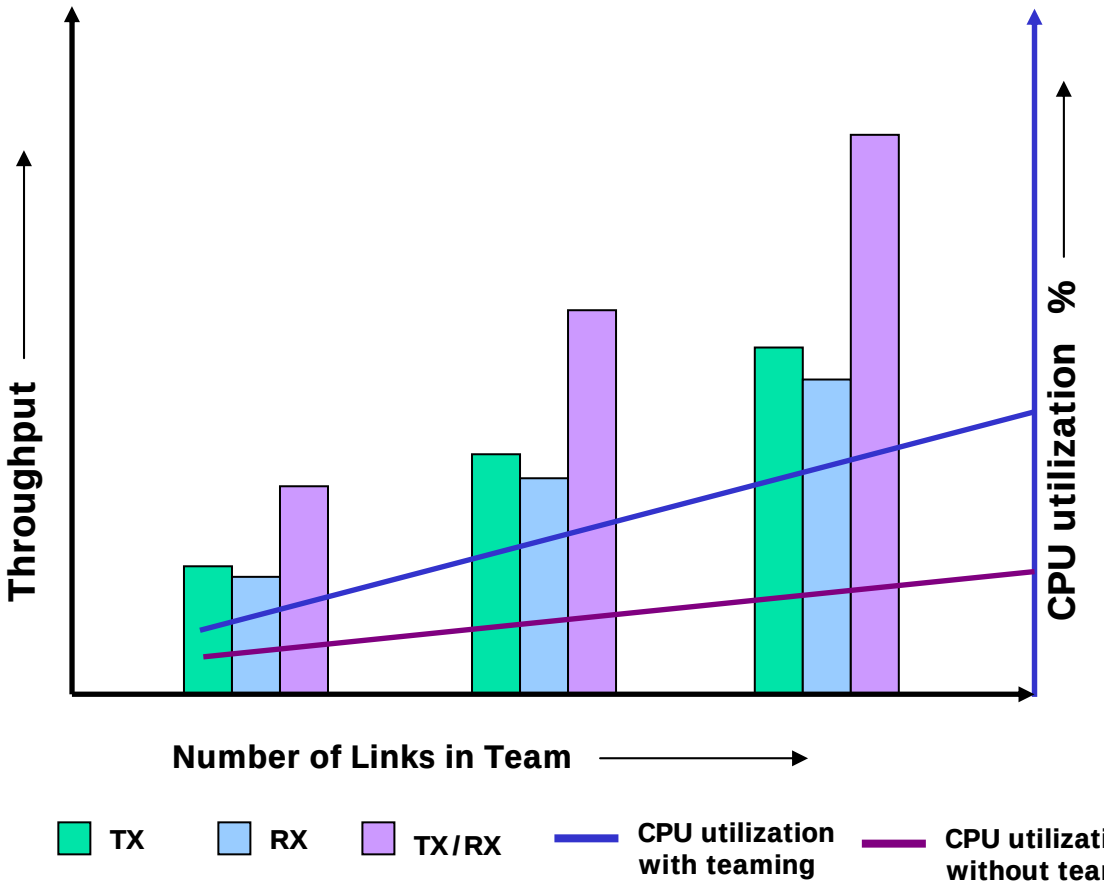

- Converged fabric in data centers
- Network, Storage and IPC converge into a single data center fabric
  - Each virtual fabric has different characteristics
  - E.g. Storage has no drop behavior, no out of order delivery
  - Low latency for IPC
  - Bandwidth guarantees per virtual fabric
- LAG adds additional level of complexity to traffic distribution
- Adds CPU overhead due to intelligent processing
- Does not reduce latency for high performance computing

# Context rich LAN controllers



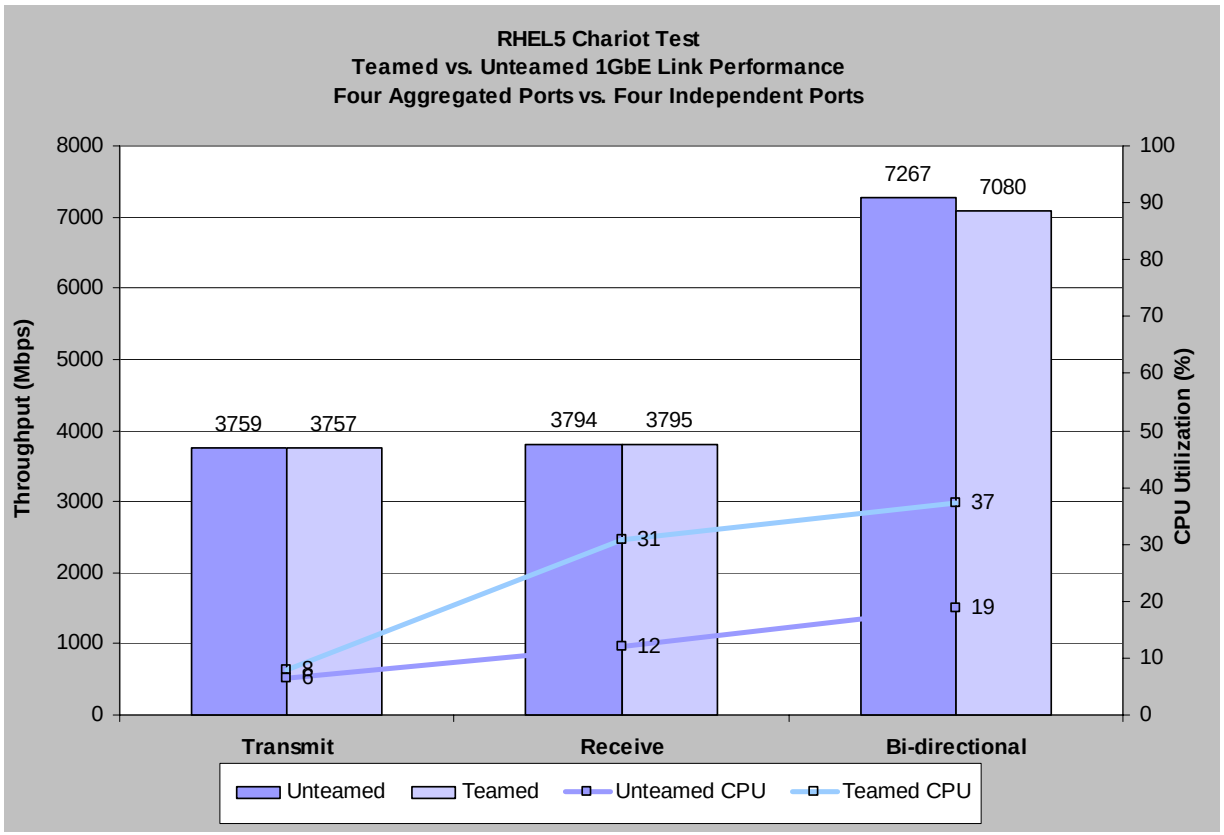

- Modern NICs store flow contexts locally in the NIC
- Provides protocol acceleration capabilities for many protocols Example:
  - IPsec, TCP/IP, Storage functions, remote DMA, etc.
- Expects bidirectional flow to terminate in the same NIC where the context is stored for efficient operation
- LAG does not guarantee inbound traffic to arrive in the same NIC that sent the outbound traffic

# LAG Performance in Servers




- Performance in Servers
  - Throughput
  - Latency
  - CPU utilization
- CPU utilization is important
  - More CPU available for real applications
- Performance may vary depending on various factors
  - E.g. OS, configuration, traffic type, features, etc.,
  - Throughput does not always scale linearly
  - No Latency improvement
- LAG impacts CPU performance

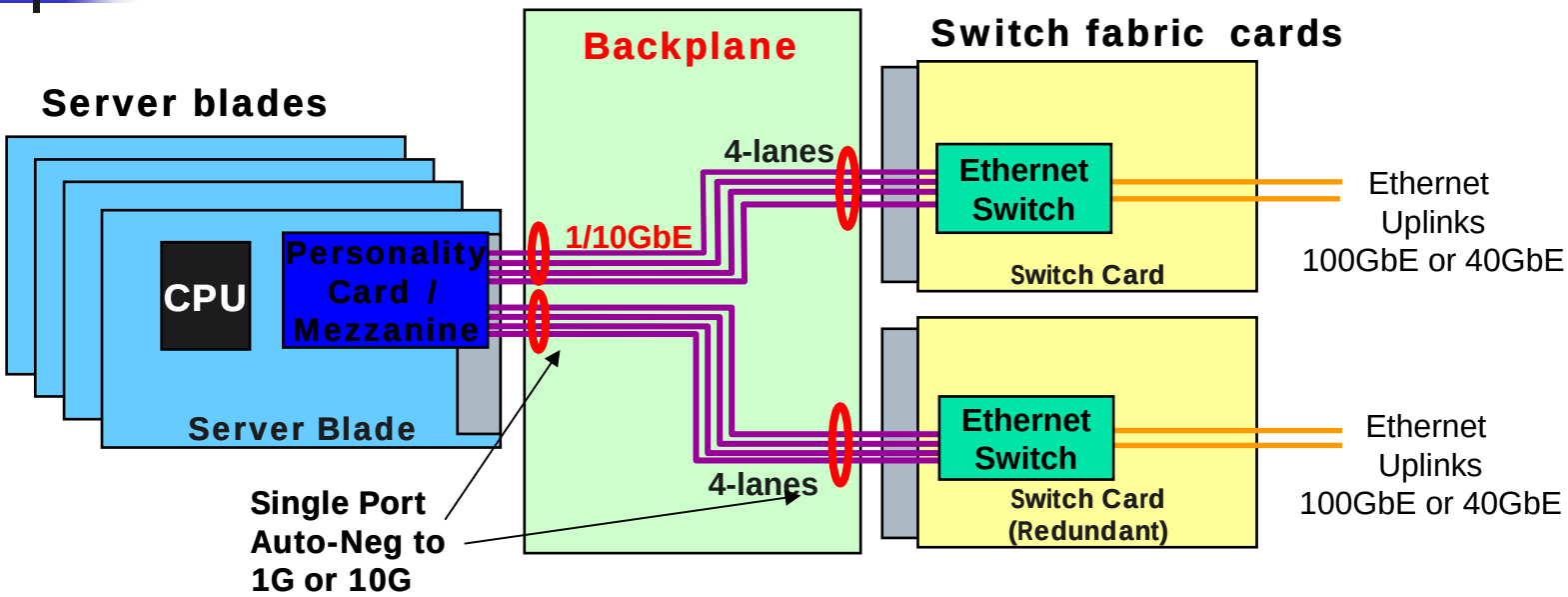
# Teaming performance benchmark




- OS: Red Hat Enterprise Linux 5
- Ixia Chariot test suite
- Server System: Intel Xeon
- 1000 Mbps NIC(s)

- Throughput or CPU utilization may vary based on OS, drivers, traffic types, features, vendor, benchmark configuration, etc.,

# Modular computing
## Blade servers and aTCA




- Mezzanine card to each Switch links constitute a single port
  - A single port type is negotiated when blades are plugged in
  - 1000BASE-KX, or 10GBASE-KX4, or 10GBASE-KR
  - “40GBASE-KR” is a natural extension to add backplane capacity
- LAG cannot be used in a standard way, as x4 links belong to same port
- Teaming will be used in blades to provide fault tolerance and failover

# Summary

- LAG is good, but is not a replacement to higher speed links
- LAG will still be used to provide high availability and fault tolerance in Servers
- LAG adds to CPU utilization, does not improve latency, adds complexity to configuration/management
- Advanced features like security difficult to implement with LAG
- Higher link speed 40GbE needed to match and keep up with Server performance