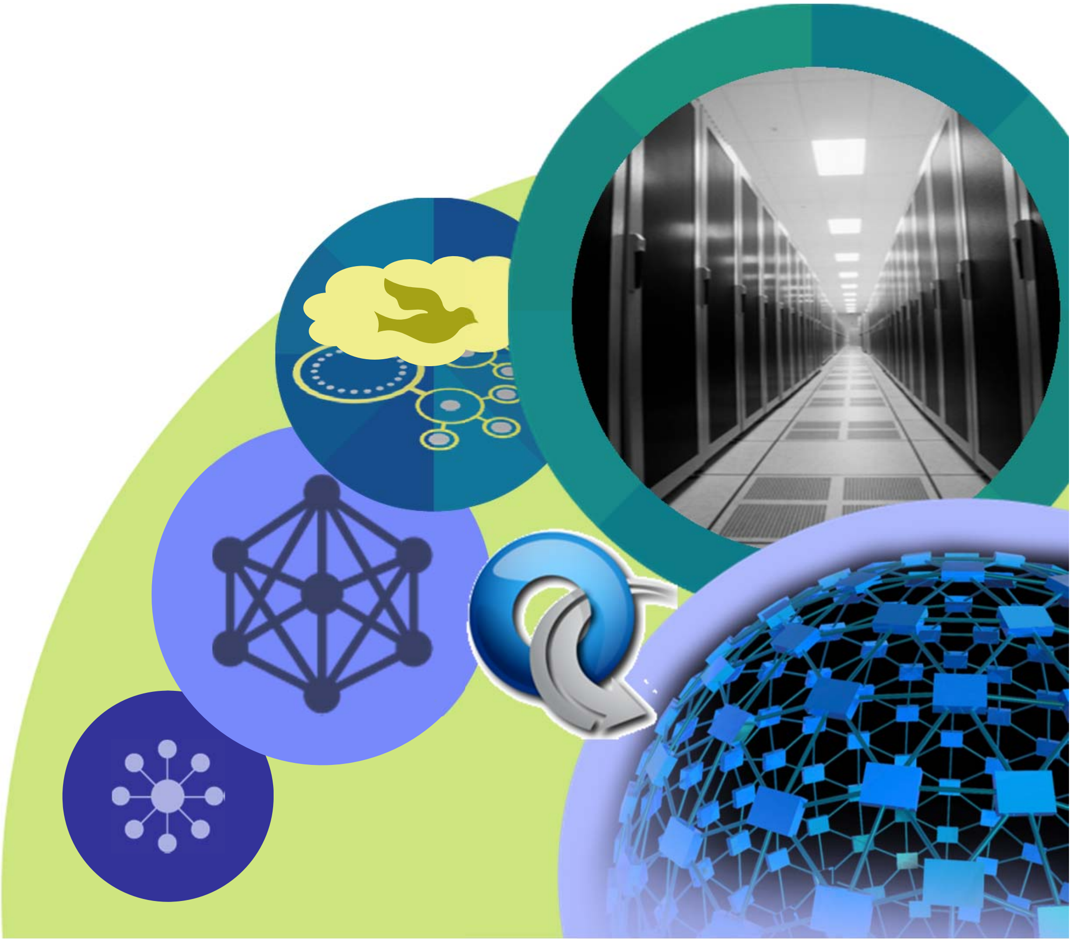
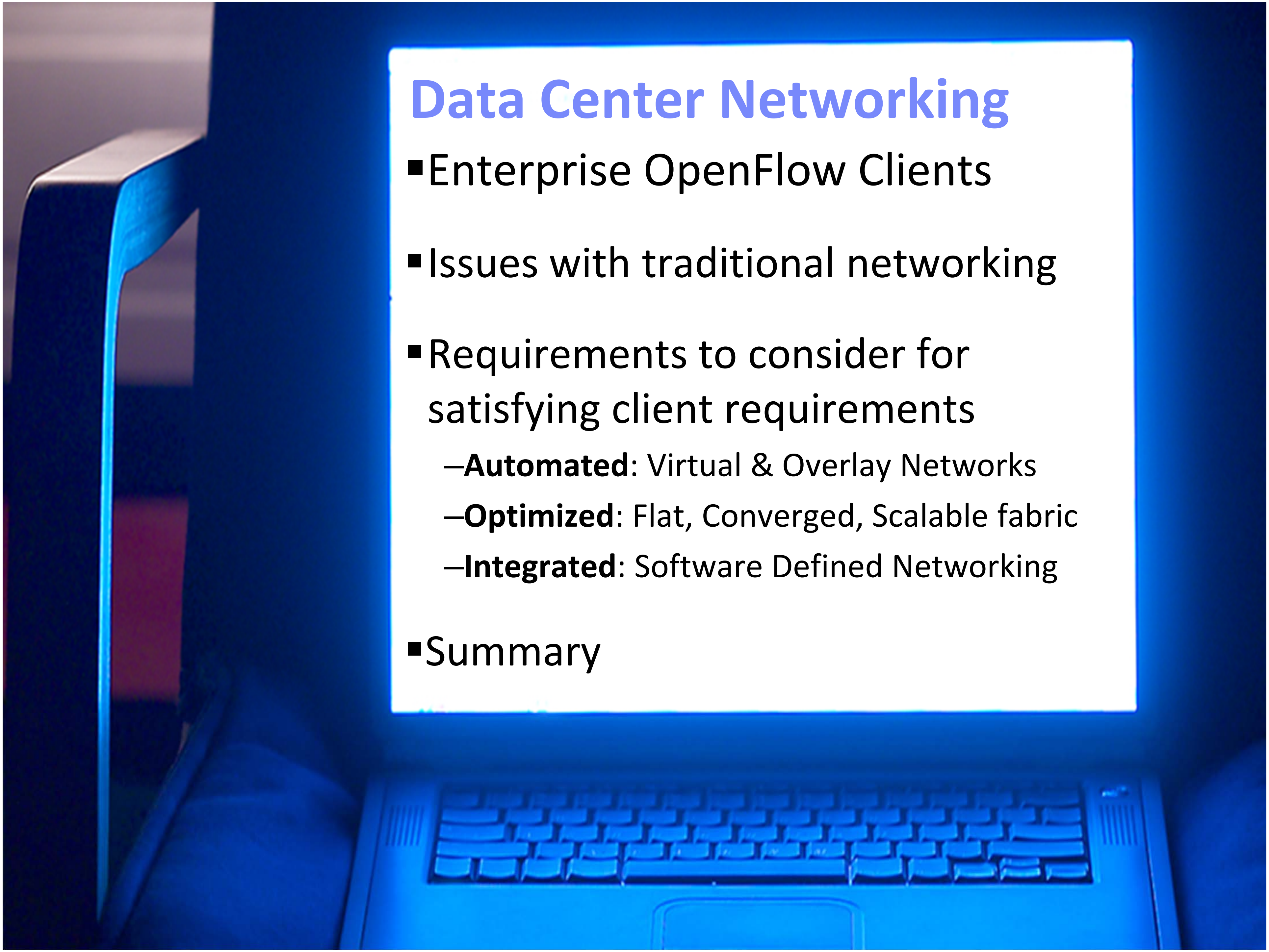
OpenFlow in Enterprise Data Centers

Products, Lessons and Requirements



Renato Recio

IBM Fellow &
System Networking CTO

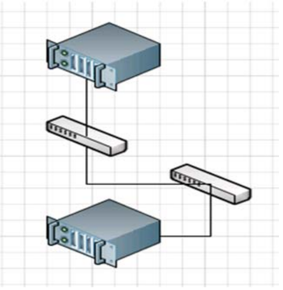


Data Center Networking

- Enterprise OpenFlow Clients
- Issues with traditional networking
- Requirements to consider for satisfying client requirements
  - **Automated**: Virtual & Overlay Networks
  - **Optimized**: Flat, Converged, Scalable fabric
  - **Integrated**: Software Defined Networking
- Summary

OF Customer: Tervela

Provider of a market-leading distributed data fabric for global trading, risk analysis and e-commerce



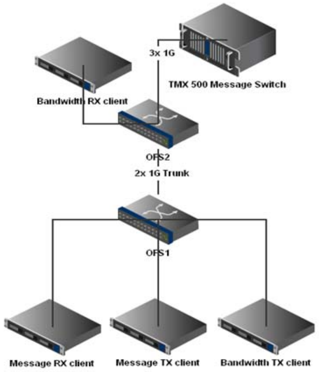
Test 1: OpenFlow deliver fast packet forwarding

Key Benefits

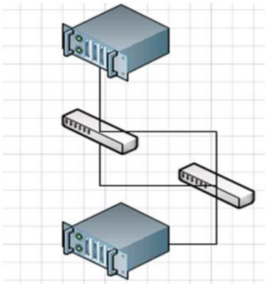
Deterministic Latency

Predictable Network Performance

Rapid Convergence



Test 2: OpenFlow switches segregate traffic



Test 3: OpenFlow switches Manage multiple trunks

4x 40 GE + 48x 10GE

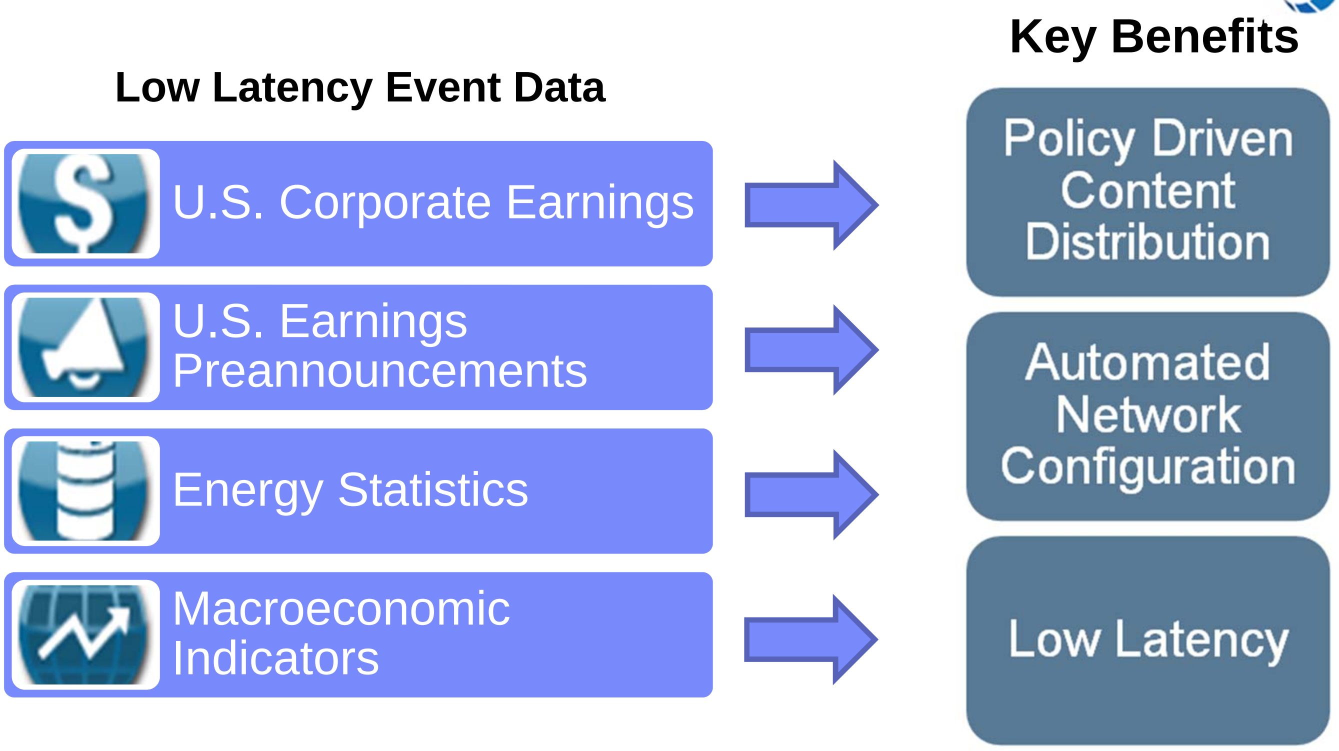
Tervela's testing validated the **IBM and NEC OpenFlow** solution ensures predictable performance of Big Data for complex and demanding business environments.

© 2012 IBM Corporation

OF Customer: Selerity

Ultra-low latency, real-time financial information provider

SELERITY


Key Benefits

Low Latency Event Data

U.S. Corporate Earnings

U.S. Earnings Preannouncements

Energy Statistics

Macroeconomic Indicators

Policy Driven Content Distribution

Automated Network Configuration

Low Latency

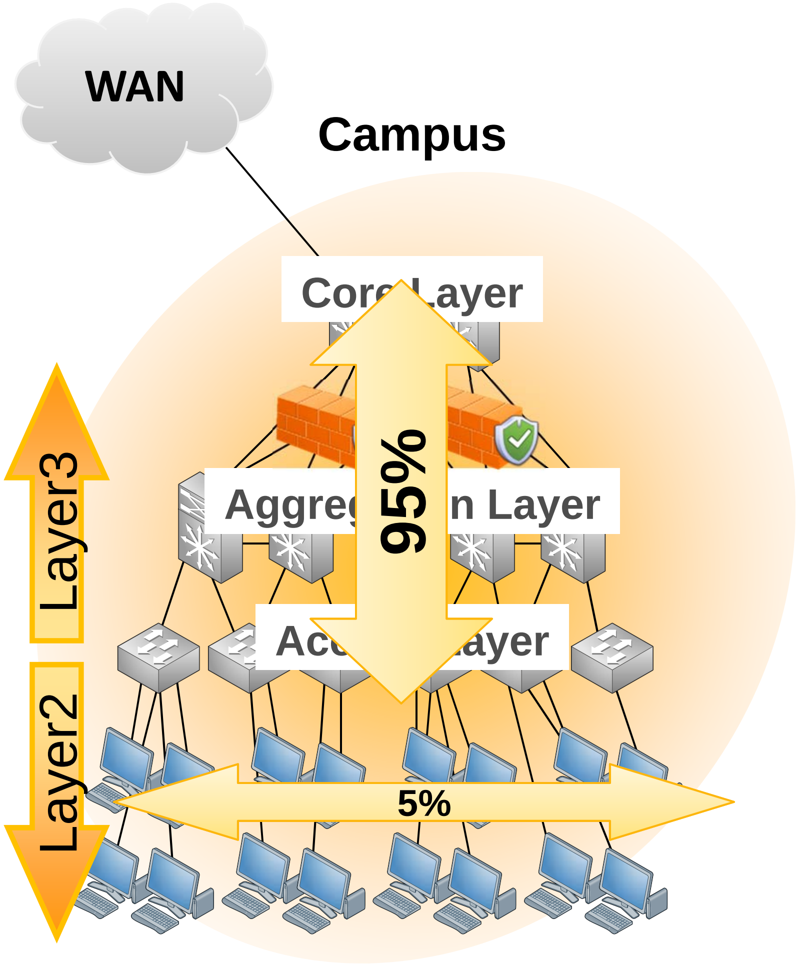
4x 40 GE + 48x 10GE

Selerity's **IBM and NEC's OpenFlow** solution improves real-time decision-making for global financial markets.

© 2012 IBM Corporation

The Beauty of Trees

- In the beginning, Ethernet was used to interconnect stations (e.g. dumb terminals), initially through repeater & hub topologies…



And… eventually through switched topologies.

- Ethernet campus evolved into a tree structure
  - Typically: core, services, aggregation & access planes.
  - Traffic is mostly North-South (directed outside campus).
  - To avoid spanning tree problems, campus networks typically are divided at access.

The industry liked the tree structure & applied it to DC

© 2012 IBM Corporation

The Beauty of Trees

- In the beginning, Ethernet was used to interconnect stations (e.g. dumb terminals), initially through repeater & hub topologies…

And… eventually through switched topologies.

- Ethernet campus evolved into a tree structure
  - Typically: core, services, aggregation & access planes.
  - Traffic is mostly North-South (directed outside campus).
  - To avoid spanning tree problems, campus networks typically are divided at access.

The industry liked the tree structure & applied it to DC

© 2012 IBM Corporation

The Repotted Campus Tree

Soo… Campus Ethernet tree was repotted to the Enterprise Data Center. ***which…***

- Has different traffic patterns:
  - 50-75% East-West* in DC
  - 95% North-South in Campus
- Has different fabric performance needs
  - Lossless traffic for storage
  - Low latency & high bandwidth for clusters
- Evolved into a virtual compute model, with different demands:
  - From static workloads
    → to dynamic workloads
    → to multi-tenant, dynamic workloads

…which today results in

complex and/or inefficient service plane
(e.g. to protect East-West traffic)

*IMC 2010 ACM paper "Network Traffic Characteristics of Data Centers in the Wild"

© 2012 IBM Corporation

Problems with the Repotting

Discrete & Decoupled

- **Discrete** components and **piece parts**
- **Multiple managers** and management **domains**
- **Box level** point Services (e.g. IPS, FW)

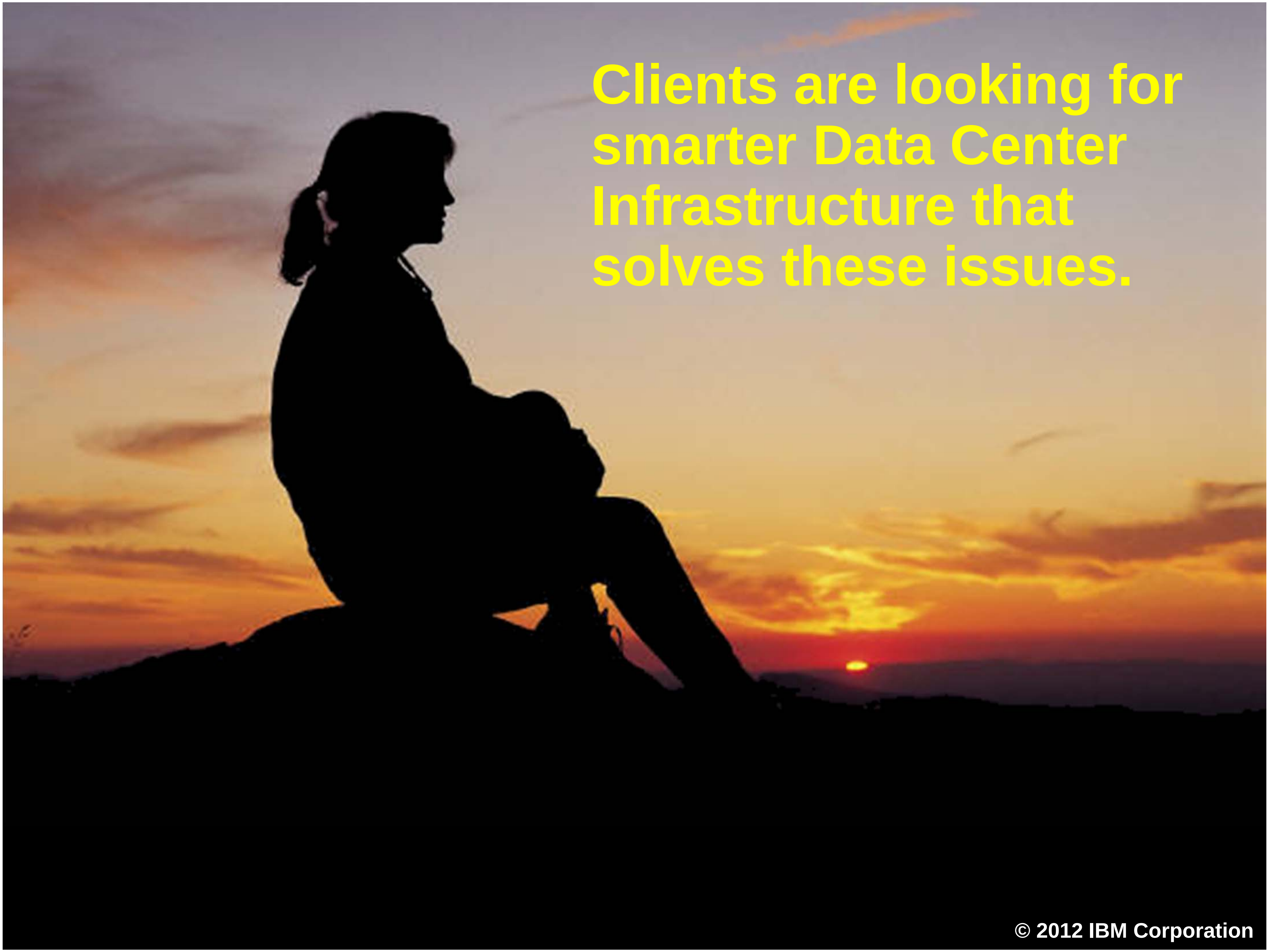
Manual & Painful

- **Dynamic workload** management **complexity**
- **Multi-tenancy complications**
- SLAs & security are **error-prone**

Limited Scale

- **Too many network types, with** too many **nodes & tiers**
- **Inefficient** switching
- **Expensive** network resources

© 2012 IBM Corporation

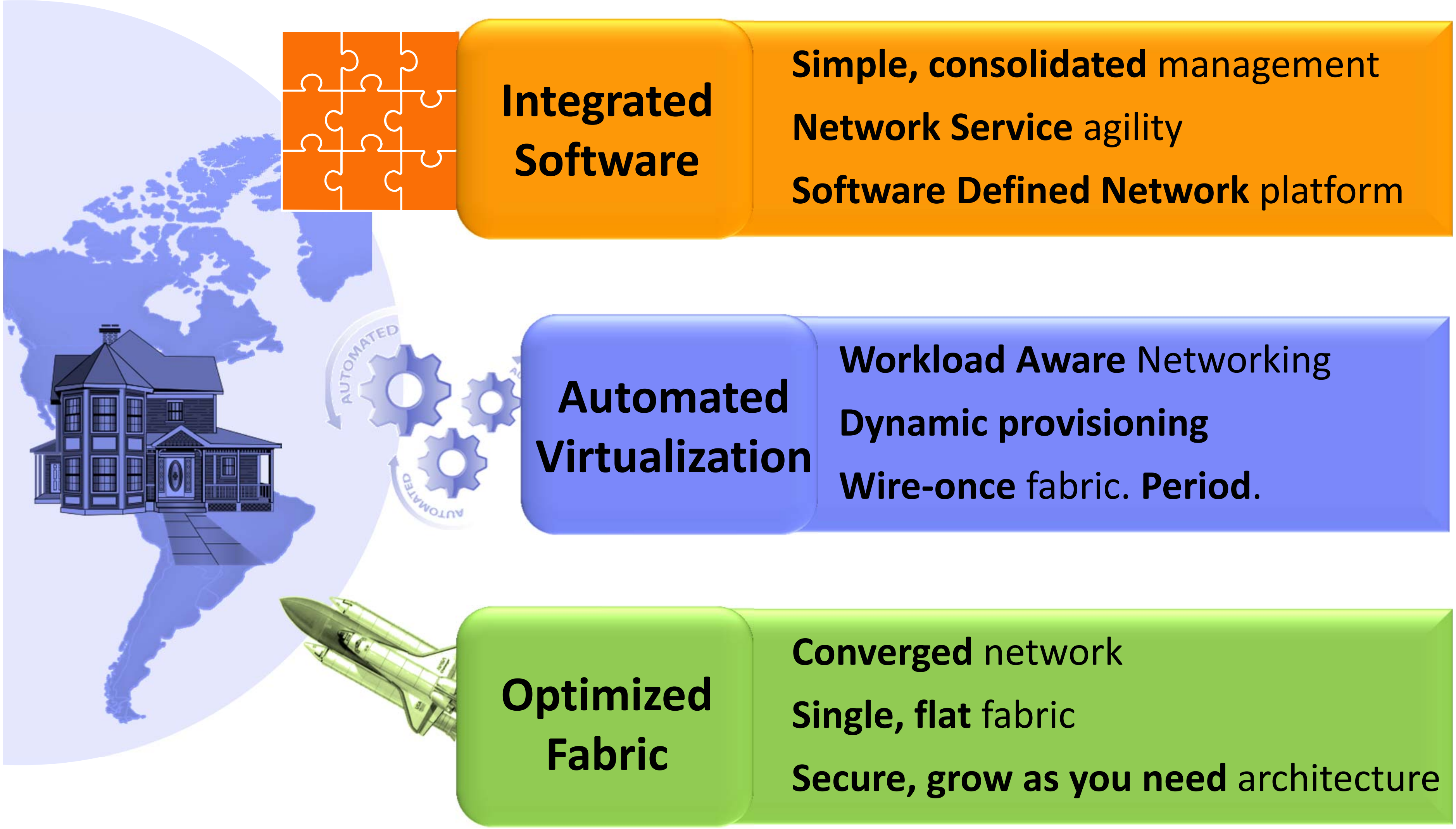


Clients are looking for smarter Data Center Infrastructure that solves these issues.

© 2012 IBM Corporation

Data Center Network Requirements

…and associated client value



Integrated Software

- **Simple, consolidated** management
- **Network Service** agility
- **Software Defined Network** platform

Automated Virtualization

- **Workload Aware** Networking
- **Dynamic provisioning**
- **Wire-once** fabric. **Period**.

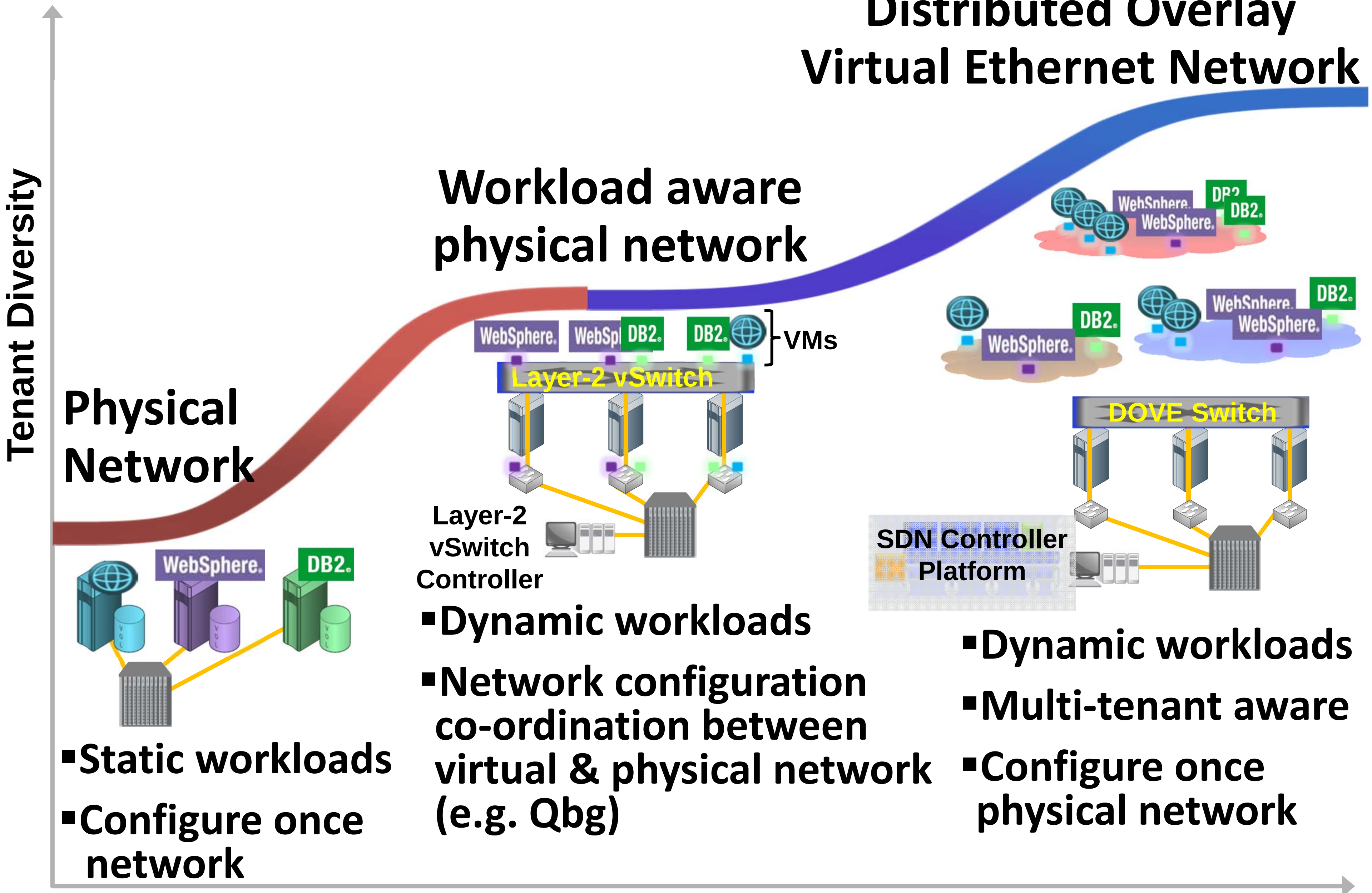
Optimized Fabric

- **Converged** network
- **Single, flat** fabric
- **Secure, grow as you need** architecture

© 2012 IBM Corporation

Automated Virtualization

Virtualization Requirement Trends



Physical Network

- Static workloads
- Configure once network

Workload aware physical network

- Dynamic workloads
- Network configuration co-ordination between virtual & physical network (e.g. Qbg)

Distributed Overlay Virtual Ethernet Network

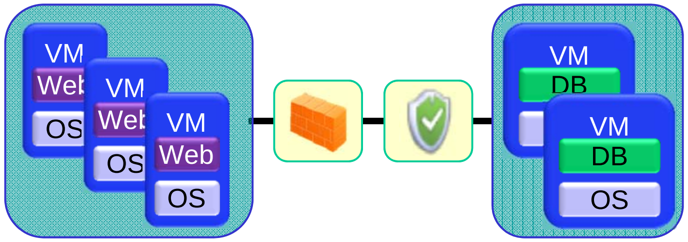
- Dynamic workloads
- Multi-tenant aware
- Configure once physical network

© 2012 IBM Corporation

Example constructs for providing DOVE Network Requirements

Virtual Network

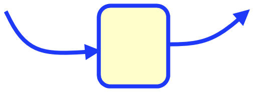
Interconnected workload groups



- Connects a set of workload groups and associated middleboxes

Services Middlebox

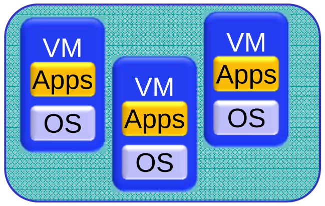
Interconnects Workloads



- Network service provider (e.g. Firewall, IPS, ADC)
- Virtual or physical

Workload Group

vNIC port set



- Logical grouping of workloads
- Workloads share network services

Workload

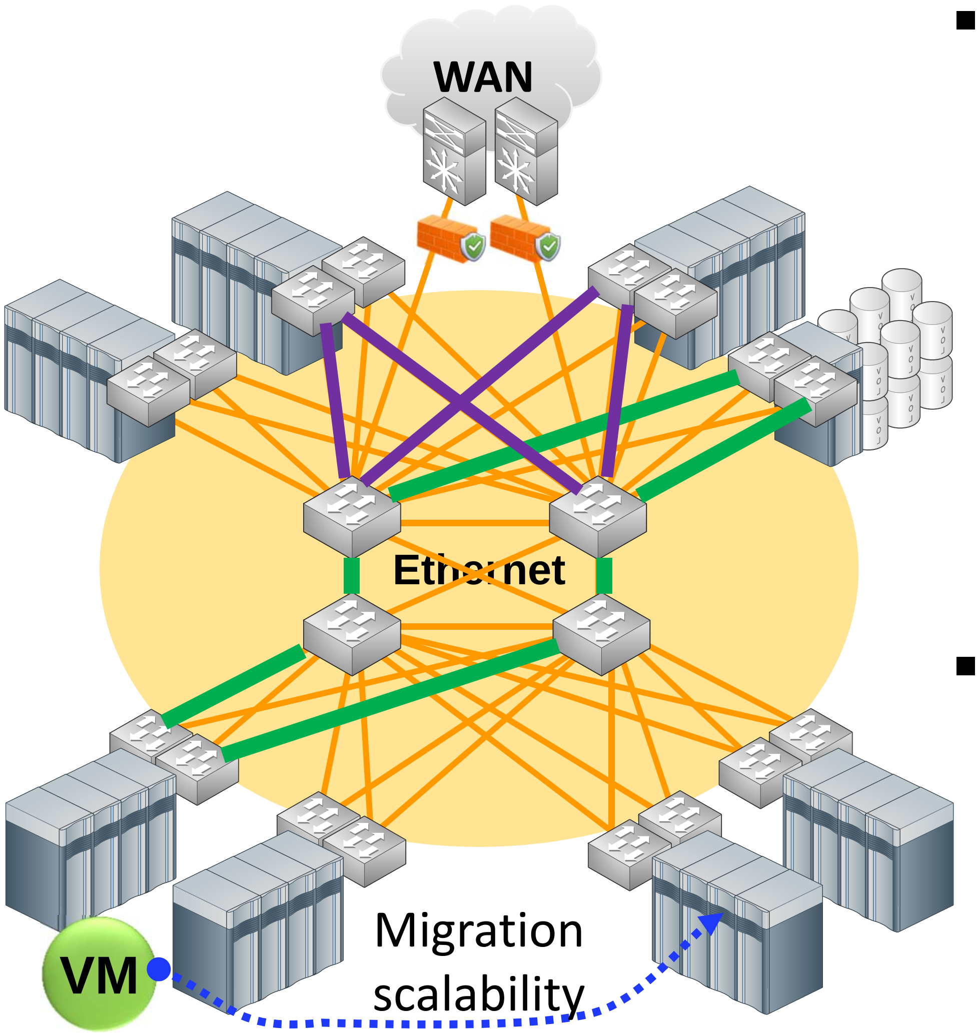
Virtual NIC port

- Layer-2 address `(00:23:45:67:00:23)`
- Layer-3 address `(129.2.200.5)`
- Port QoS attributes `(e.g. # Gbps)`

© 2012 IBM Corporation

Optimized Fabric

Data Center Fabric Requirements



VMware, KVM, XEN,
PowerVM, Hyper-V, zVM

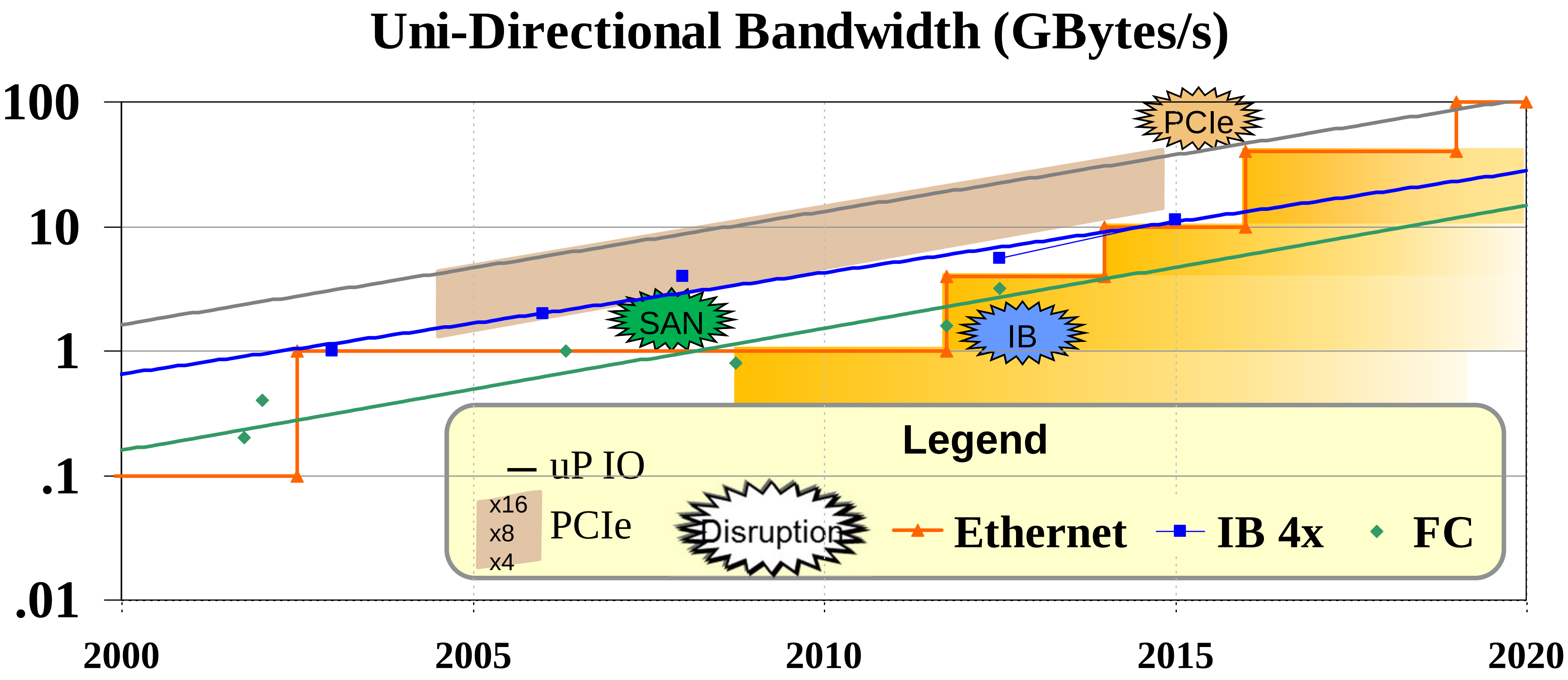
🟪 Shared multi-path 🟩 Disjoint multi-path

- Scalable fabric
  - Multi-pathing (shared & disjoint)
  - Large cross-section bandwidth
  - HA, with fast convergence
  - Switch clustering (less switches to manage)
  - Secure fabric services, for physical and virtual workloads
- Converged network
  - Storage: FCoE, iSCSI, NAS & FC-attach
  - Cluster: RDMA over Ethenet
  - Link: flow control, bandwidth allocation, congestion management
- High bandwidth links
  - 10GE → 40 GE → 100 GE

© 2012 IBM Corporation

Optimized Fabric

System IO and Fabric Trends



- Ethernet performance growth is causing disruptions in DC fabrics:
  - 10 GE & CEE → Disrupting storage market (Fibre Channel SAN)
  - 40 GE & CEE → Will further disrupt cluster market (InfiniBand)
  - 400 GE & CEE → Will disrupt server IO market & structure in 4-6 years.

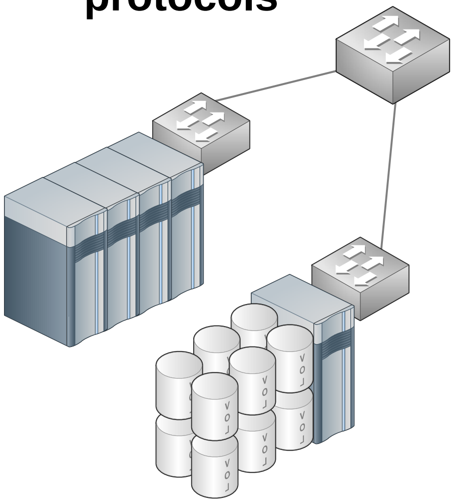
© 2012 IBM Corporation

Optimized Fabric

Link Configuration Requirement

To provide network convergence & fabric virtualization capabilities, data center links need to be configured, which today is performed through LLDP, DCBX and ECP.

Per link, discovery & configuration protocols

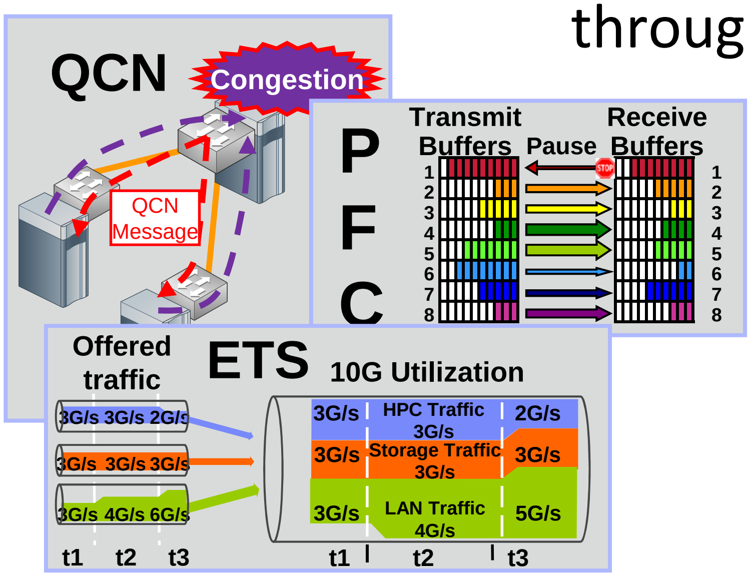


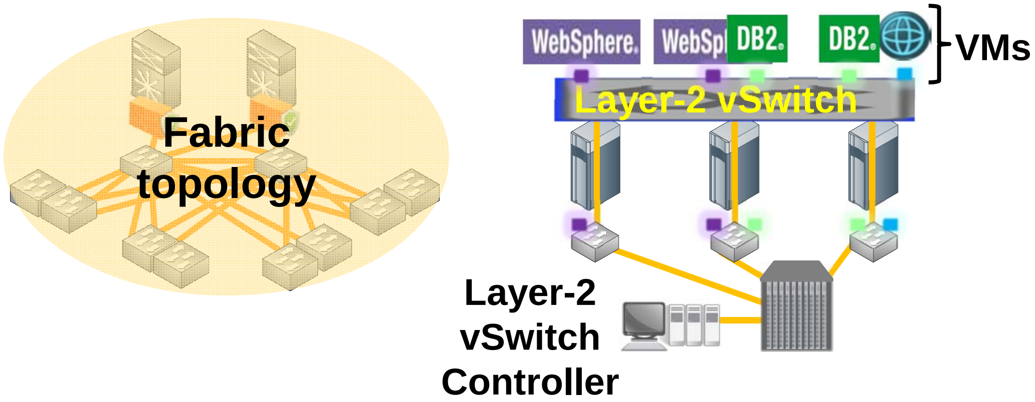
© 2012 IBM Corporation

Optimized Fabric

Link Configuration Requirement

To provide network convergence & fabric virtualization capabilities, data center links need to be configured, which today is performed through LLDP, DCBX and ECP.





- DCBX is used to discover and configure:
  - Priority flow control
  - Enhanced Transmission Selection
  - Per priority Quantized Congestion Notification feedback settings
- LLDP is used to discover and configure:
  - For management tools (MIBs): devices, neighbors, topology
  - For IEEE 802.1Qbg: reflective-relay, number of S-channels, …
- Qbg is used to discover & configure
  - port profiles (a.k.a. VSI Types) associated with a VM, …

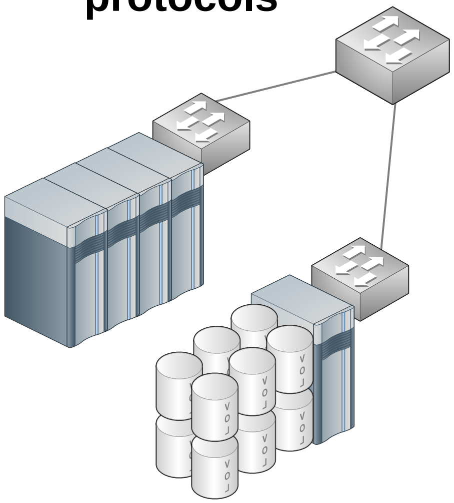
© 2012 IBM Corporation

Optimized Fabric

Link Configuration Requirement… Continued

Some of the options for performing link configuration for a pure OpenFlow fabric

Per link, discovery & configuration protocols



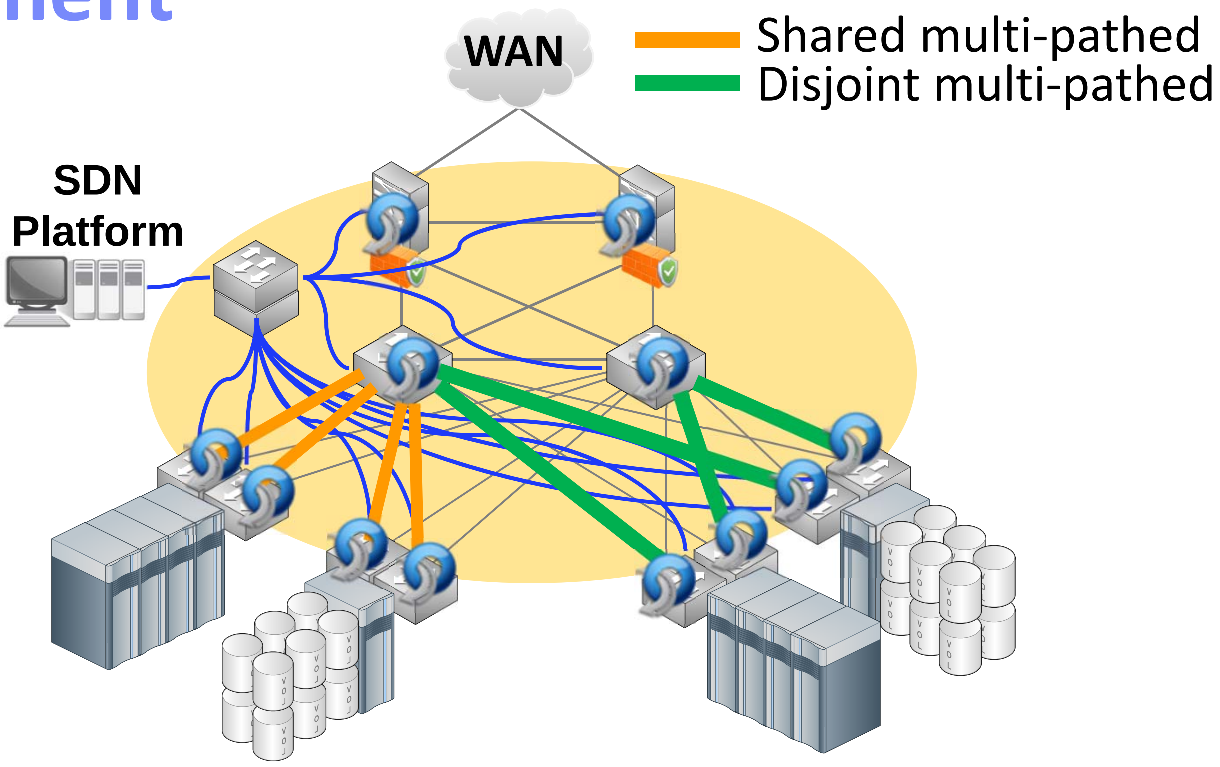
- Proxy all LLDP, DCBX and ECP messages to SDN Controller Platform (SNDCP)
  - Scaling issues
- Proxy less frequent frames to SDNCP (e.g. Qbg ECP/VDP) & have switch perform frequent frames
  - Rototills switch’s LLDP/DCBX/ECP processing
  - Scaling may still be an issue
- Let switch run link layer algorithms (LLDP, DCBX and Qbg), but have forwarding off.
  - In order for SDNCP to perform pathing service, requires efficient, real-time way of extracting LLDP …not an issue of switch has sufficient CPU resources.

…

© 2012 IBM Corporation

Optimized Fabric

OpenFlow based multipathing Requirement



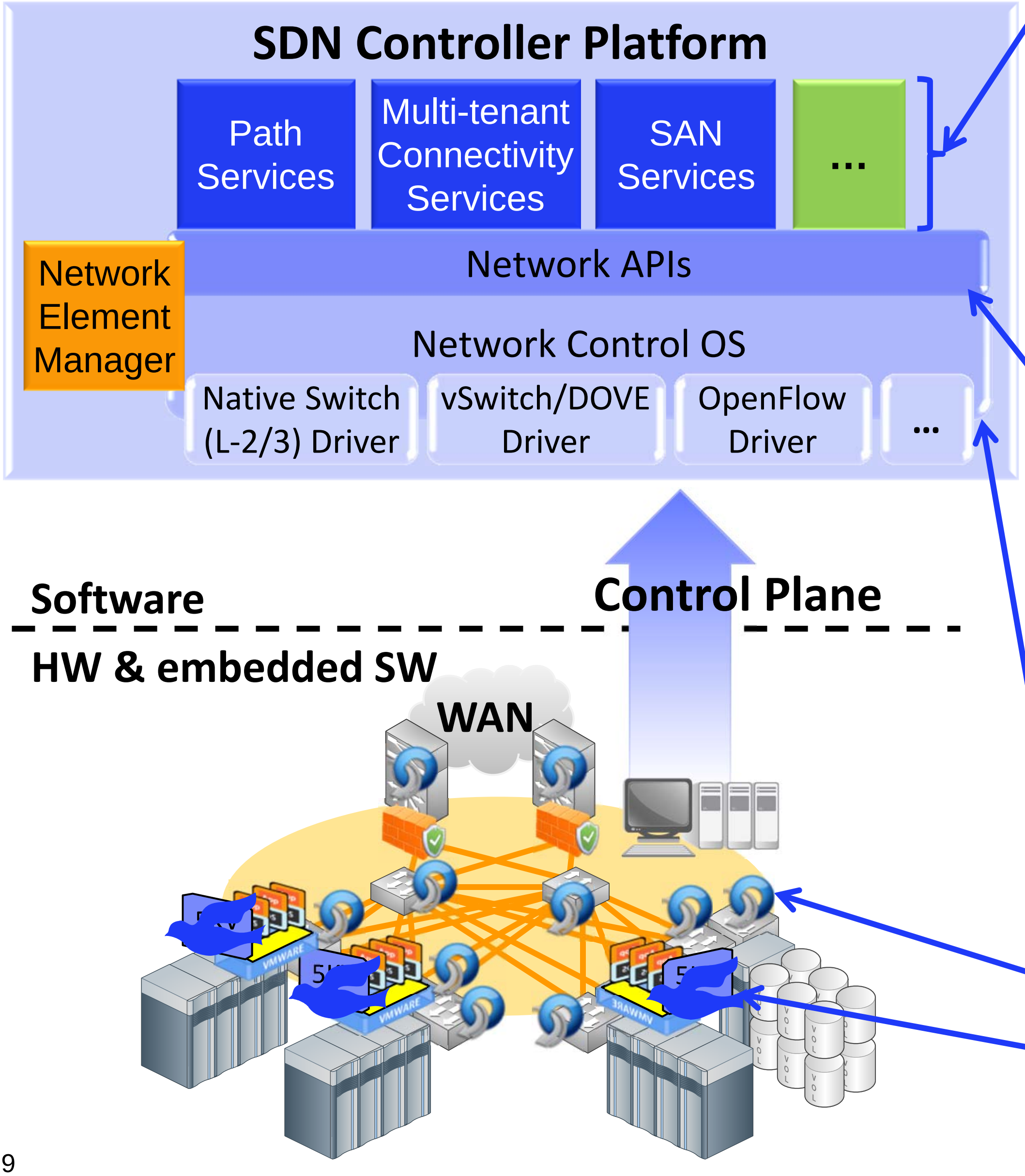
- SDN Platform Controller:
  - Discovers switches and switch adjacencies.
  - Computes physical paths, including disjoint paths
  - Processes all ARPs for IP/Enet (and optionally all FIPs for FCoE), ideally with new ability to request disjoint pathing
  - For virtual environment without DOVE,serves as VSI manager for Qbg
  - Configures switch forwarding tables
  - …

- Each switch:
  - Uses one of the options from the previous page to bring up link and discover/propagate CEE (and Qbg) settings.
  - Has layer-2 forwarding off;
  - Connects to OF Controller.

© 2012 IBM Corporation

Integrated Software

Software Defined Networking Technologies



SDN Controller Platform

Path Services

Multi-tenant Connectivity Services

SAN Services

...

Network Element Manager

Network APIs

Network Control OS

Native Switch (L-2/3) Driver

vSwitch/DOVE Driver

OpenFlow Driver

...

Software

Control Plane

HW & embedded SW

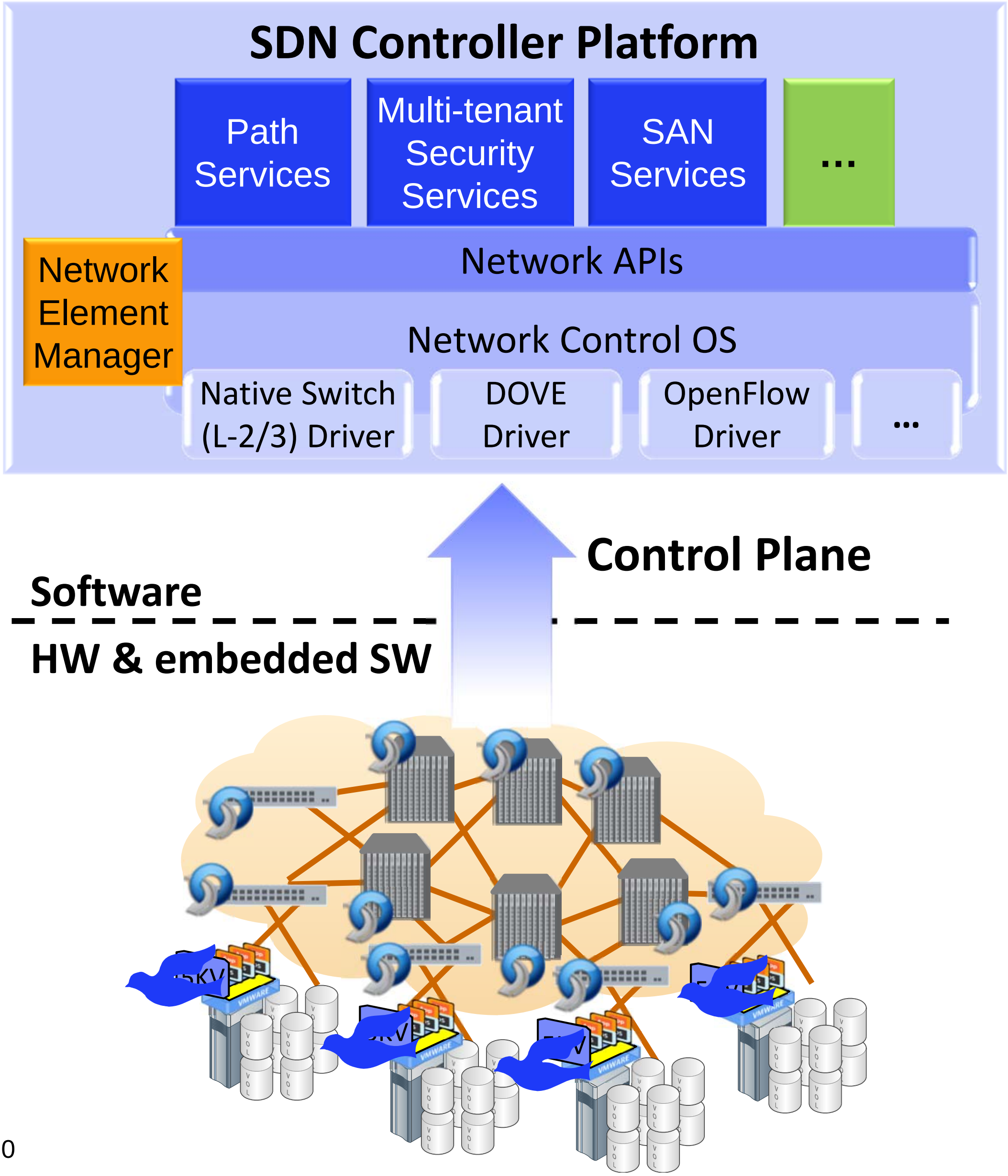
WAN

- **Network functions delivered as services**
  - Multi-tenant connectivity
  - Security
  - Load balancing
  - ...
- **Network APIs provide an abstract interface into underlying controller**
  - Distributes, configures & controls state between services and controller
  - Provides multiple abstract views
- **Network Operating System drives set of devices**
  - Physical devices (e.g. TOR)
  - Virtual devices (e.g. DVS 5000v)

© 2012 IBM Corporation

Integrated Software

SDN Summary



SDN Controller Platform

Path Services

Multi-tenant Security Services

SAN Services

...

Network Element Manager

Network APIs

Network Control OS

Native Switch (L-2/3) Driver

DOVE Driver

OpenFlow Driver

...

Control Plane

Software

HW & embedded SW

- **Network Services value:**
  - Eco-system for network Apps vs today’s closed switch model
- **DOVE Network value:**
  - Cloud scale resource provisioning
  - De-couples virtual network from physical network
- **OpenFlow value:**
  - De-couples switch’s control plane from data plane
  - Data center wide physical network control

© 2012 IBM Corporation

Thank You