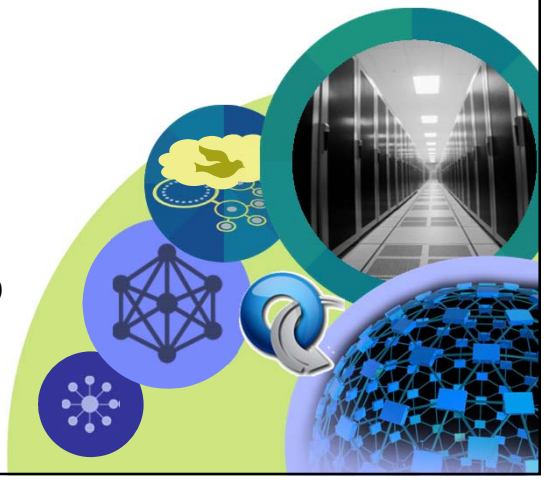
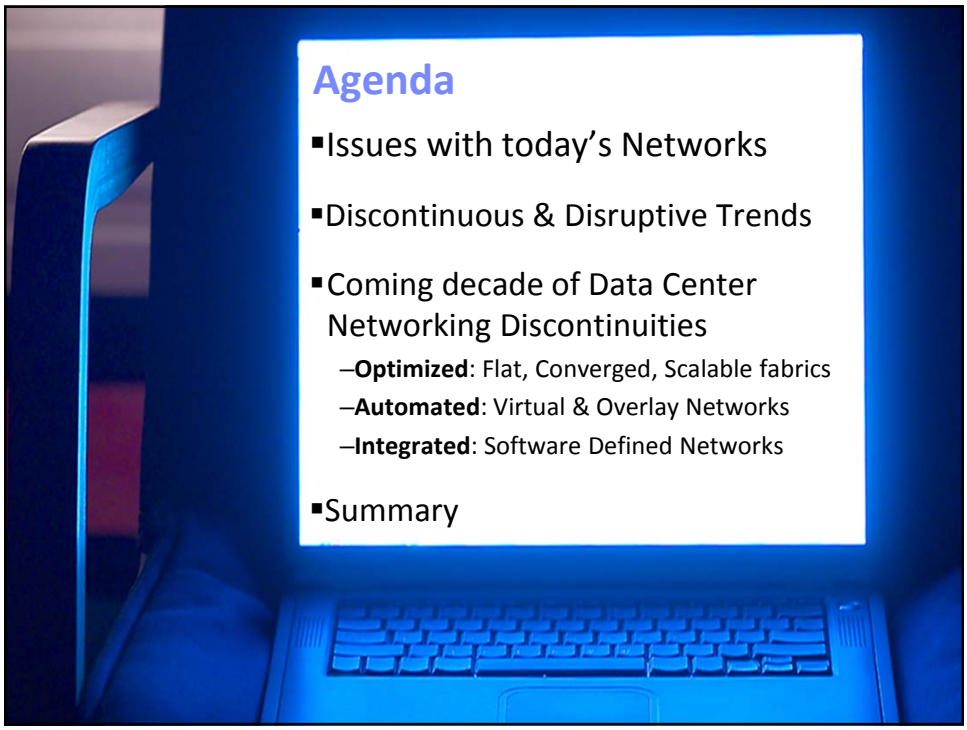
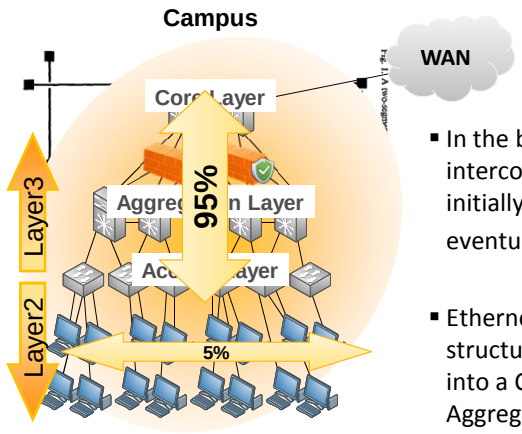
The Coming Decade of Data Center Networking Discontinuities

Renato Recio
IBM Fellow &
System Networking CTO





Early Ethernet Campus Evolution

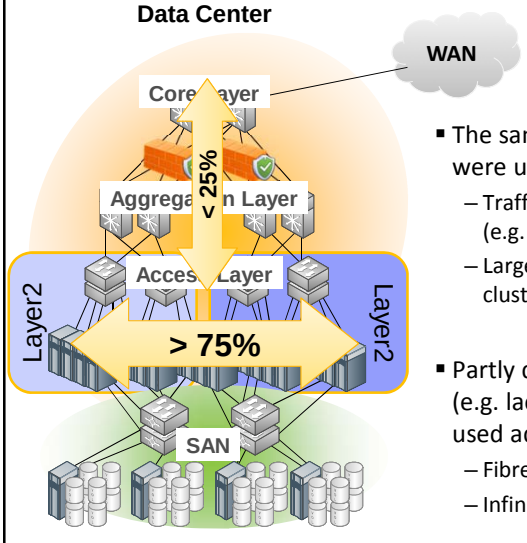


- In the beginning, Ethernet was used to interconnect stations (e.g. dumb terminals), initially through repeater & hub topologies, eventually through switched topologies.
- Ethernet campuses evolved into a structured network typically divided into a Core, Service (e.g. firewall), Aggregation and Access Layer, where:
  - Traffic pattern is mostly North-South (directed outside campus vs peer-peer).
  - To avoid spanning tree problems, campus networks typically are divided at access.

3

© 2012 IBM Corporation

Ethernet in the Data Center



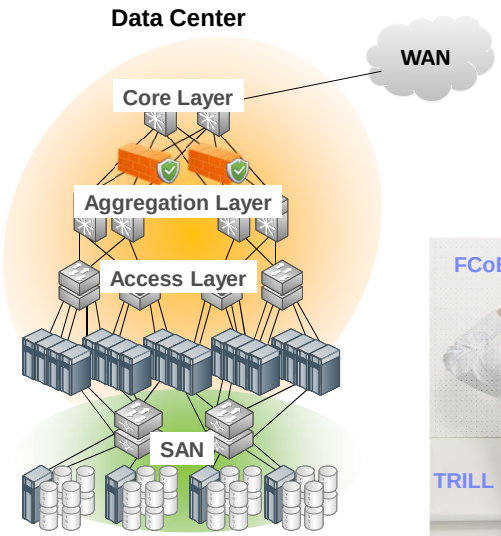
- The same structured network topologies were used in the data center, but…
  - Traffic pattern is mostly East-West (e.g. Application to Database tier).
  - Large layer-2 domains needed for clustering and Virtual Machine mobility.
- Partly due to Ethernet limitations (e.g. lack of flow control), the data center used additional networks, such as
  - Fibre Channel Storage Area Networks (SAN)
  - InfiniBand cluster networks.

4

© 2012 IBM Corporation

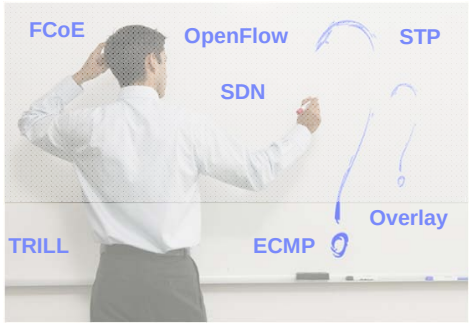
Ethernet in the Data Center

IBM



Net: today's Data Center Network looks similar to Campus networks

But does it still fit to the current requirements of a modern Data Center?

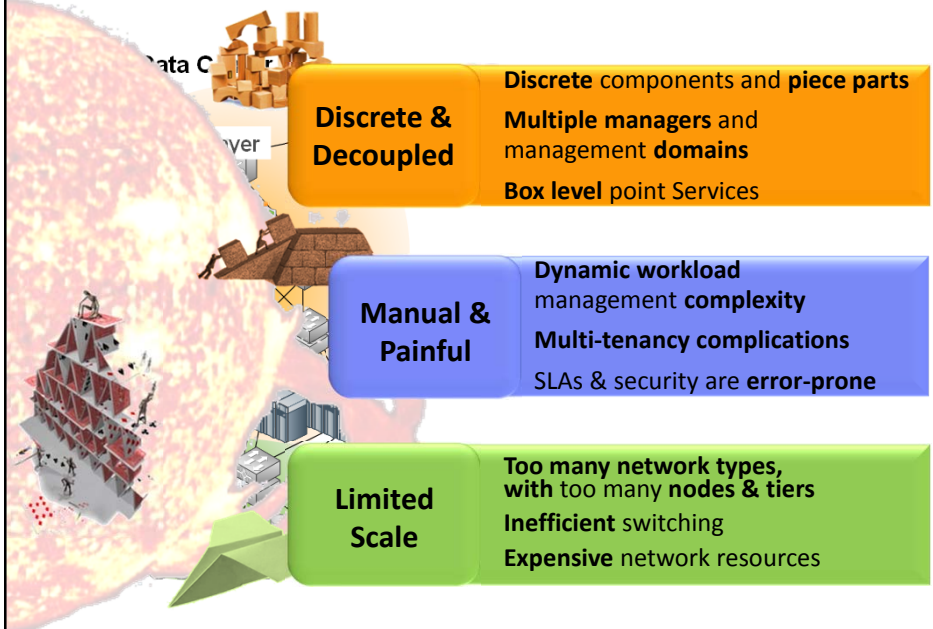


5

© 2012 IBM Corporation

Traditional Data Center Network Issues

IBM



Discrete & Decoupled

Discrete components and **piece parts**

Multiple managers and management **domains**

Box level point Services

Manual & Painful

Dynamic workload management **complexity**

Multi-tenancy complications

SLAs & security are **error-prone**

Limited Scale

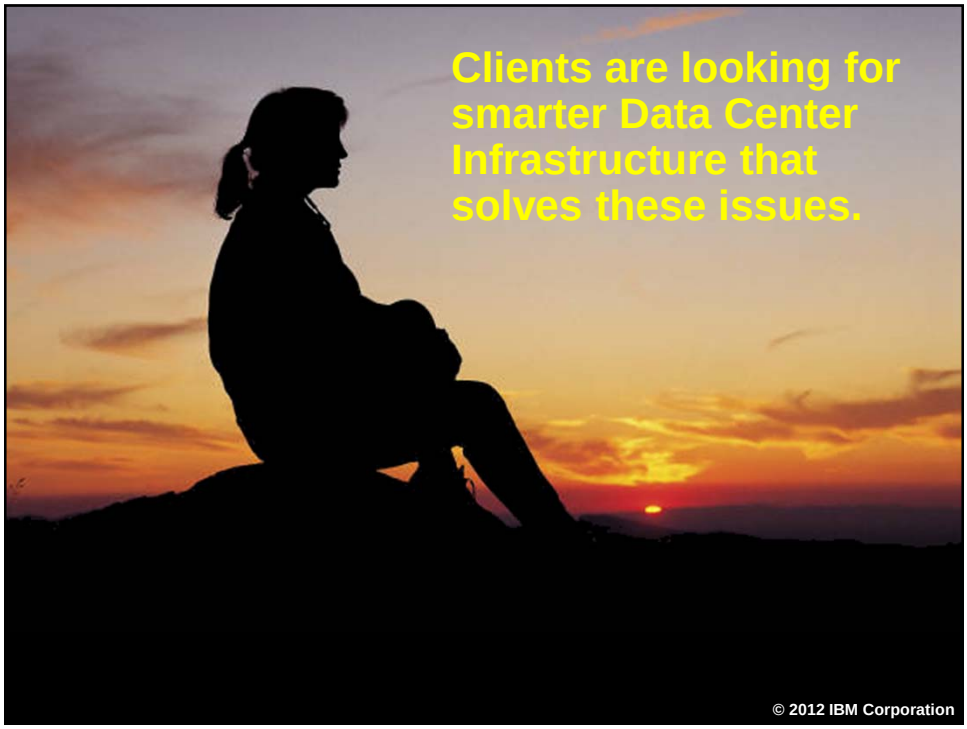
Too many network types, with too many **nodes & tiers**

Inefficient switching

Expensive network resources

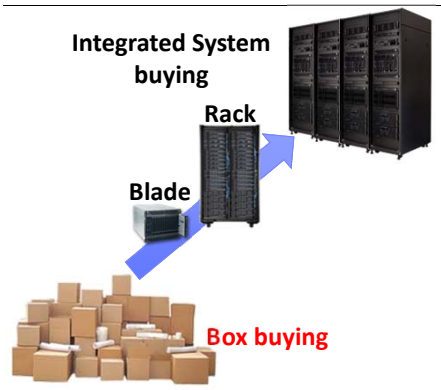
6

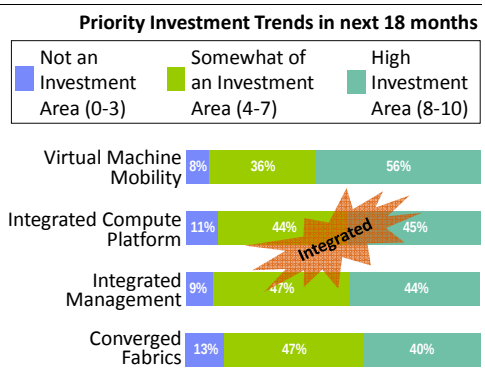
© 2012 IBM Corporation



Customers Want Larger Integrated Systems

IBM





Clients are seeking solutions to the complexities associated with inefficient networking, server sprawl and manual virtualization management.

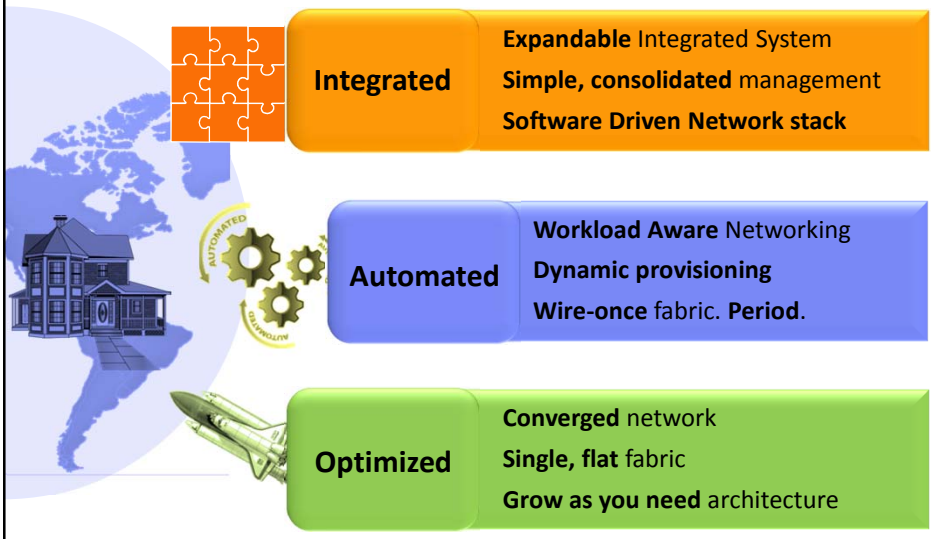
Integrated system pre-packages server, storage, network, virtualization and management and provides an automated, converged & virtualized solution with fast time to value & simple maintenance

8

© 2012 IBM Corporation

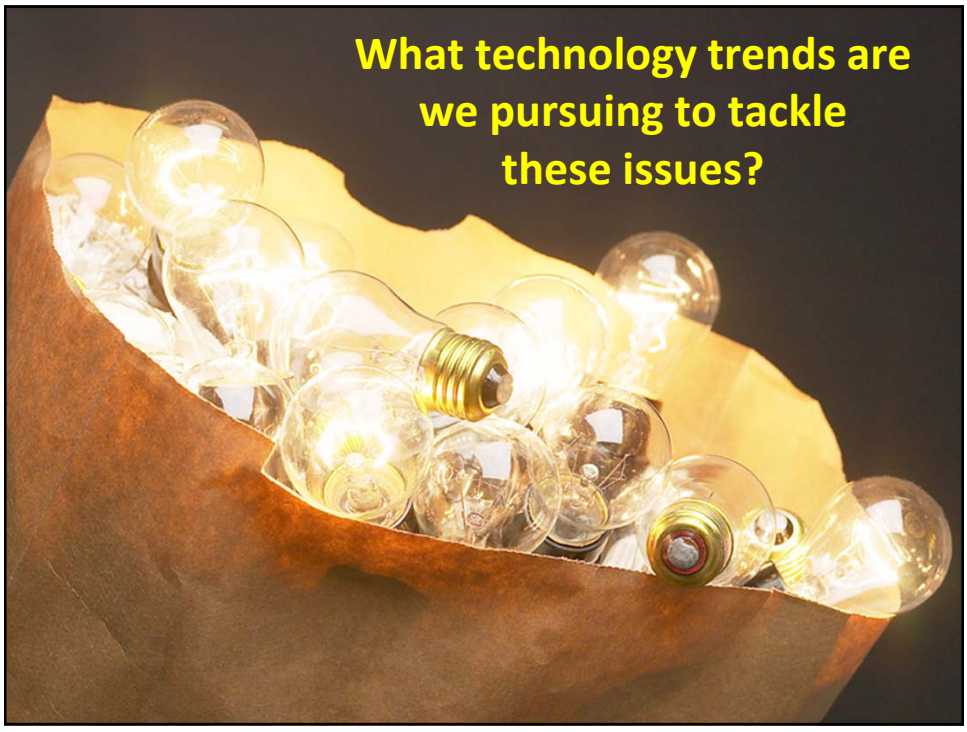
Smarter Data Center Infrastructure

IBM



9

© 2012 IBM Corporation



Discontinuous Technologies

Discontinuity – a: the property of being not mathematically continuous; b: an instance of being not mathematically continuous; *especially a value of an independent variable at which a function is not continuous*

11

Discontinuous Technologies Examples

Distributed Overlay Virtual Ethernet (DOVE) Networks

TRILL, with disjoint multi-pathing

Converged Enhanced Ethernet

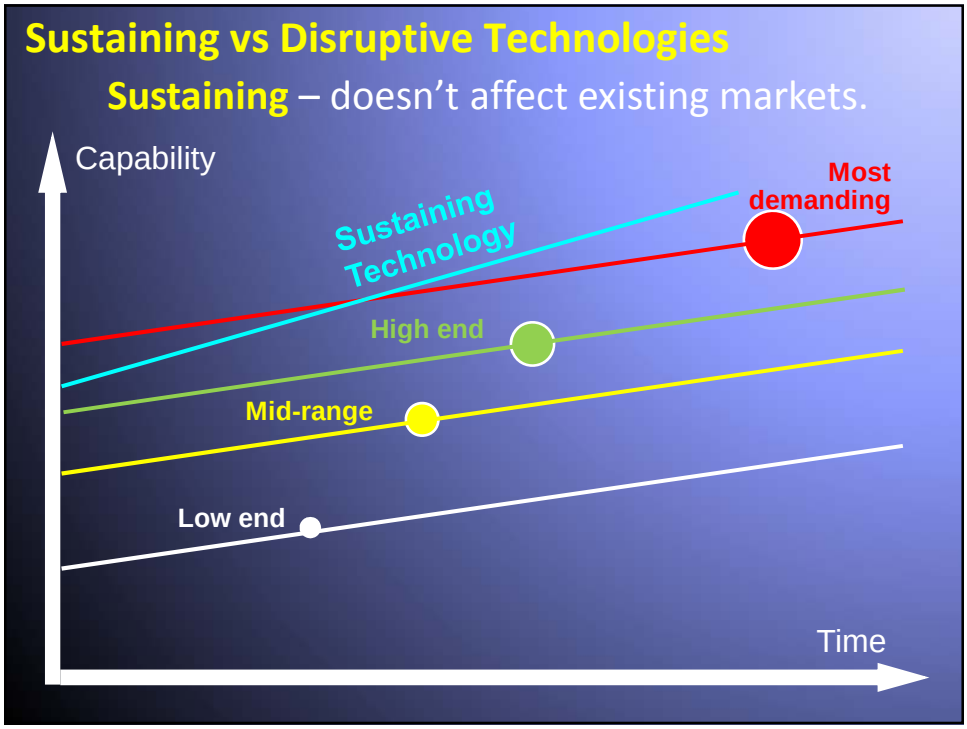
Fibre Channel over Ethernet

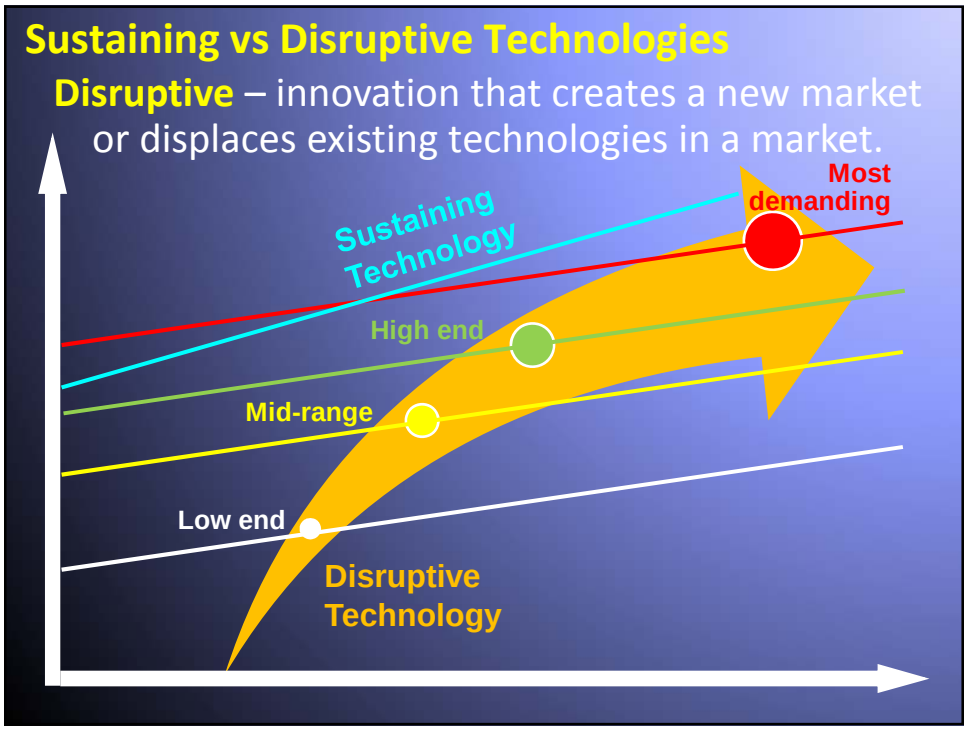
Software Defined Networking

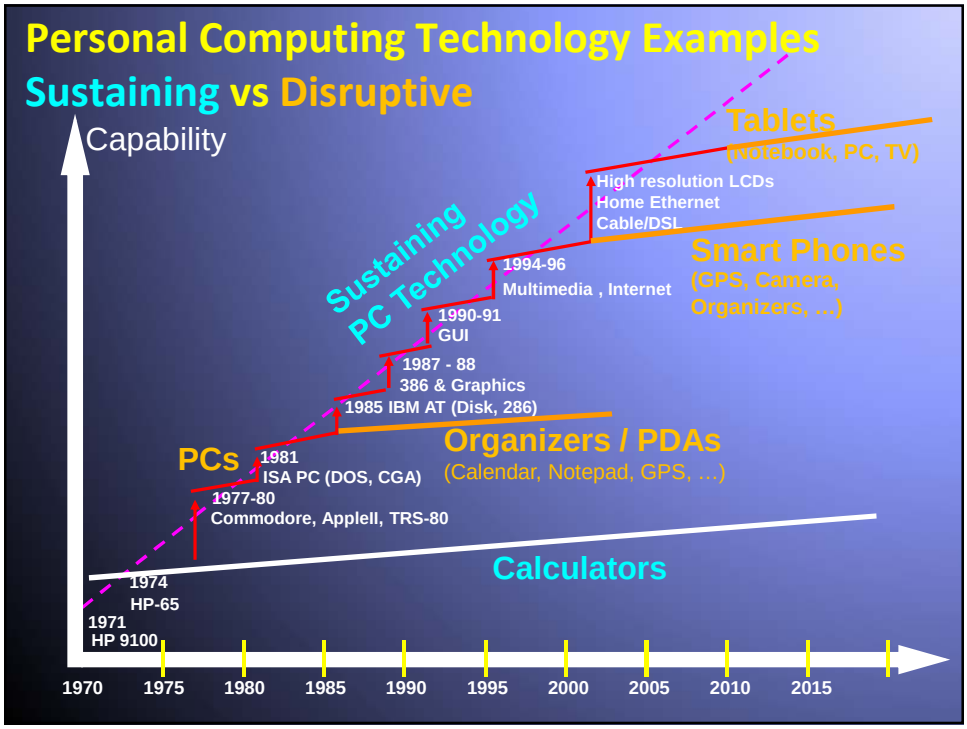
OpenFlow

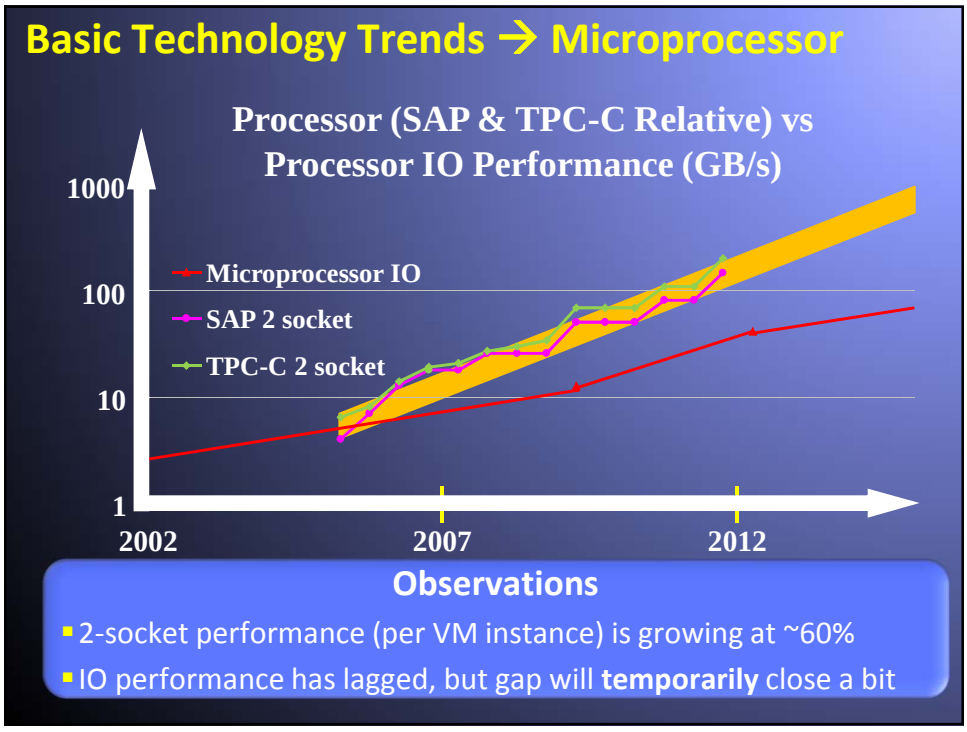
(more later ☺)

12

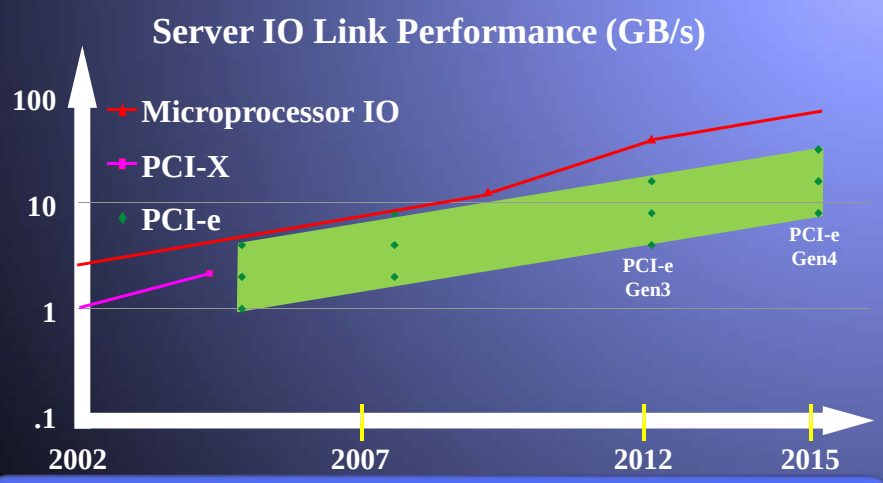

Sustaining vs Disruptive Technologies
Sustaining – doesn't affect existing markets.
Capability
Most demanding
Sustaining Technology
High end
Mid-range
Low end
Time


Sustaining vs Disruptive Technologies
Disruptive – innovation that creates a new market or displaces existing technologies in a market.
Most demanding
Sustaining Technology
High end
Mid-range
Low end
Disruptive Technology


Personal Computing Technology Examples
Sustaining vs Disruptive
Capability
Tablets
(Notebook, PC, TV)
High resolution LCDs
Home Ethernet
Cable/DSL
Smart Phones
(GPS, Camera, Organizers, …)
Sustaining PC Technology
1994-96
Multimedia , Internet
1990-91
GUI
1987 - 88
386 & Graphics
1985 IBM AT (Disk, 286)
Organizers / PDAs
(Calendar, Notepad, GPS, …)
PCs
1981
ISA PC (DOS, CGA)
1977-80
Commodore, AppleII, TRS-80
Calculators
1974
HP-65
1971
HP 9100
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015


Basic Technology Trends → Microprocessor
Processor (SAP & TPC-C Relative) vs
Processor IO Performance (GB/s)
1000
100
10
1
Microprocessor IO
SAP 2 socket
TPC-C 2 socket
2002
2007
2012
Observations
2-socket performance (per VM instance) is growing at ~60%
IO performance has lagged, but gap will temporarily close a bit

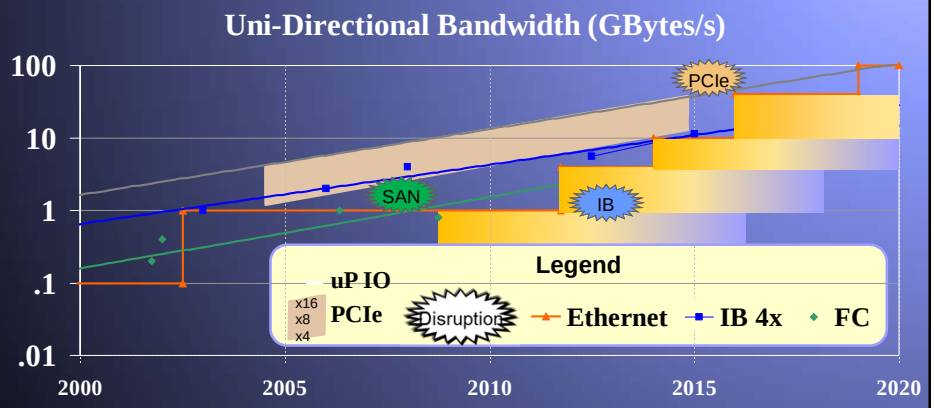
Basic Technology Trends → Server IO Links



Observations

- PCIe Generation 3 is tracking traditional uP IO bandwidth trend
- 1 PCIe Gen3 can support 2x 40GE (4x 40GE, but not at link rate)

Basic Technology Trends → Server IO & Fabrics



Observations

- Ethernet, with Data Center enhancements (more next), is a disruptive technology for SAN & cluster markets (now), and PCIe market in the future.

Emerging Disruptive IP/Ethernet Technologies

Converged stacks: CEE, FCoE, RoCE

Layer-2 based multi-pathing: TRILL, Proprietary fabrics & OpenFlow

Software Defined Networking stack

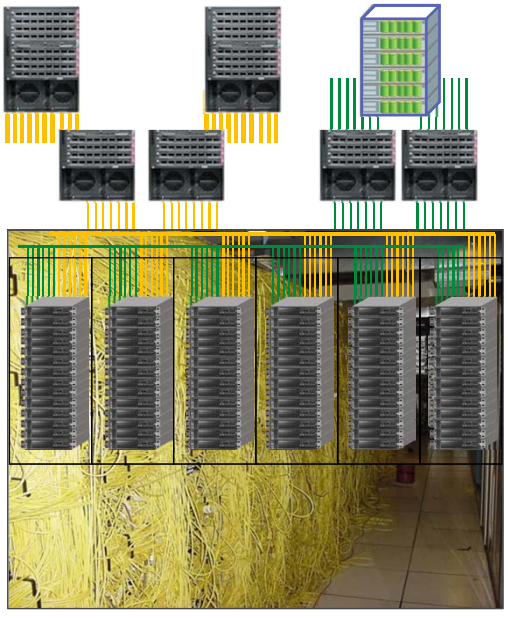
Disrupted markets:

- **Fibre Channel SAN**
- **InfiniBand Clusters**
- **Enterprise Networks**

Limited Scale

Traditional Data Center Networks

IBM



1. Multiple fabrics (SAN, LAN, …) because Ethernet lacks convergence services:
   - Lossy, no bandwidth guarantee, no multi-pathing, not VM aware
2. Inefficient topology, with too many tiers & nodes
   - Per hop forwarding
   - Wasted paths (due to STP)
   - East-West traffic forced to go North-South (e.g. due to multi-cast containment or proprietary, VNTag, approach)
3. Server-direct-to-core makes cabling a mess

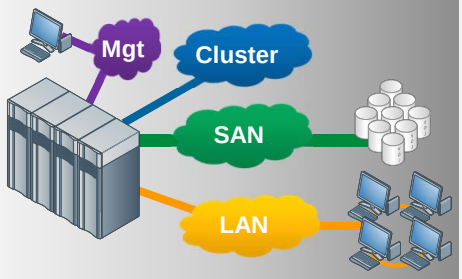
20

© 2012 IBM Corporation

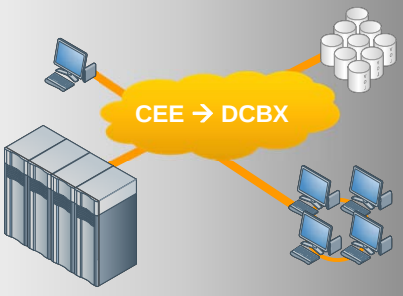
Optimized

CEE the Path to Converged Fabrics

IBM



- Multiple fabrics
- Complex management
- Inefficient RAS



- Converged fabric
- Simplified management
- Improved RAS

© 2012 IBM Corporation

Optimized

Data Center Convergence Enabling Technologies

IBM

Remote Direct Memory Access

FC Forwarders, FC Data Forwarders, FC Snooping Bridges

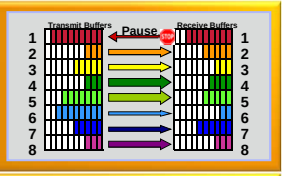
RDMA over Ethernet

Fibre Channel over Ethernet

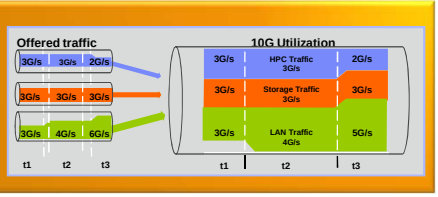
Priority Based Flow Control

Enhanced Transmission Selection

Per priority pause

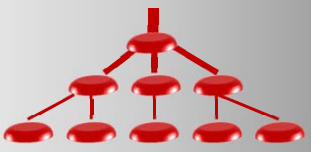


Best available & per priority bandwidth guaranteed services

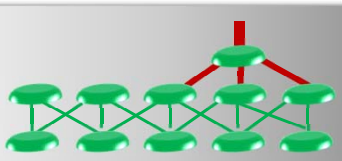


© 2012 IBM Corporation

Optimized Traditional Arcane DCN → Mesh Ahead



- Multi-tiered tree topologies
- High oversubscription
- Expensive, high bandwidth uplinks
- Small layer-2 fabric
- Robustness of higher tier products has been a concern



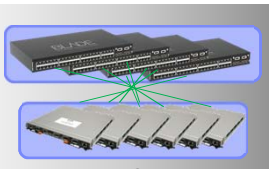
- Mesh, Clos, Jellyfish topologies
- Oversubscription only to WAN/core
- High cross sectional B/W (cheap TOR B/W)
- Layer-2 scaling options (more next)
- Robust, HA topologies

© 2012 IBM Corporation

Optimized Integrated System Fabric Trends

Standalone

- *Optimal local switching*
- *Low latency*
- *Converged*
- *Full L2/3*



Stacking

Single stacked switch

+

- *Active-active multi-link redundancies*
- *Single logical switch*

Technology Trend

Fabric

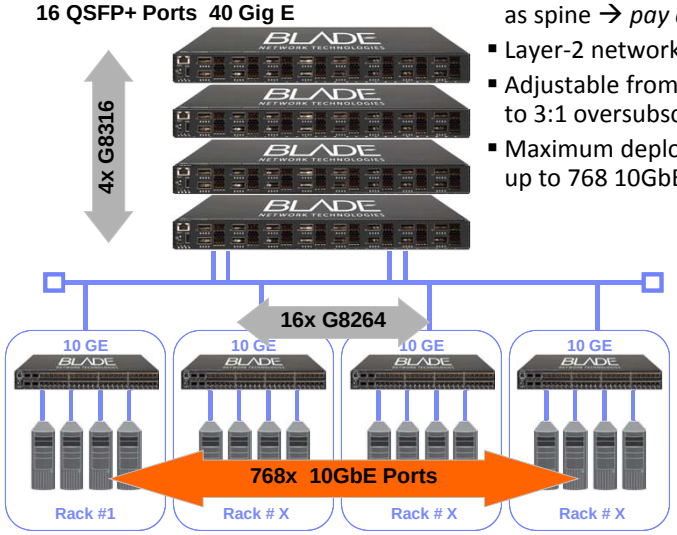
Switch cluster

+

- *Arbitrary topologies (fat-tree, Clos, Jellyfish…)*
- *Cross-sectional bandwidth scaling*
- *Multiple redundant paths*
- *Scales to many ports*
- *Unifies physical & virtual*

© 2012 IBM Corporation

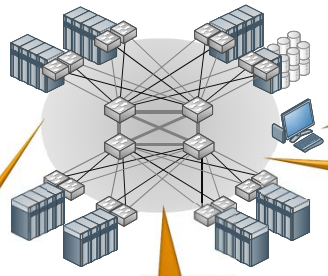
Optimized Stacking Example: 40GE TOR Switch



- Stacked 40GE TOR switch (e.g. G8316) as spine → *pay as you grow*
- Layer-2 network state is at rack level
- Adjustable from non-blocking to 3:1 oversubscription
- Maximum deployment up to 768 10GbE Server Ports

© 2012 IBM Corporation

Optimized Fabric Technology Options



Layer-3

- *Established technology*
- *Standards based*
- *Distributed control plane*
- *Large scalability*
- *HA with fast convergence*
- *Small layer 2 without DOVE network*
- *Many devices to manage*

TRILL

- *Large layer-2*
- *Distributed control plane*
- *Large scalability*
- *HA with fast convergence*
- *Emerging technology (some proprietary)*
- *Single disjoint multi-path fabric may need new RFC*

OpenFlow

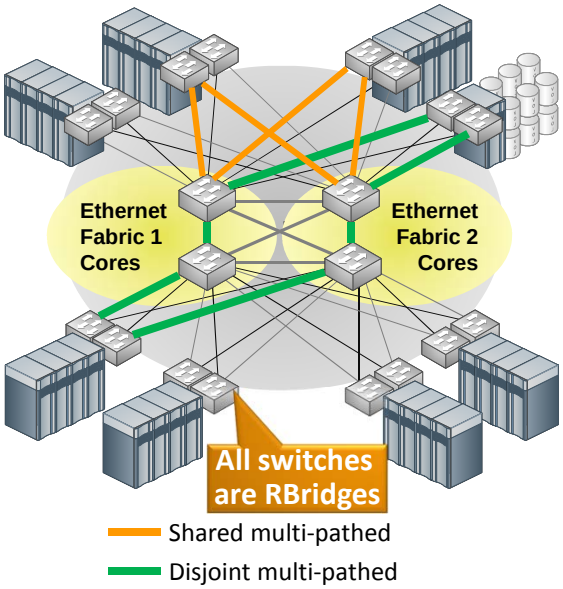
- *Large layer-2*
- *Distributed control plane*
- *Large scalability*
- *HA with fast convergence*
- *Enables network functions delivered as Services → e.g. disjoint multi-pathing (more later)*
- *Emerging technology*
- *Client acceptance barrier*

© 2012 IBM Corporation

Optimized

Shared vs Disjoint Multi-pathing

IBM

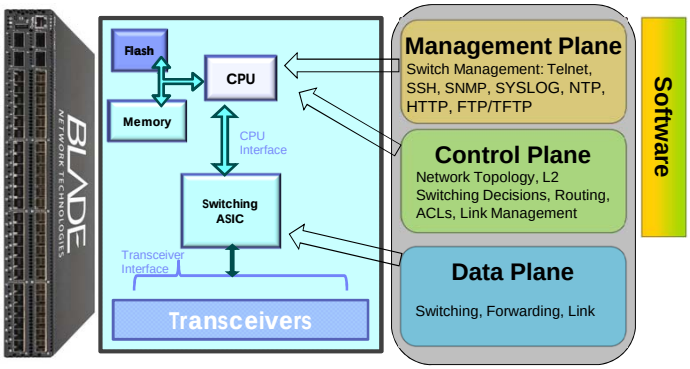


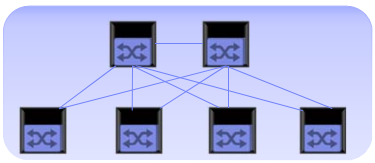
- Data Center storage and cluster fabrics require full path redundancy.
- Completely redundant Ethernet fabrics meet this requirement, but come with an administration burden (e.g. 2x SAN configuration & maintenance).
- Dual-TRILL fabrics are emerging today.
  - Single TRILL fabric, with disjoint multi-pathing would eliminate administration burden.

© 2012 IBM Corporation

Traditional Network

IBM



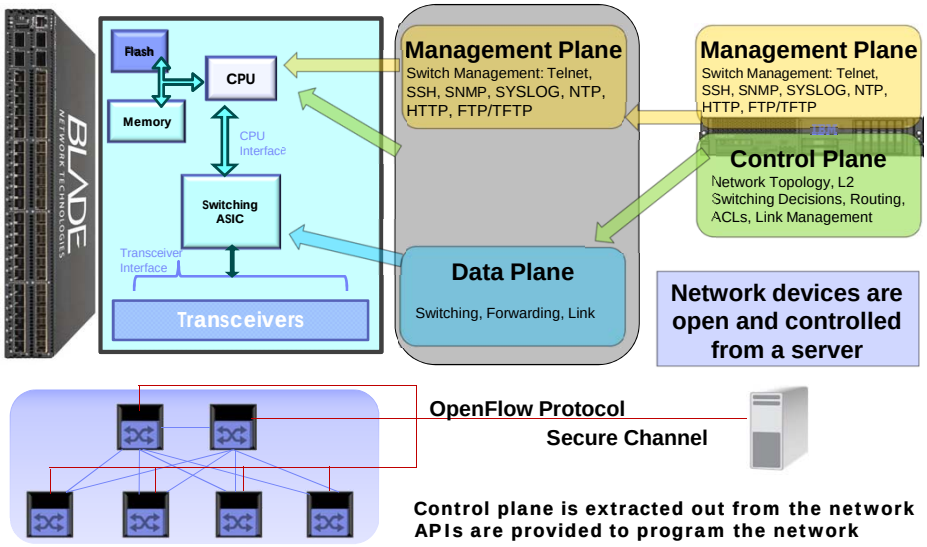


Network devices are closed system

Each Network element has its own control and management plane

© 2012 IBM Corporation

OpenFlow Network

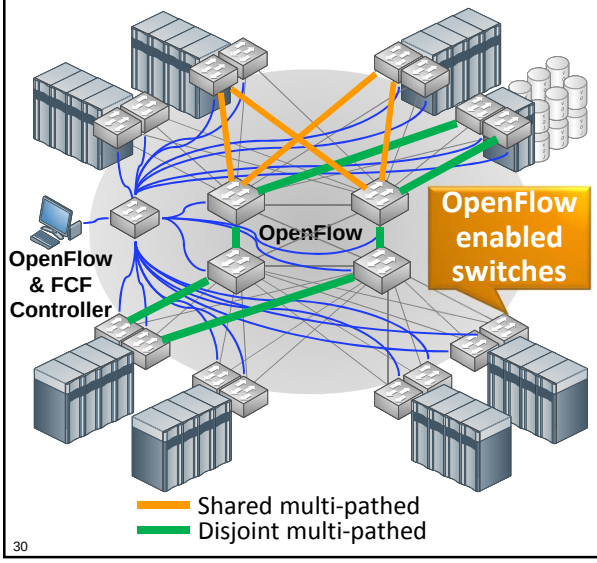


© 2012 IBM Corporation

Optimized

OpenFlow Based Fabric Overview

OpenFlow can also be used to create a disjoint multi-pathing fabric.



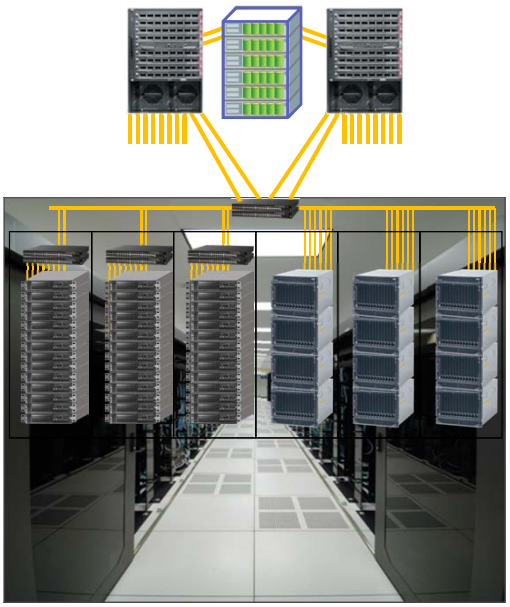
- Each switch has Layer-2 forwarding turned off.
- Each switch connects to OpenFlow Controller.
- OFC discovers switches and switch adjacencies.
- OFC computes disjoint physical paths and configures switch forwarding tables.
- All ARPs go to OFC.

© 2012 IBM Corporation

Optimized

Smart Data Center Networks

IBM



1. Converged fabric (LAN, SAN, Cluster, ...)
   - Lossless
   - Bandwidth allocation
   - Disjoint multi-pathing
   - VM workload aware
2. Efficient topology, with few tiers & nodes
   - Efficient forwarding
   - Disjoint multi-path enabled
   - Optimized East-West flows
3. Servers connect to switches within rack (eliminates cabling mess)

31

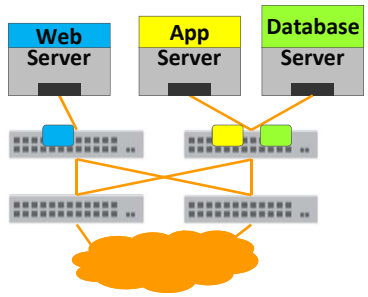
© 2012 IBM Corporation

Manual & Painful

Virtualization increased network management complexity

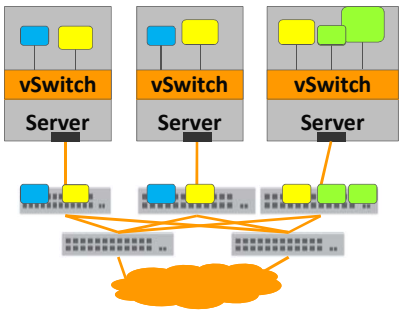
IBM

Before Virtualization



- Static workloads ran on bare-metal OS
- Each workload had network state associated with it.
- Physical network was static & simple (configured once)

Physical Network with vSwitches



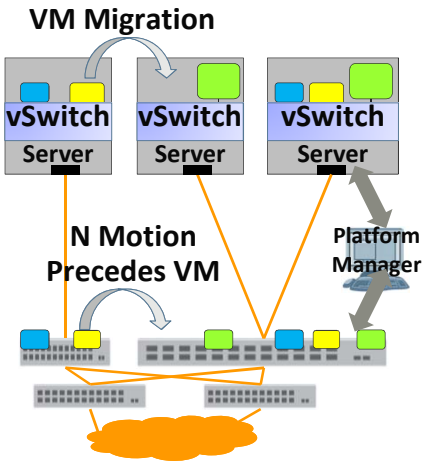
- Server virtualization = dynamic workloads
- VM's network state resides in vSwitch/DCN
- Physical network is dynamic and more complex (VMs come up dynamically & move)

32

© 2012 IBM Corporation

Automated Hypervisor vSwitch Automation

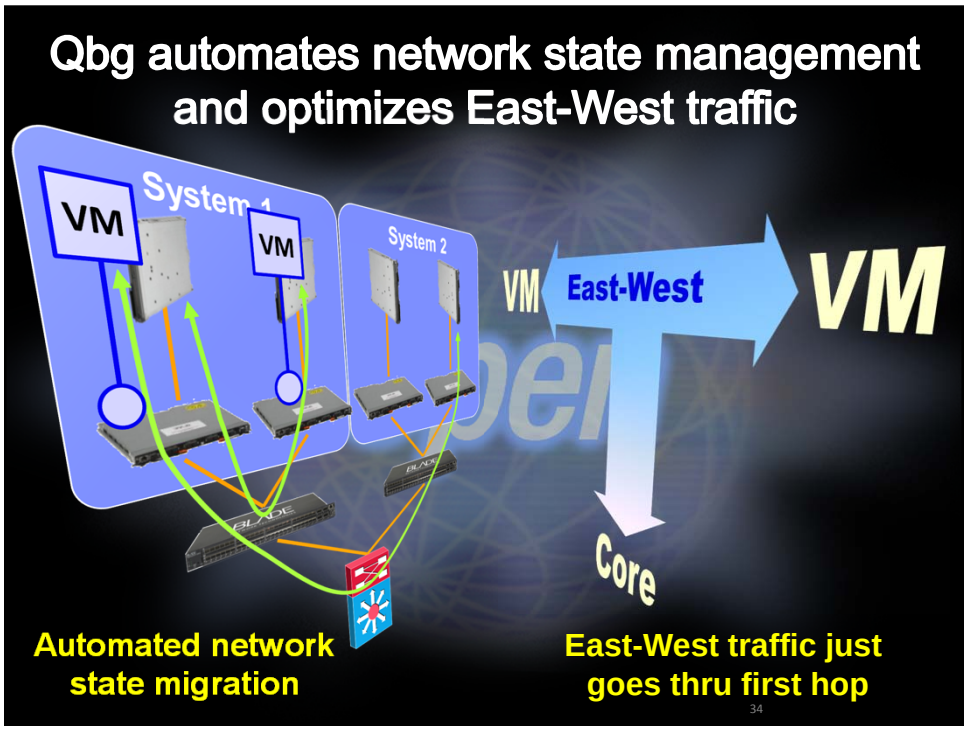
IBM



1. Per-VM switching in HW
2. Hypervisor Vendor Agnostic
3. Platform Manager Integration
4. Standards based (IEEE 802.1Qbg)
5. Network state migrates ahead of VM

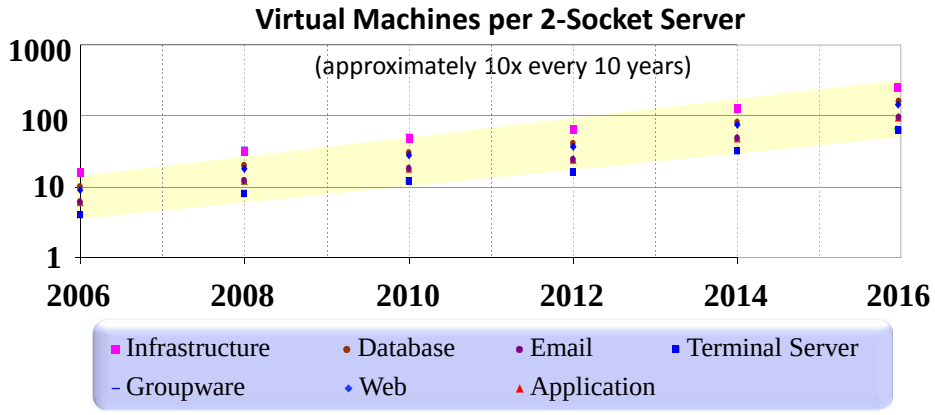
33

© 2012 IBM Corporation



Automated Network Virtualization Trends

IBM



- Number of VMs per socket is rapidly growing (10x every 10 years).
  - Increases amount of VM-VM traffic in Enterprise Data Centers (e.g. co-resident Web, Application & Database).
  - VM growth increases network complexity associated with creating/migrating: layer-2 (VLANs, ACLs…) & layer-3 (e.g. Firewall, IPS) attributes.

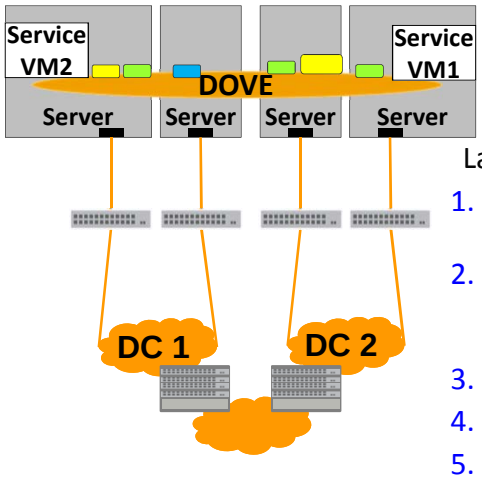
35

© 2012 IBM Corporation

Automated Hypervisor Network Virtualization

IBM

Technology Trend



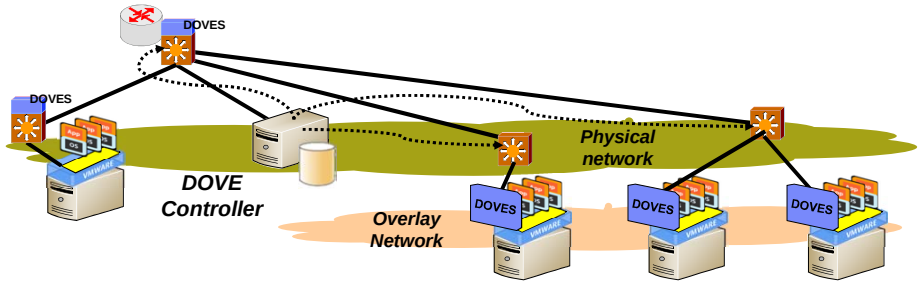
Layer-2 vSwitch features, plus:

1. Layer-3 Distributed Overlay Virtual Ethernet (DOVE)
2. Simple "configure once" network (physical network doesn't have to be configured per VM).
3. De-couples virtual from physical
4. Multi-tenant aware
5. Enables cross-subnet virtual appliance services (e.g. Firewall, IPS)

36

© 2012 IBM Corporation

Automated Overview of DOVE Technology Elements

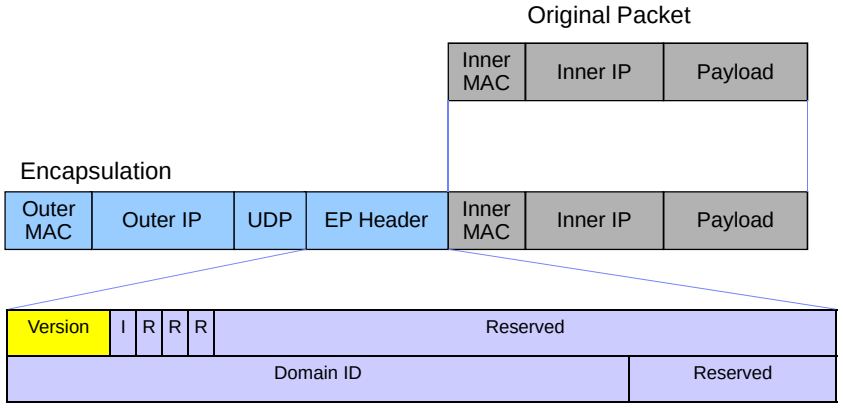


- DOVE Controller
  - Performs management & a portion of control plane functions across DOVE Switches
- DOVE Switches (DOVES)
  - Provides layer-2 over UDP overlay (e.g. based on OTV/VXLAN)
  - Performs data and some control plane functions
  - Runs in Hypervisor vSwitch or gateways
  - Provides interfaces for Virtual Appliances to plug into (Analogous to appliance line-cards on a modular switch)

37 © 2012 IBM Corporation

Automated DOVE Technology

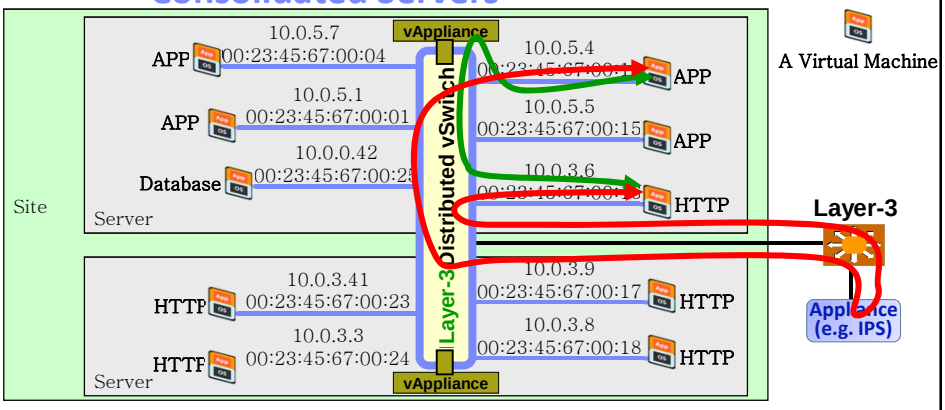
VXLAN based Encapsulation Example



Encapsulation Protocol (EP) Header (e.g. VXLAN based)
(VXLAN extension in Yellow → necessary IETF version field)

38 © 2012 IBM Corporation

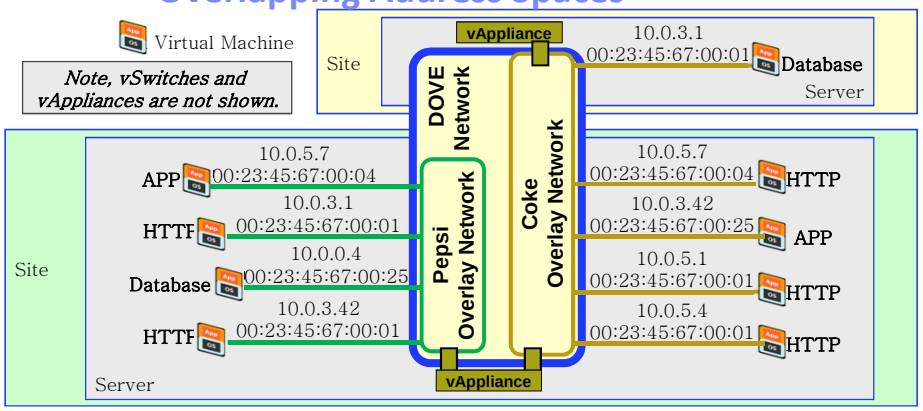
Automated Improving Networking Efficiency for Consolidated Servers



- Hypervisor vSwitches enable addition of virtual appliances (vAppliances), which provide secure communication across subnets (e.g. APP to Database tier).
  - However, all traffic must be sent to an external Layer-3 switch, which is inefficient considering VM/socket growth rates and integrated servers.
- To solve this issue requires cross-subnet communications in Hypervisor's vSwitch.

39 © 2012 IBM Corporation

Automated Multi-Tenant with Overlapping Address Spaces



- Multi-tenant, Cloud environments require multiple IP address spaces within same server, within a Data Center and across Data Centers (see above).
  - Layer-3 Distributed Overlay Virtual Ethernet (DOVE) switches enable multi-tenancy all the way into the Server/Hypervisor, with overlapping IP Address spaces for the Virtual Machines.

40 © 2012 IBM Corporation

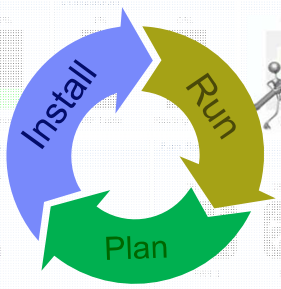
Integrated

Systems Network Element Manager SNEM

IBM

A comprehensive tool to automate Virtualized Data Center Network workflows

Engineer
- Physical
- Firmware
- Configuration
- Advanced PD



Operator
- Monitoring
- Initial PD
- Automation

Planner (or Engineer)

Investment = business direction + application requirements + utilization trends

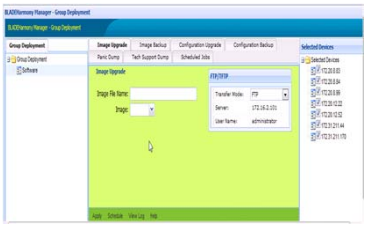
© 2012 IBM Corporation

Integrated

SNEM Examples

IBM

Engineer

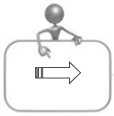


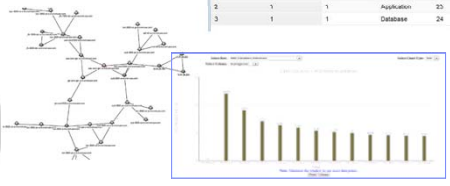
Perform Efficient Firmware or Configuration Updates to Multiple Switches

Operate

Automate VM network resident port profiles and converged fabric Quality of Service

Plan





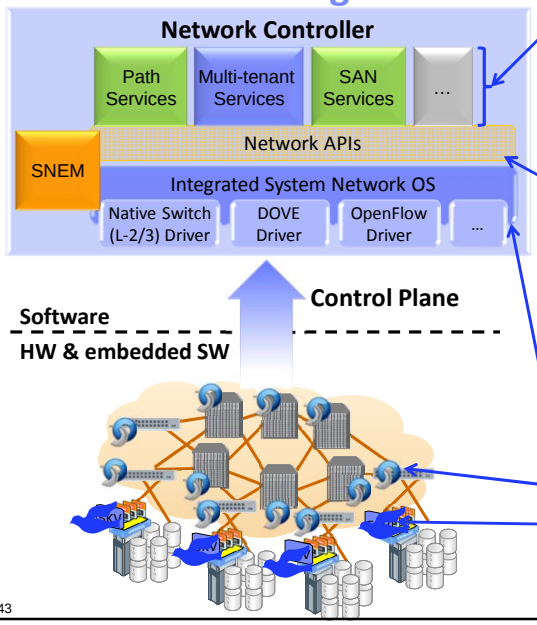
Performance trend & root-cause analysis, fault management, ..

© 2012 IBM Corporation

Integrated

Software Defined Networking Technologies

IBM



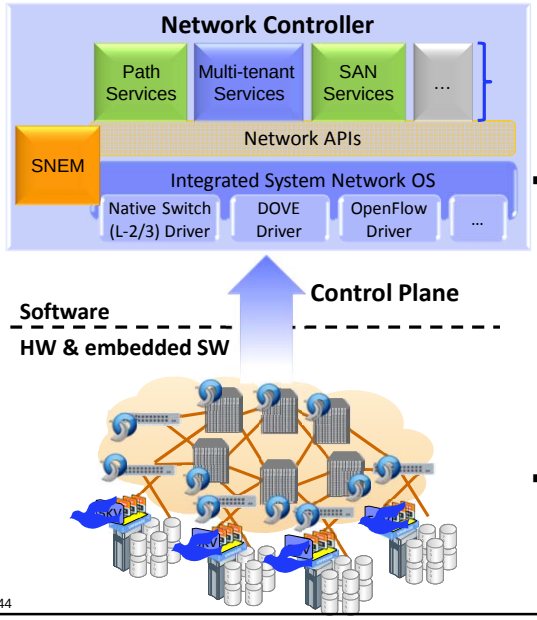
- **Network functions delivered as services**
  - Multi-tenant VM security
  - Virtualized load balancing
- **Network API's provides an abstract interface into underlying controller**
  - Distributes, configures & controls state between services & controllers
  - Provides multiple abstract views
- **Network Operating System drives set of devices**
  - Physical devices (e.g. TOR)
  - Virtual devices (e.g. vSwitch)

© 2012 IBM Corporation

Integrated

Software Defined Networking Value

IBM



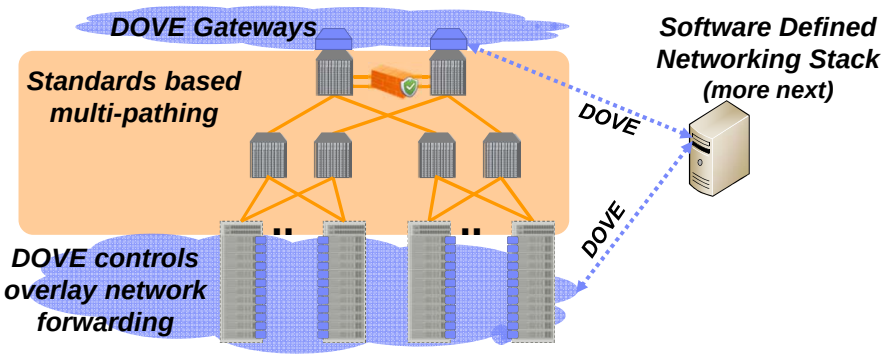
- **Network Services value:**
  - Eco-system for network Apps vs today's closed switch model (think smart-phone App model)
  - Examples: SAN (FC) services; Application performance monitor…
- **DOVE Network value:**
  - Virtualized network resource provisioning
  - De-couples virtual network from physical network
  - Simple "configure once" network (physical network doesn't have to be configured per VM).
  - Cloud scale (e.g. multi-tenant)
- **OpenFlow value:**
  - De-couples switch's control plane from data plane
  - Data center wide physical network control

© 2012 IBM Corporation

Integrated

DOVE Technology + Multi-pathing

IBM



- DOVE network simplifies virtual machine network
  - Enables multi-tenancy all the way to the VM
  - Enables single MAC Address per physical server (2 for HA)
  - Significantly reduces size of physical network TCAM & ACL tables
  - Increases layer-2 scale within Data Center and across Data Centers, by decoupling VM's layer-2 from physical network
  - Qbg automates layer-2 provisioning , DOVE automates layer 3-7 provisioning
- Standards based multi-pathed physical network

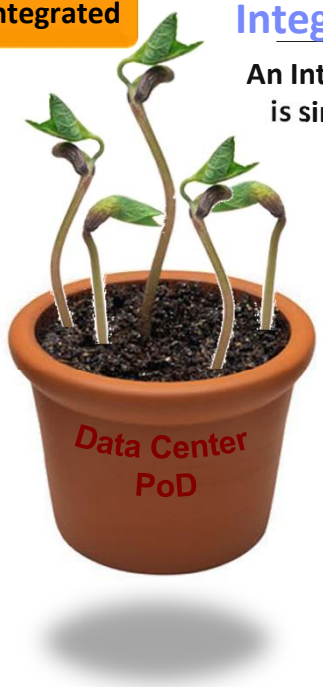
45

© 2012 IBM Corporation

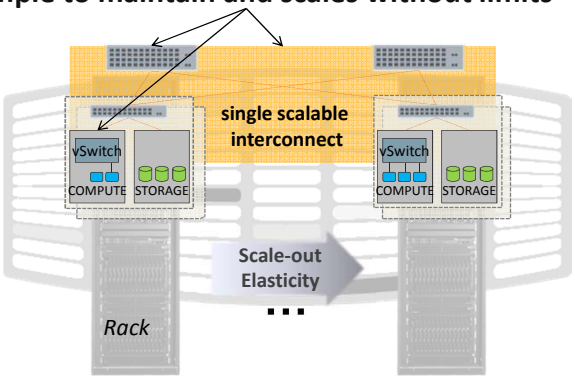
Integrated

Integrated System

IBM



An Integrated System: provides fast time to value, is simple to maintain and scales without limits



Optimized: Converged, multi-path fabric High-performance integrated systems

Automated: Network virtualization

Integrated: Software Defined Networking

46

© 2012 IBM Corporation

Thank You

Renato J Recio
IBM Fellow & Systems Networking CTO

11400 Burnett Road
Austin, TX 78758
512 973 2217

recio us ibm com

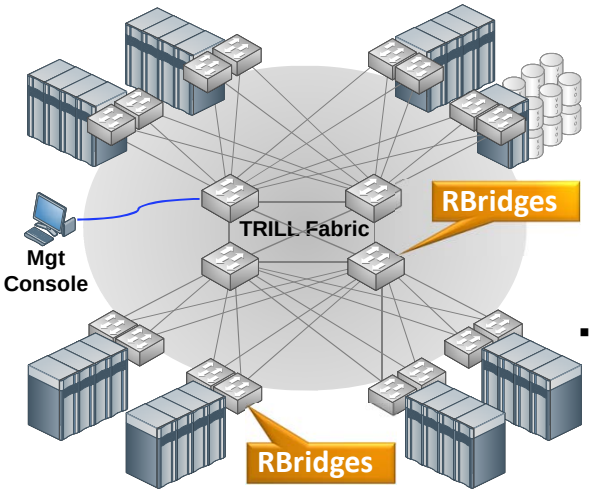
47

© 2012 IBM Corporation

Optimized

TRILL Based Fabric Overview

IBM



- TRILL provides multi-pathing through the network
  - Uses IS-IS to provide shortest layer-2 path
  - Works with arbitrary topologies
  - Easy implementation – minimum configuration
  - Scalable deployment, from few to large number of switches
- Switches that support TRILL use IETF Routing Bridge (Rbridge) protocol
  - Rbridges discovers each other and distribute link state
  - Rbridges use TRILL header to encapsulate packets

49 TRILL = Transparent Interconnection of Lots of Links

© 2012 IBM Corporation

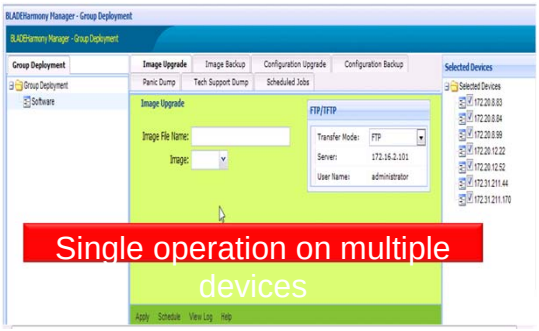
Integrated

Firmware & Configuration

IBM

- Group management
  1. Simultaneous management operations to multiple switches
  2. Back-up, restore, compare & maintain configuration history
  3. Firmware upgrades
  4. Control operations
  5. CLI script execution

Engineer
- Perform Efficient Firmware or Configuration Updates to Multiple Switches

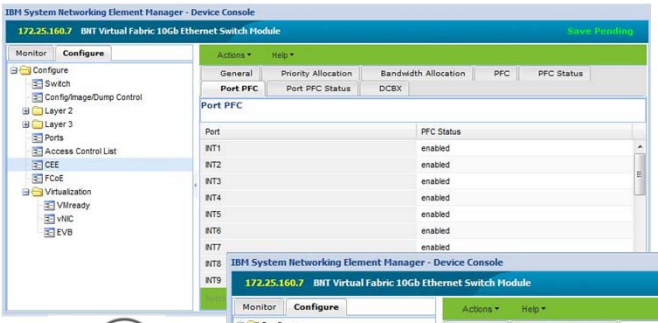


50

© 2012 IBM Corporation

Integrated Converged Network Setup

IBM



Converged Enhanced Ethernet management

- Priority Flow Control (PFC)
- Bandwidth allocation by Priority Group [a.k.a. Enhanced Transmission Selection (ETS)]

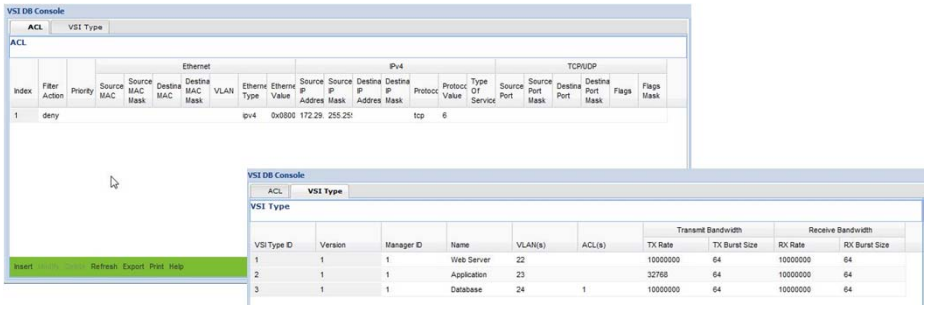
Operator

- Monitoring
- Initial PD
- Automation

51

Integrated Port Profile (VSI Database) Setup

IBM



Operator

- Monitoring
- Initial PD
- Automation

- Full VM Migration with connectivity persistence requires Port Profile (VSI database in 802.1Qbg terms) automation
- Qbg set up:
  - ✓ VLANs for each VSI type
  - ✓ Access Control Lists
  - ✓ Send & Receive Bandwidth

52

© 2012 IBM Corporation

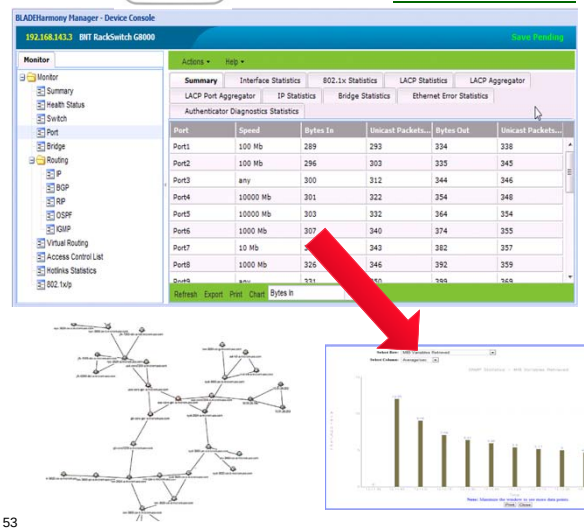
Integrated

Performance Management

IBM

Planner (or Engineer)

Investment = business direction + application requirements + utilization trends



Real-time performance monitoring of switch statistics

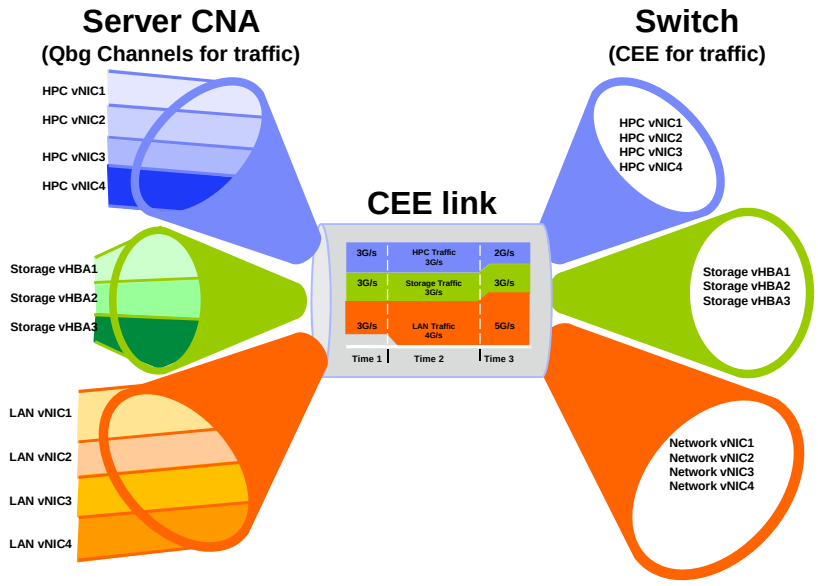
Fault management & root-cause analysis

Performance data trend charts

© 2012 IBM Corporation

Example of Qbg Channels and CEE

IBM



© 2012 IBM Corporation