A Scalable Networking Architecture for Improving Performability Within and Across Data Centers
A Systems Perspective

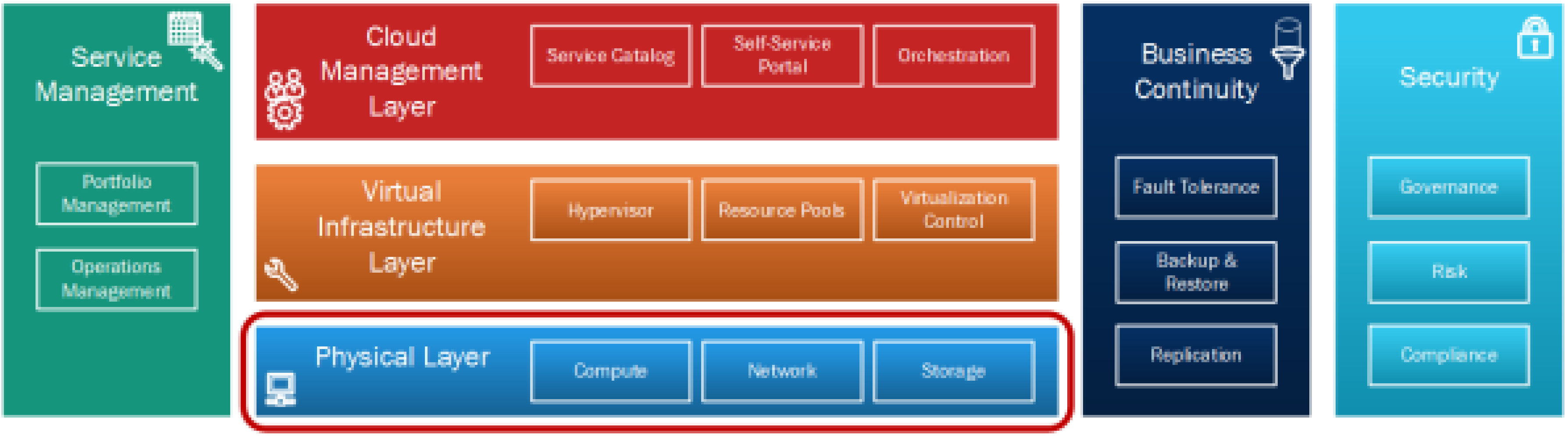
Steven W Hunter, PhD, PE
IBM Fellow

May 11, 2015

Some Industry Trends

- Workloads are increasingly being scaled out with growing dependence on the network and its characteristics
- Cloud services and data being distributed globally to meet performance, latency, and high availability requirements
- Significant scaling of compute, network, and data resources
- Increasing number and lengths of fiber as data centers grow in size
- Software defined technologies being deployed to improve management and reduce costs

Software Defined Data Center (SDDC)



- Data Center: “A large group of networked computer servers typically used by organizations for the remote storage, processing, or distribution of large amounts of data.”
- The goal of the SDDC platform is to provide efficient and flexible access to the underlying infrastructure
- Service Management, Business Continuity and Security are also part of goal
- Multiple activities are underway toward the SDDC vision with OpenStack being a key component using open technology

OpenStack

- An open environment for deploying Infrastructure Services

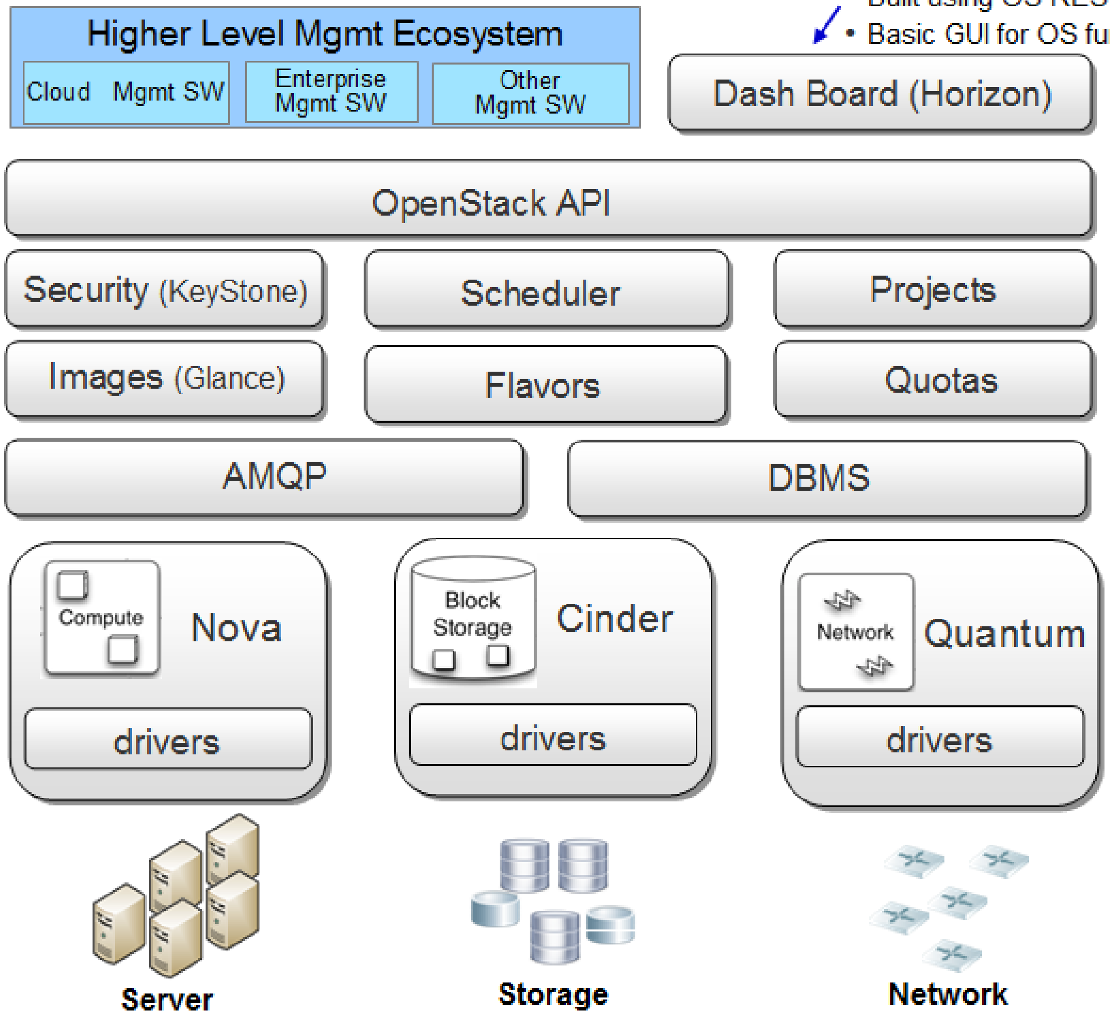
Cloud Management APIs
- Focus on providing IaaS
- Broad Eco System

Management Services
- Image Management
- Virtual Machine Placement
- Account Management

Foundation (Middleware)
- AMQP Message Broker
- Database for Persistence

Virtualization Drivers
- Adapters to hypervisors
- Server, storage, network
- Vendor Led Drivers

Simple Console
- Built using OS REST API
- Basic GUI for OS functions



Higher Level Mgmt Ecosystem

Cloud Mgmt SW | Enterprise Mgmt SW | Other Mgmt SW

Dash Board (Horizon)

OpenStack API

Security (KeyStone) | Scheduler | Projects

Images (Glance) | Flavors | Quotas

AMQP | DBMS

Nova: Compute, drivers

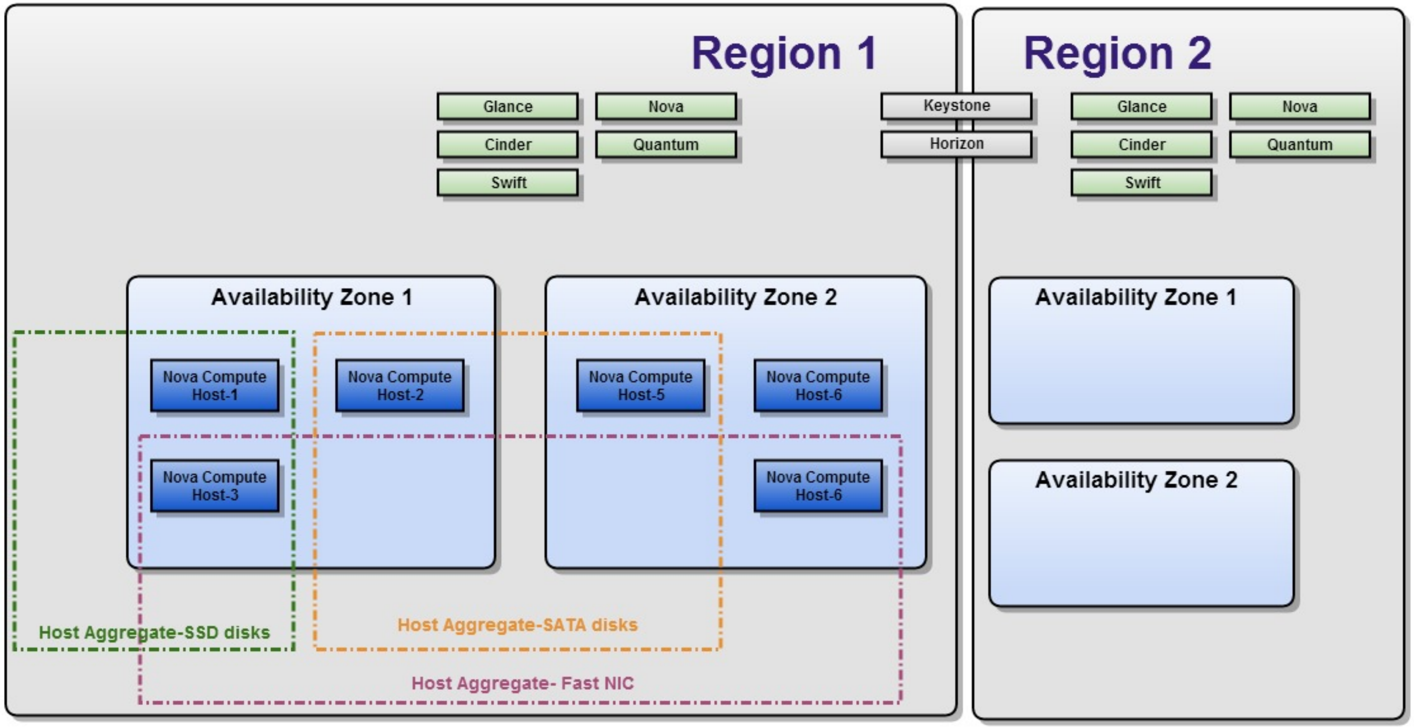
Cinder: Block Storage, drivers

Quantum: Network, drivers

Server | Storage | Network

openstack CLOUD SOFTWARE

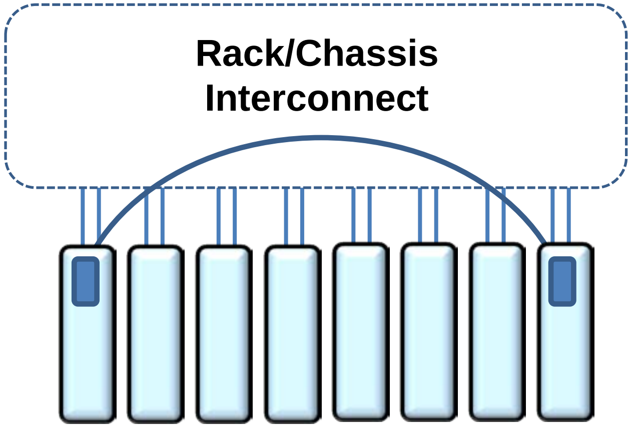
Availability Zones/Regions with OpenStack



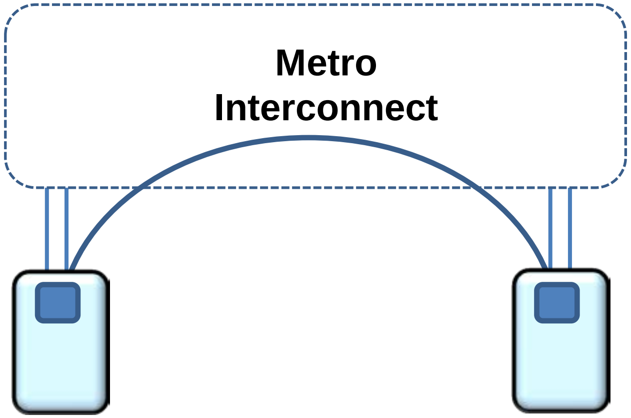
- Availability Zones/Regions are composed of “host aggregates” and are defined in to achieve high availability, disaster recovery, and performance
- A host aggregate is a grouping of hosts with associated metadata
- The option exists for a host to be in more than one host aggregate
- A host aggregate <u>may</u> be exposed to users in the form of an availability zone
- An availability zone name is an option when creating a host aggregate

System Examples

Local



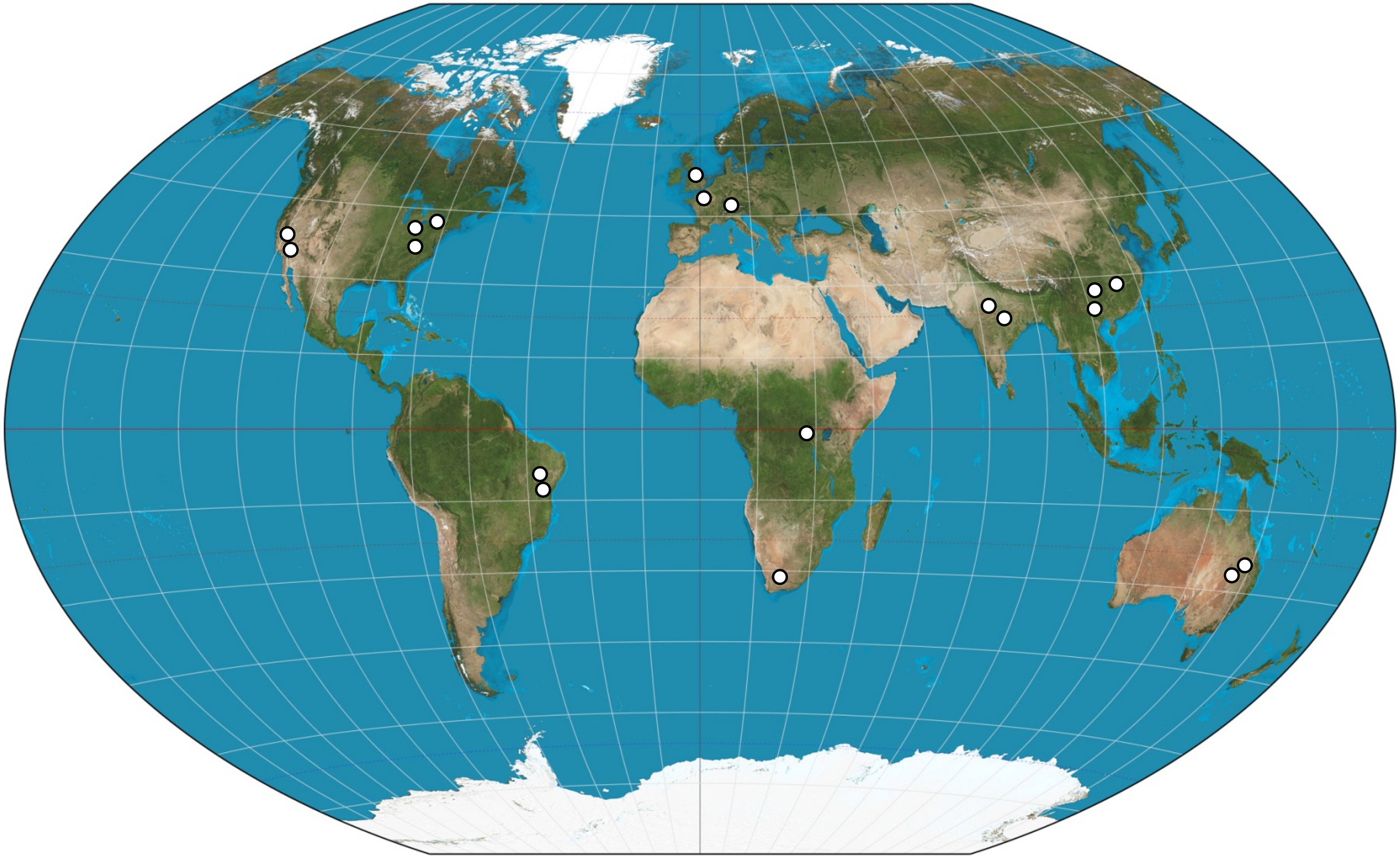
Distributed



- System models provide for scalability, performance, resource sharing, high availability, disaster recovery and/or other solutions
- Enabled by low latency, increasing bandwidth, and virtualization
- SDN, RDMA and Optical technology improving data movement efficiency
- Commercial optical interconnects are impacting today’s computing systems

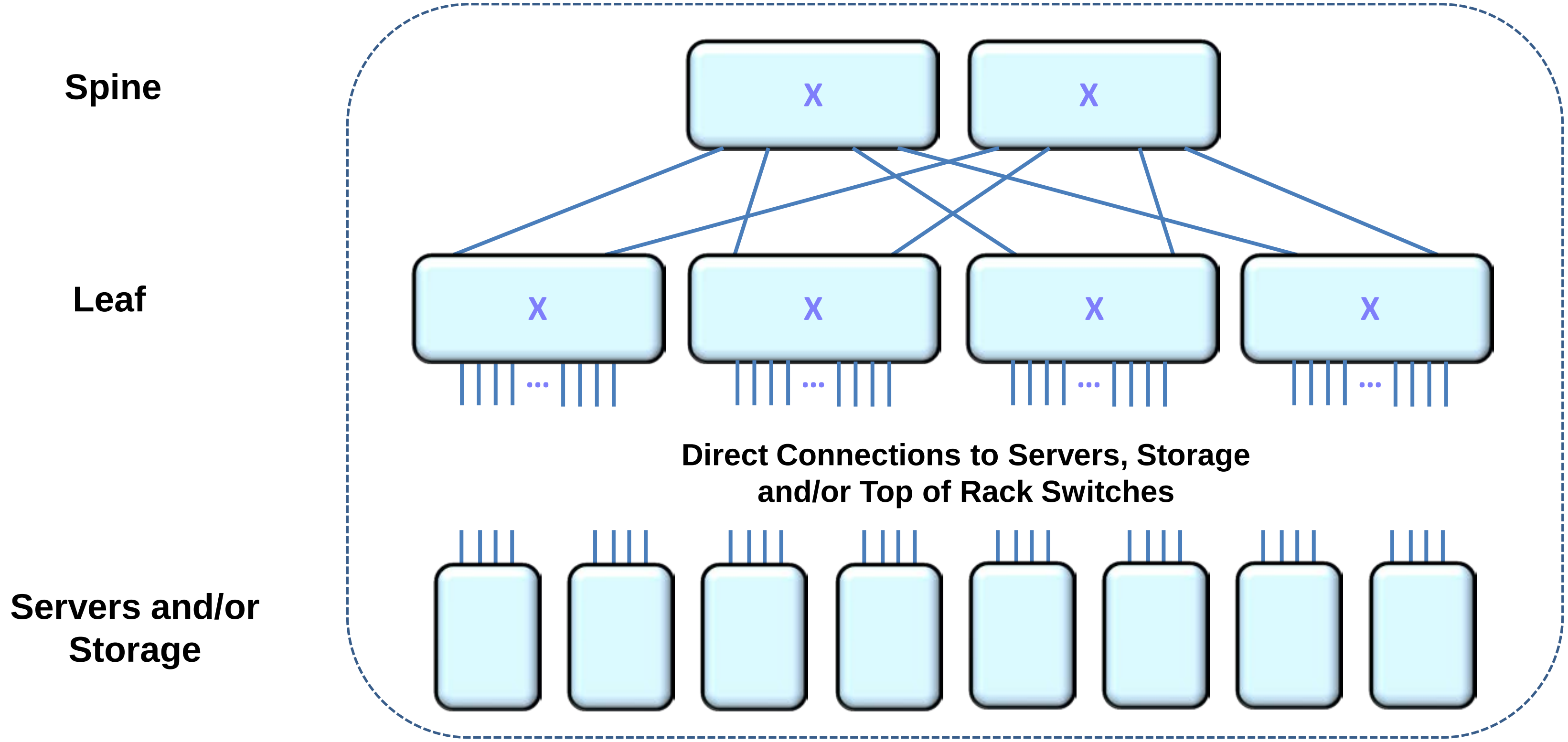
Machine	Name	# Fibers
BG/Q	Sequoia	660k
BG/Q	Mira	330k
BG/Q	JUQUEEN	165k
BG/Q	Fermi	65k
Power 775	DARPA	600k

Global Application of Availability Zones/Regions (Example)



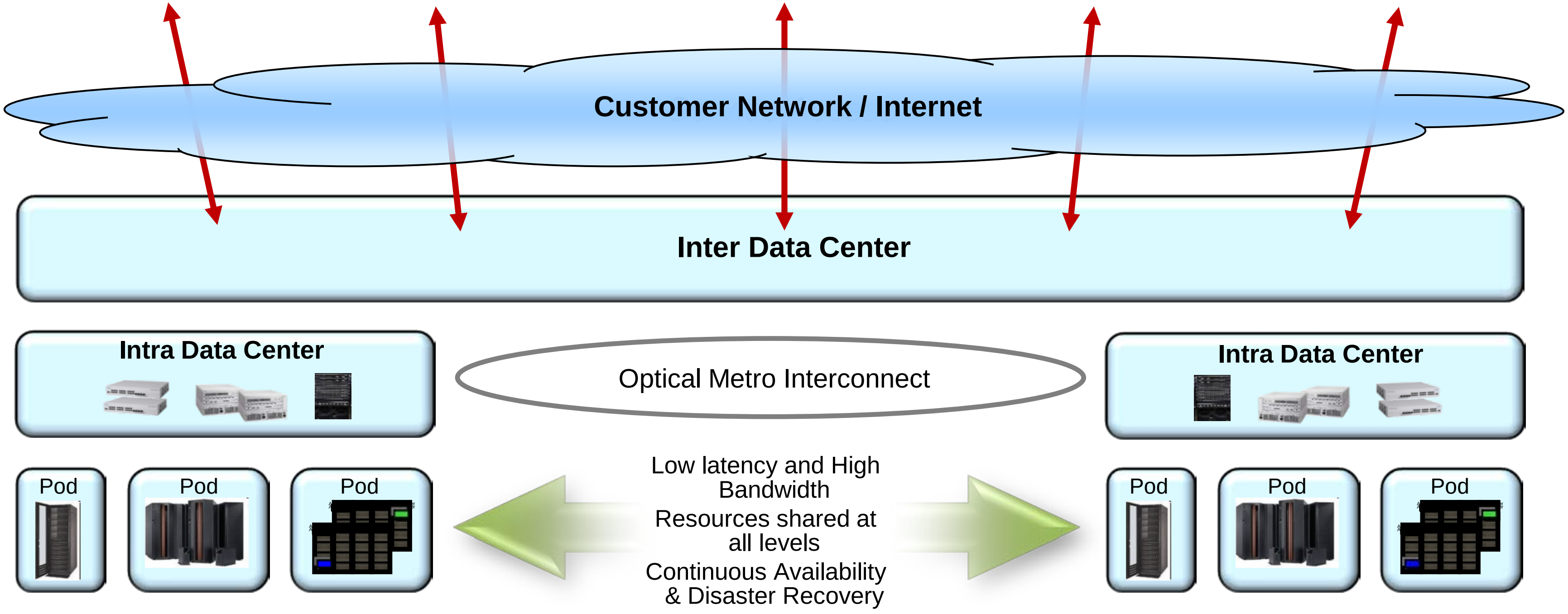
- Localized disasters and data locality are two factors driving global scale
- Latency (speed of light) becomes more of a significant issue globally

Data Center Infrastructure Example with Spine/Leaf Network


Spine
X
X
Leaf
X
X
X
X
...
...
...
...
Direct Connections to Servers, Storage and/or Top of Rack Switches
Servers and/or Storage

- Designed for scalability, high availability, and balance of resources
- Number of Spine and Leaf nodes may vary depending on size of deployment
- Ethernet is the most common interconnect with a variety of speeds supported (e.g., 1G, 10G, 25G, 40G, 50G, and 100G)

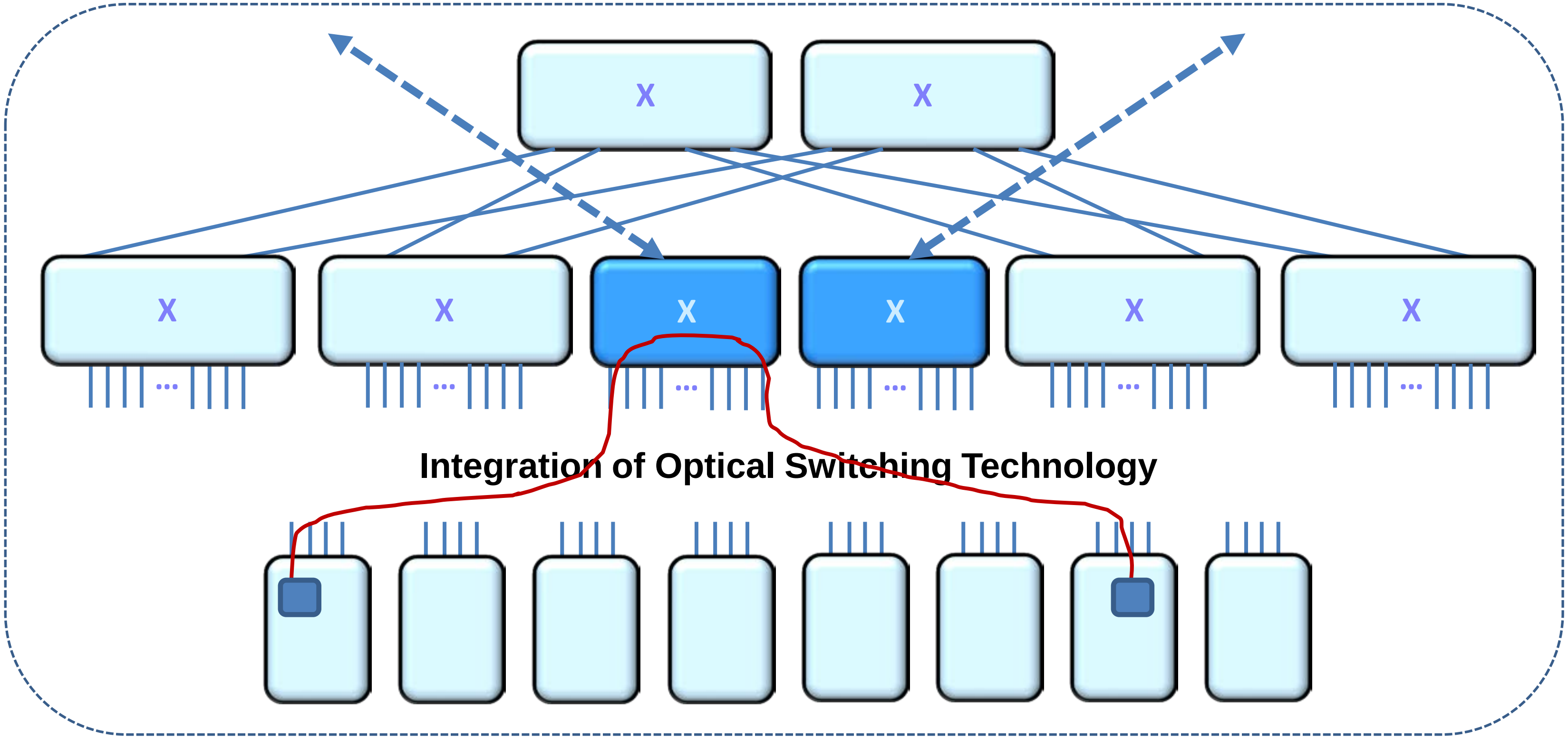
Flexible Distributed System Framework / Architecture


Customer Network / Internet
Inter Data Center
Intra Data Center
Optical Metro Interconnect
Intra Data Center
Pod
Pod
Pod
Low latency and High Bandwidth
Resources shared at all levels
Continuous Availability & Disaster Recovery
Pod
Pod
Pod

Goals:

- Provide a broader set of options/services for scalability, performance, high availability, and capacity expansion at multiple levels
- Consistent set of services providing 'network containment' of applications well bounded latency, scaling independence, and better established fault domains

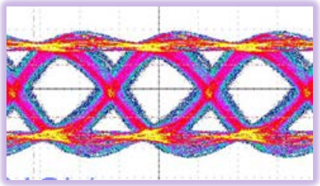
Hybrid Architecture Example



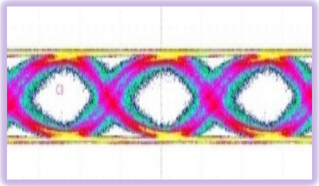
- Not a new idea. Multiple research activities have been underway
- Integrate optical switching as a bypass or fast path to provide additional options and/or services
- Remote systems can be coupled dynamically with a guaranteed network service

Some Advantages of Optically vs Electrically Switched Links

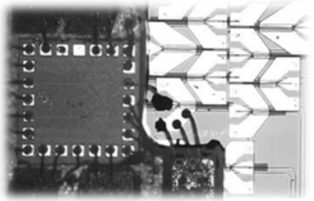
- Interconnect speed, power, and density



64G VCSEL link

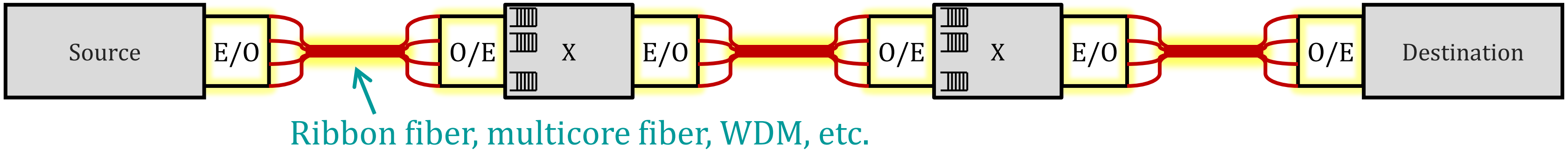


1 pJ/b VCSEL link at 25G

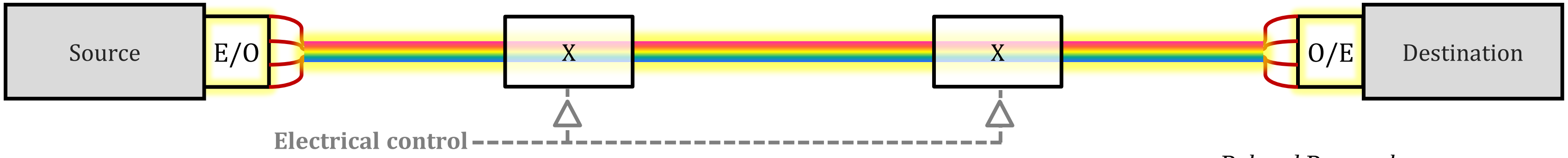


10 km at 25G SiPh link

- Lower latency and potential for power efficiency, and bandwidth density, but slower configuration and loss of granularity with optical switching
  - At each hub (electrical): de-multiplex, receive, switch, transmit, re-multiplex; Sub-streams are switched independently; Communication power scales with hops (multiple E/O/Es)

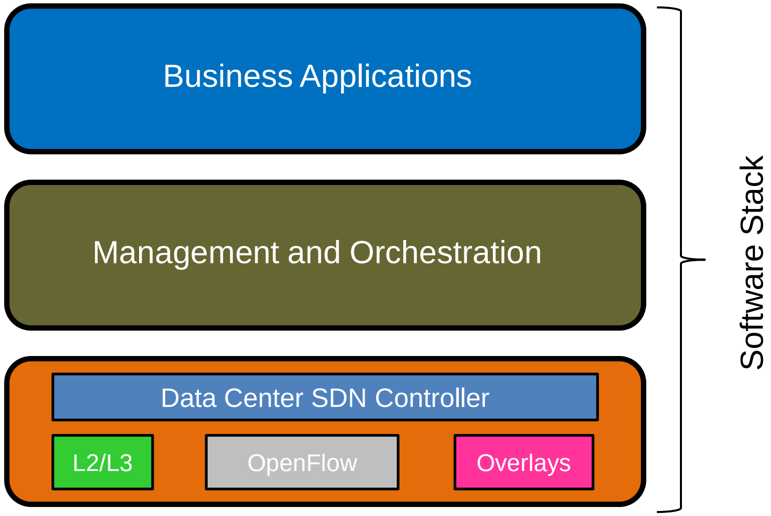


  - At each hub: route data; Sub-streams (e.g. WDM channels) routed aggregately; Link power can scale sub-linearly with hop count; Electrical low speed control only



Related Research
Schares, IBM Corporation

Hybrid Architecture Example



Business Applications

Management and Orchestration

Data Center SDN Controller

L2/L3

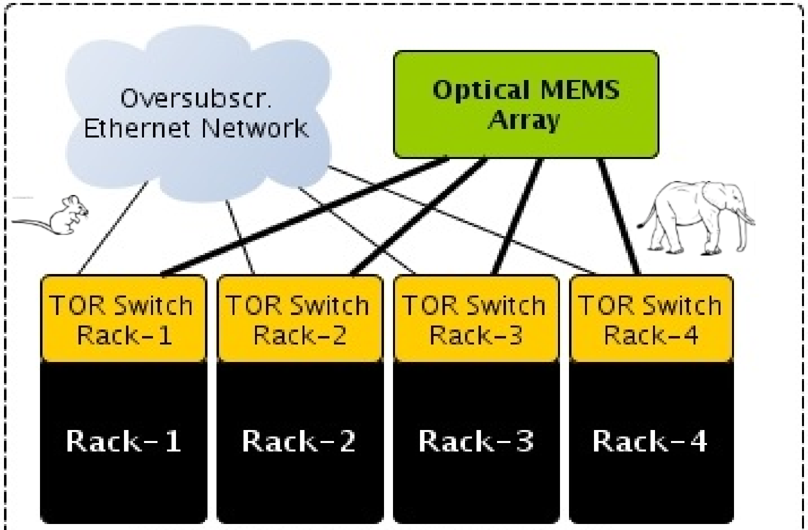
OpenFlow

Overlays

Software Stack

Hybrid Networks Value Add:

1. Dynamic bandwidth allocation
2. Provides interconnect with low latency and consistency
3. High Bandwidth, non-blocking, line rate
4. Provide control with software APIs
5. Potentially lower Capex and Opex(*)



Hybrid Network Architecture

Needs

1. Lower costs at the physical layer
2. Improve monitoring (e.g., flow characteristics) and extend control plane (**)

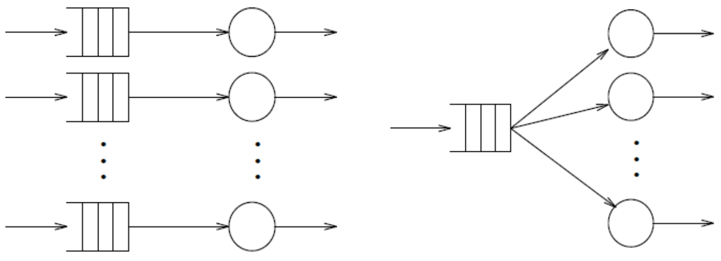
Related Research

(*) K. Katrinis et al., Euro-Par 2013.

(**) K. Katrinis et al., IEEE Summer Topicals 2013
Farrington, et.al. – Helios, SIGCOMM‘10

Modeling and Analysis Examples

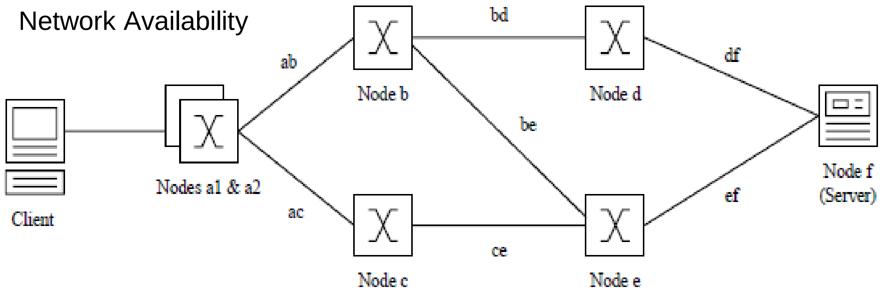
1. Availability Modeling of a Two Node Cluster
2. Analysis of software rejuvenation in cluster systems using stochastic reward nets
3. Combined Performance and Availability (Performability) Analysis of a Switched Network Application
4. Reliability Analysis Techniques Explored Through a Communication Network Example

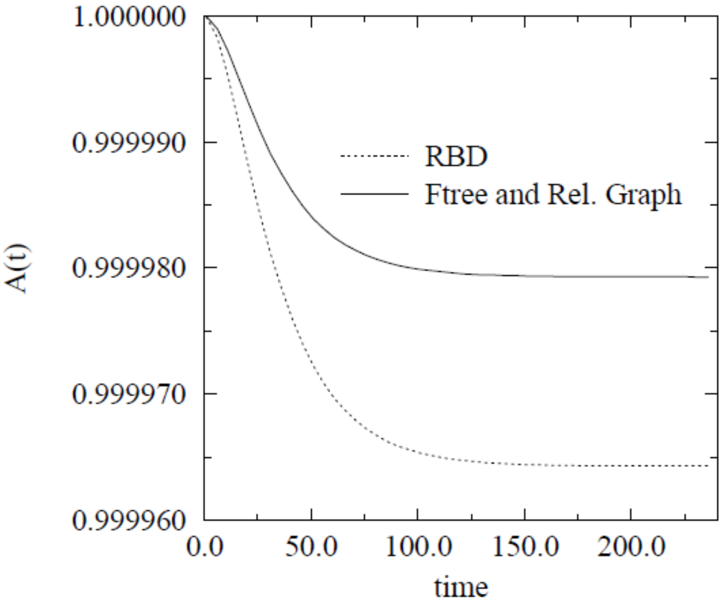


Multi-Server Cluster with Separate Queues

Multi-Server Cluster with a Common Queue

Network Availability





Instantaneous Availability of Network

Summary

- A hybrid architecture better enables a dynamic bypass / fast path
- This approach aligns well with Software Defined Technology
- Enables the extension of network service options for High Availability, Performance, and Capacity Expansion
- Modeling and analysis can provide good insight for optimizing services
- Broader traffic flow information would enhance applications/services