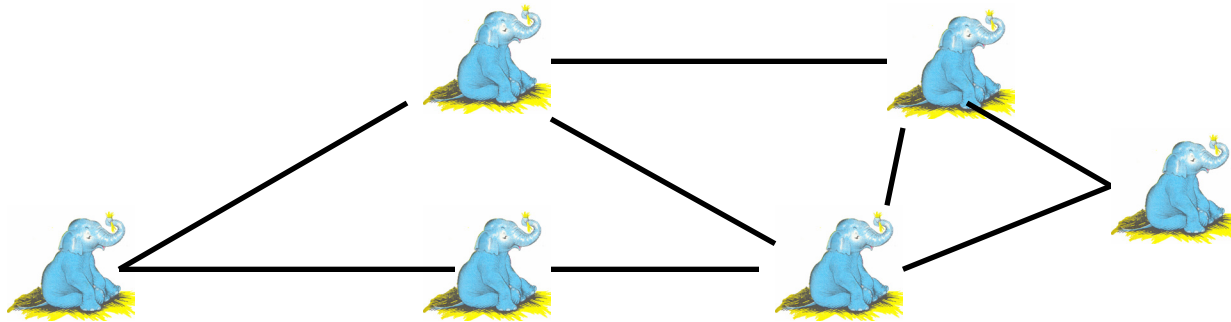


Networking Issues For Big Data



Raj Jain

Washington University in Saint Louis

Saint Louis, MO 63130

Jain@cse.wustl.edu

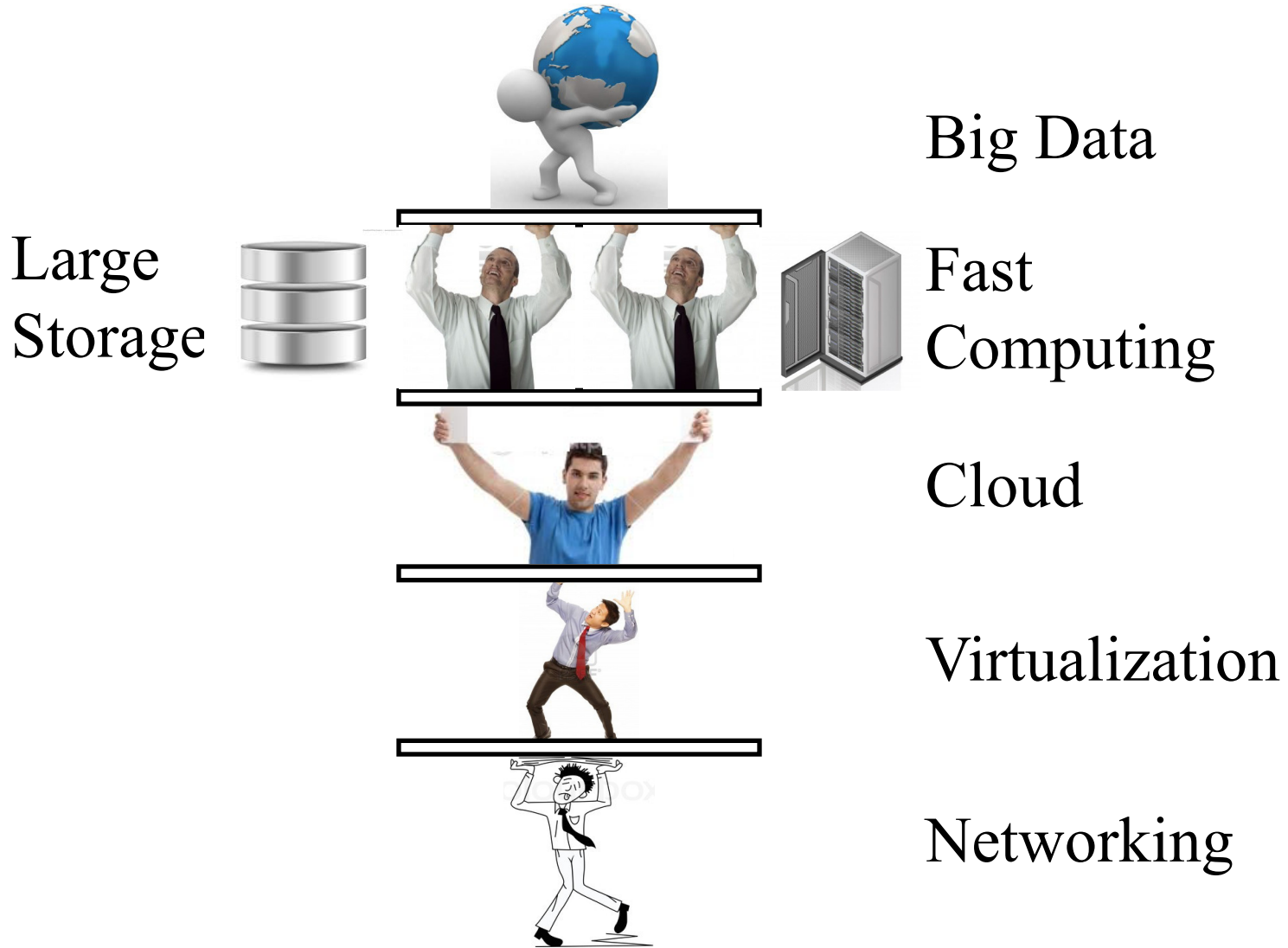
These slides and audio/video recordings of this class lecture are at:

<http://www.cse.wustl.edu/~jain/cse570-18/>



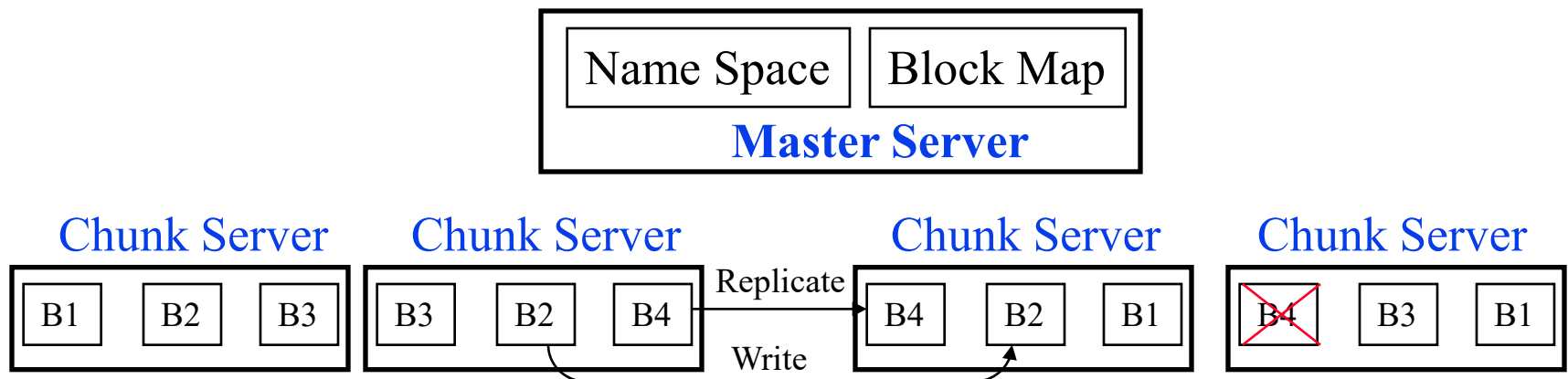
1. Why, What, and How of Big Data:
It's all because of advances in networking
2. Recent Developments in Networking and their role in Big Data (Virtualization, SDN, NFV)
3. Networking needs Big Data

Big Data Enabled by Networking



Google File System

- ❑ Commodity computers serve as “Chunk Servers” and store multiple copies of data blocks
- ❑ A master server keeps a map of all chunks of files and location of those chunks.
- ❑ All writes are propagated by the writing chunk server to other chunk servers that have copies.
- ❑ Master server controls all read-write accesses



Ref: S. Ghemawat, et al., "The Google File System", OSP 2003, <http://research.google.com/archive/gfs.html>

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-18/>

©2018 Raj Jain

BigTable

- ❑ Distributed storage system built on Google File System
- ❑ Data stored in rows and columns
- ❑ Optimized for sparse, persistent, multidimensional sorted map.
- ❑ Uses commodity servers
- ❑ Not distributed outside of Google but accessible via Google App Engine

Ref: F. Chang, et al., "Bigtable: A Distributed Storage System for Structured Data," 2006,

<http://research.google.com/archive/bigtable.html>

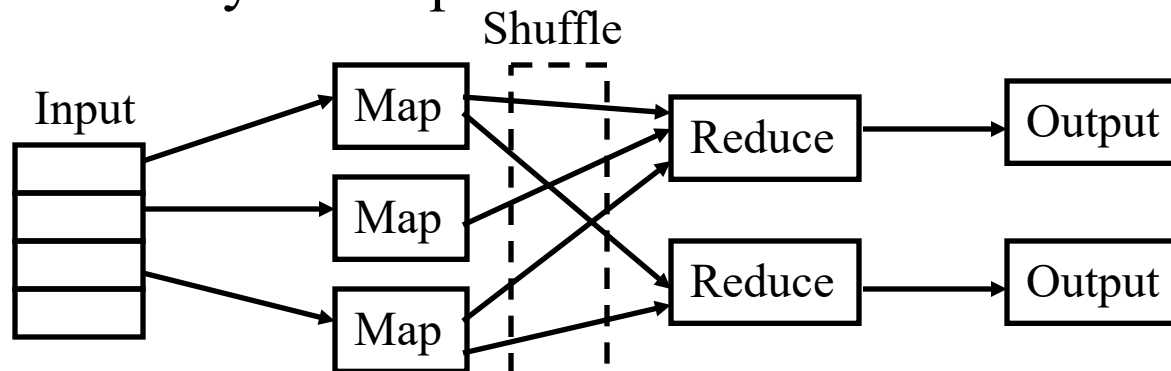
Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-18/>

©2018 Raj Jain

MapReduce

- ❑ Software framework to process massive amounts of unstructured data in parallel
- ❑ **Goals:**
 - **Distributed:** over a large number of inexpensive processors
 - **Scalable:** expand or contract as needed
 - **Fault tolerant:** Continue in spite of some failures
- ❑ **Map:** Takes a set of data and converts it into another set of key-value pairs..
- ❑ **Reduce:** Takes the output from Map as input and outputs a smaller set of key-value pairs.



Ref: J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," OSDI 2004,
<http://research.google.com/archive/mapreduce-osdi04.pdf>

MapReduce Example

- ❑ 100 files with daily temperature in two cities. Each file has 10,000 entries.
- ❑ For example, one file may have (Toronto 20), (New York 30), ..
- ❑ Our goal is to compute the maximum temperature in the two cities.
- ❑ Assign the task to 100 Map processors each works on one file. Each processor outputs a list of key-value pairs, e.g., (Toronto 30), New York (65), ...
- ❑ Now we have 100 lists each with two elements. We give this list to two reducers – one for Toronto and another for New York.
- ❑ The reducer produce the final answer: (Toronto 55), (New York 65)

MapReduce Optimization

❑ **Scheduling:**

- Task is broken into pieces that can be computed in parallel
- Map tasks are scheduled before the reduce tasks.
- If there are more map tasks than processors, map tasks continue until all of them are complete.
- A new strategy is used to assign Reduce jobs so that it can be done in parallel
- The results are combined.

❑ **Synchronization:** The map jobs should be comparables so that they finish together. Similarly reduce jobs should be comparable.

❑ **Code/Data Collocation:** The data for map jobs should be at the processors that are going to map.

❑ **Fault/Error Handling:** If a processor fails, its task needs to be assigned to another processor.

Story of Hadoop

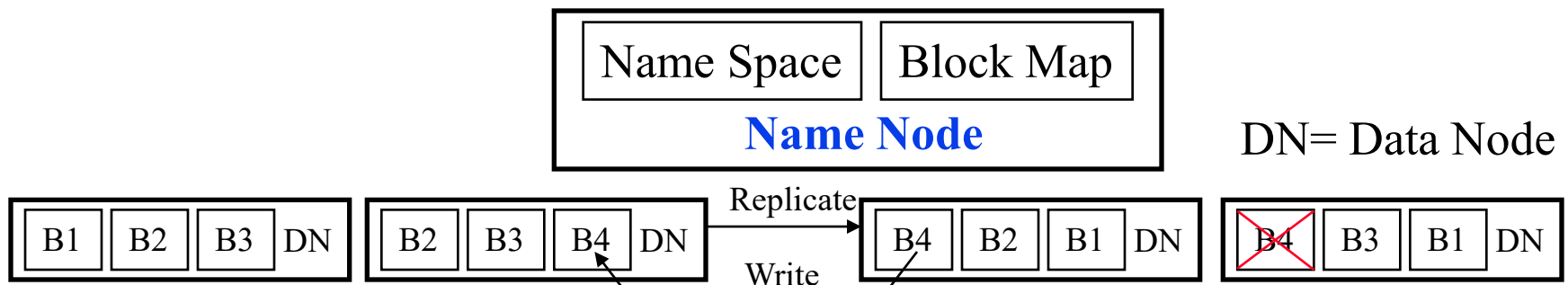
- ❑ Doug Cutting at Yahoo and Mike Caferella were working on creating a project called “Nutch” for large web index.
- ❑ They saw Google papers on MapReduce and Google File System and used it
- ❑ Hadoop was the name of a yellow plus elephant toy that Doug’s son had.
- ❑ In 2008 Amr left Yahoo to found Cloudera.
In 2009 Doug joined Cloudera.

Ref: Michael Minelli, "Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses," Wiley, 2013, ISBN:'111814760X



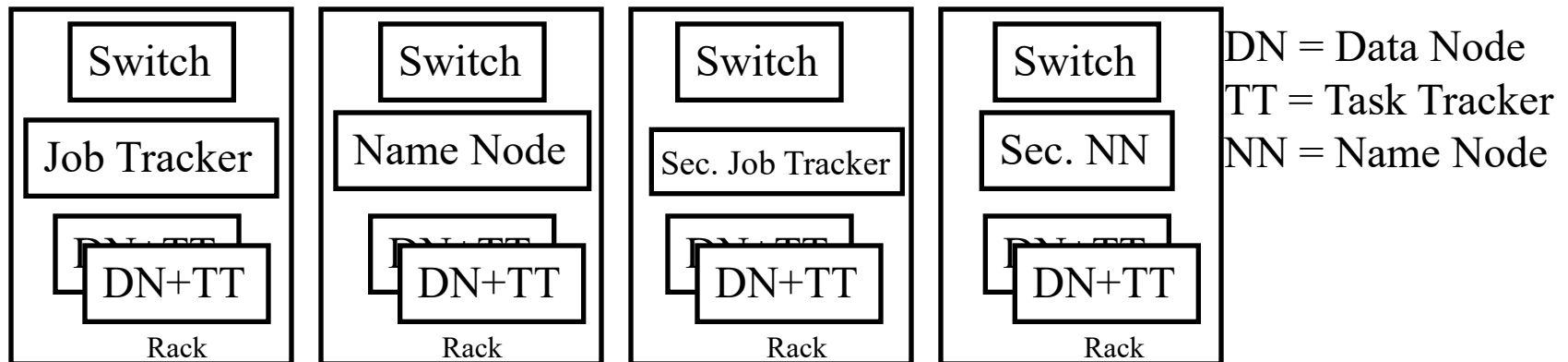
Hadoop

- ❑ An open source implementation of MapReduce
- ❑ Named by Doug Cutting at Yahoo after his son's yellow plus elephant
- ❑ Hadoop File System (**HDFS**) requires data to be broken into blocks. Each block is stored on 2 or more data nodes on different racks.
- ❑ **Name node**: Manages the file system name space
⇒ keeps track of blocks on various **Data Nodes**.



Hadoop (Cont)

- ❑ **Job Tracker:** Assigns MapReduce jobs to task tracker nodes that are **close** to the data (same rack)
- ❑ **Task Tracker:** Keep the work as close to the data as possible.



Hadoop (Cont)

- ❑ Data nodes get the data if necessary, do the map function, and write the results to disks.
- ❑ Job tracker then assigns the reduce jobs to data nodes that have the map output or close to it.
- ❑ All data has a check attached to it to verify its integrity.

Networking Requirements for Big Data

1. **Code/Data Collocation:** The data for map jobs should be at the processors that are going to map.
2. **Elastic bandwidth:** to match the variability of volume
3. **Fault/Error Handling:** If a processor fails, its task needs to be assigned to another processor.
4. **Security:** Access control (authorized users only), privacy (encryption), threat detection, all in real-time in a highly scalable manner
5. **Synchronization:** The map jobs should be comparables so that they finish together. Similarly reduce jobs should be comparable.

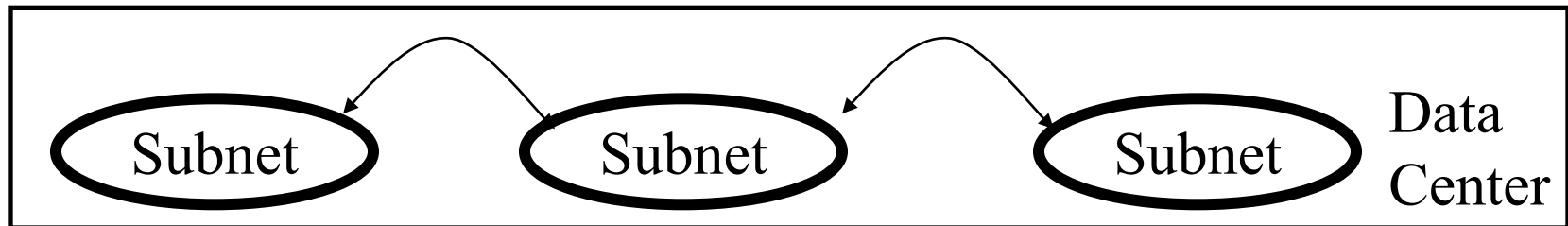
Recent Developments in Networking

1. High-Speed: 100 Gbps Ethernet
⇒ 400 Gbps ⇒ 1000 Gbps
⇒ Cheap storage access. Easy to move big data.
2. Virtualization
3. Software Defined Networking
4. Network Function Virtualization

Virtualization (Cont)

- Recent networking technologies and standards allow:
 1. Virtualizing Computation
 2. Virtualizing Storage
 3. Virtualizing Rack Storage Connectivity
 4. Virtualizing Data Center Storage
 5. Virtualizing Metro and Global Storage

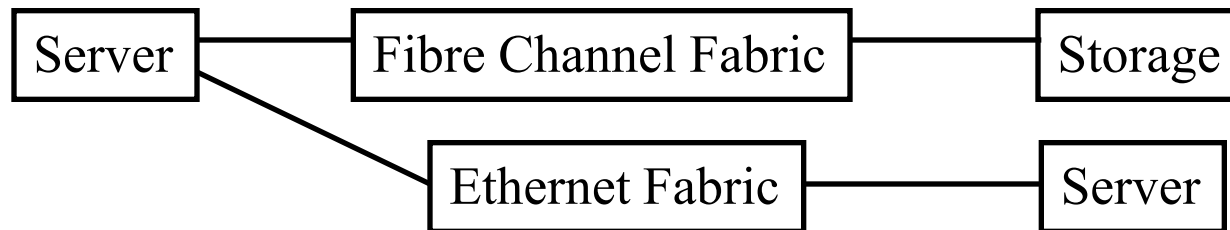
1. Virtualizing Computation



- ❑ Initially data centers consisted of multiple IP subnets
 - Each subnet = One Ethernet Network
 - Ethernet addresses are globally unique and do not change
 - IP addresses are locators and change every time you move
 - If a VM moves inside a subnet \Rightarrow No change to IP address \Rightarrow Fast
 - If a VM moves from one subnet to another \Rightarrow Its IP address changes \Rightarrow All connections break \Rightarrow Slow \Rightarrow Limited VM mobility
- ❑ IEEE 802.1ad-2005 Ethernet Provider Bridging (PB), IEEE 802.1ah-2008 Provider Backbone Bridging (PBB) allow Ethernets to span long distances \Rightarrow Global VM mobility

2. Virtualizing Storage

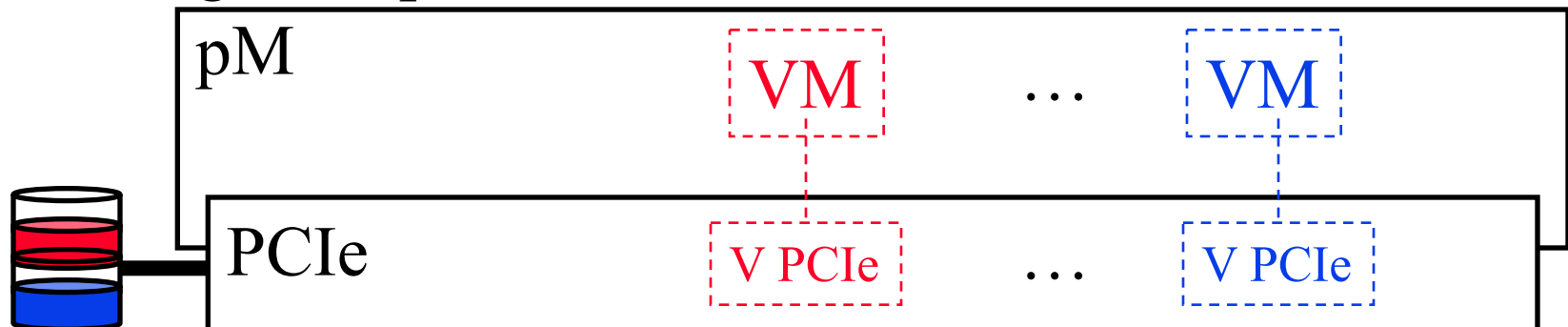
- Initially data centers used Storage Area Networks (Fibre Channel) for server-to-storage communications and Ethernet for server-to-server communication



- IEEE added 4 new standards to make Ethernet offer low loss, low latency service like Fibre Channel:
 - Priority-based Flow Control (IEEE 802.1Qbb-2011)
 - Enhanced Transmission Selection (IEEE 802.1Qaz-2011)
 - Congestion Control (IEEE 802.1Qau-2010)
 - Data Center Bridging Exchange (IEEE 802.1Qaz-2011)
- Result: Unified networking \Rightarrow Significant CapEx/OpEx saving

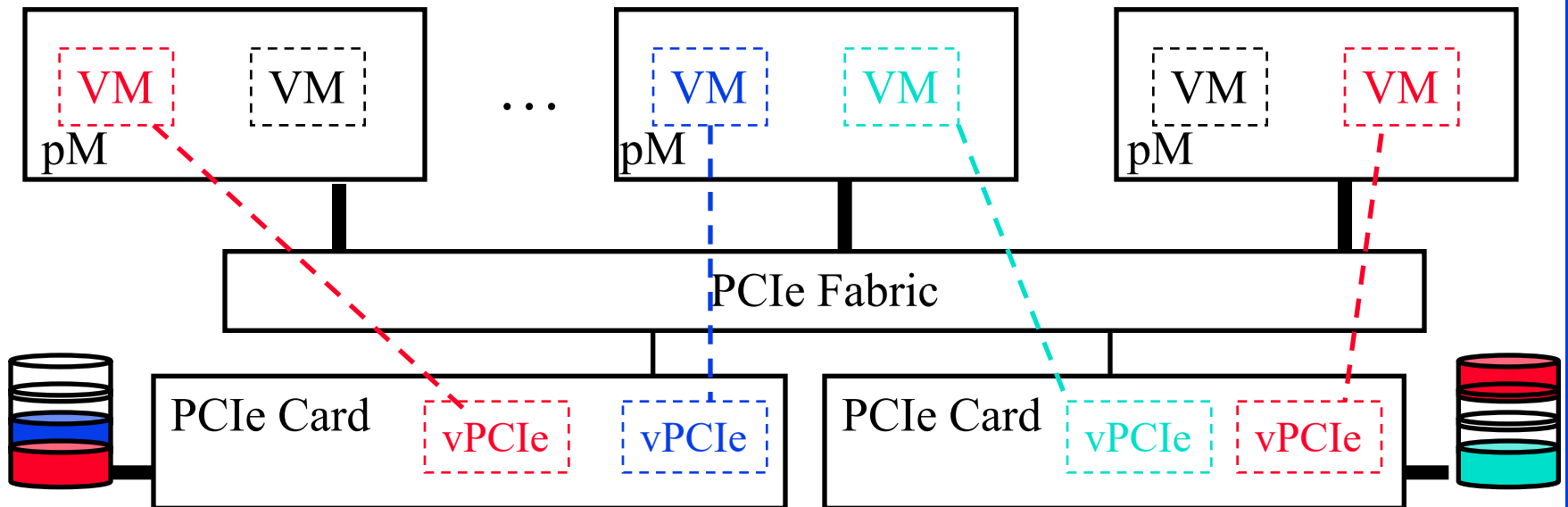
3. Virtualizing Rack Storage Connectivity

- ❑ MapReduce jobs are assigned to the nodes that have the data
- ❑ Job tracker assigns jobs to task trackers in the **rack** where the data is.
- ❑ High-speed Ethernet can get the data in the same rack.
- ❑ Peripheral Connect Interface (PCI) Special Interest Group (SIG)'s Single Root I/O virtualization (**SR-IOV**) allows a storage to be virtualized and shared among multiple VMs.



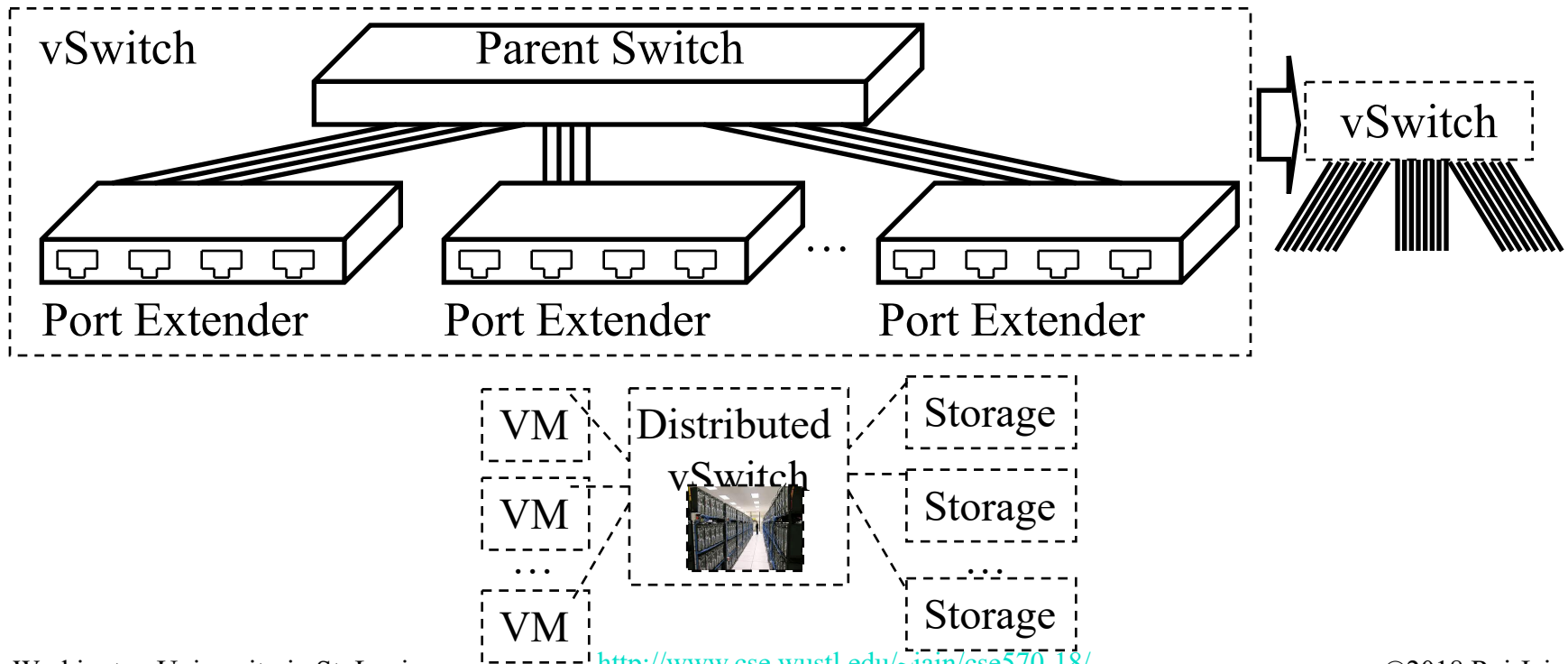
Multi-Root IOV

- ❑ PCI-SIG Multi-Root I/O Virtualization (**MR-IOV**) standard allows one or more PCIe cards to serve multiple servers and VMs in the same rack
- ❑ Fewer adapters \Rightarrow Less cooling. No adapters \Rightarrow Thinner servers



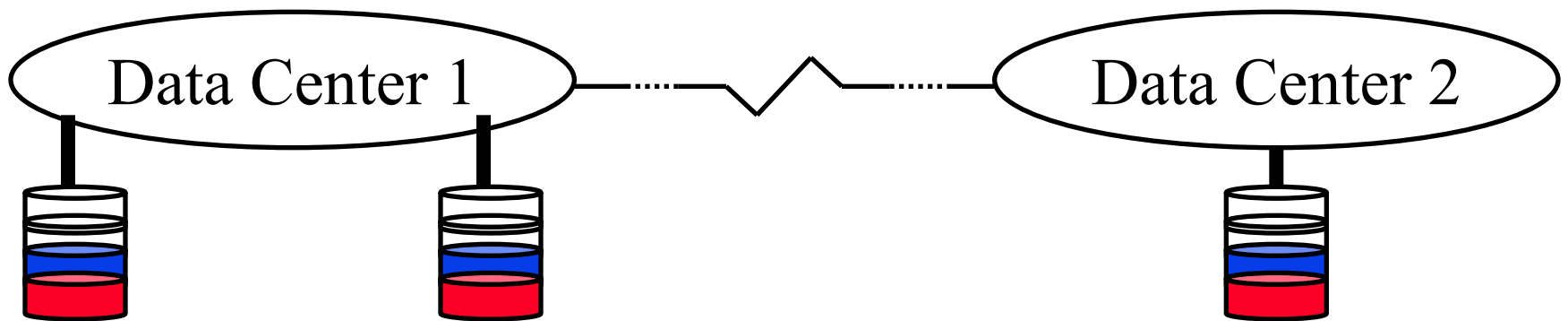
4. Virtualizing Data Center Storage

- ❑ IEEE 802.1BR-2012 Virtual Bridgeport Extension (VBE) allows multiple switches to combine in to a very large switch
- ❑ Storage and computers located anywhere in the data center appear as if connected to the same switch



5. Virtualizing Metro Storage

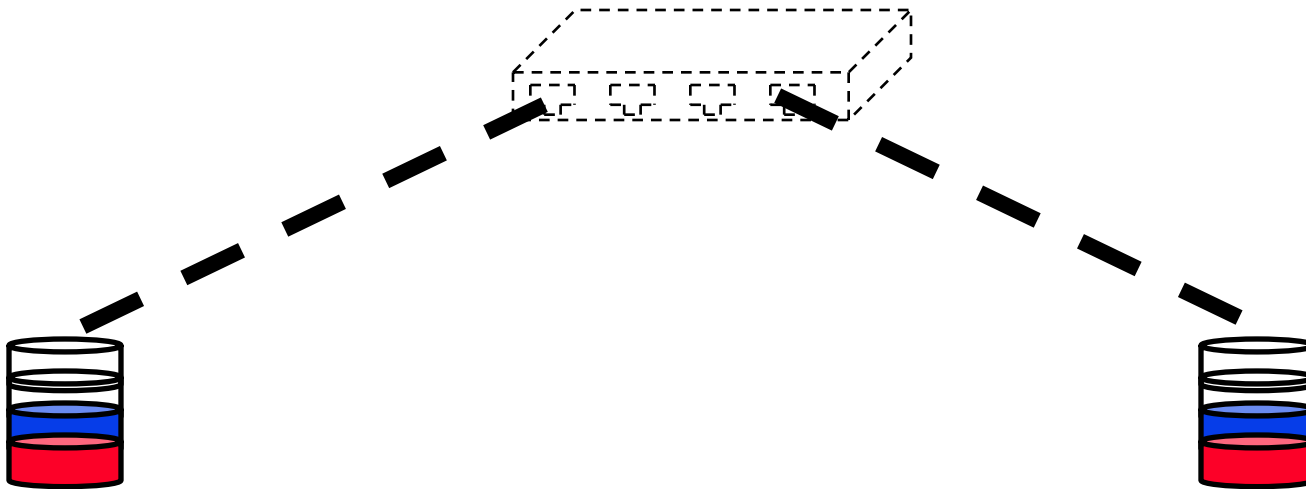
- ❑ Data center Interconnection standards:
 - Virtual Extensible LAN (VXLAN),
 - Network Virtualization using GRE (NVGRE), and
 - Transparent Interconnection of Lots of Link (TRILL)
- ⇒ data centers located far away to appear to be on the same Ethernet



Ref: <http://tools.ietf.org/html/draft-mahalingam-dutt-dcops-vxlan-04>, <http://tools.ietf.org/html/draft-sridharan-virtualization-nvgre-03>, RFC 5556

Virtualizing the Global Storage

- ❑ Energy Science Network (ESNet) uses virtual switch to connect members located all over the world
- ❑ Virtualization \Rightarrow Fluid networks \Rightarrow The world is flat \Rightarrow You draw your network \Rightarrow Every thing is virtually local



Ref: I. Monga, "Software Defined Networking for Big-data Science,"

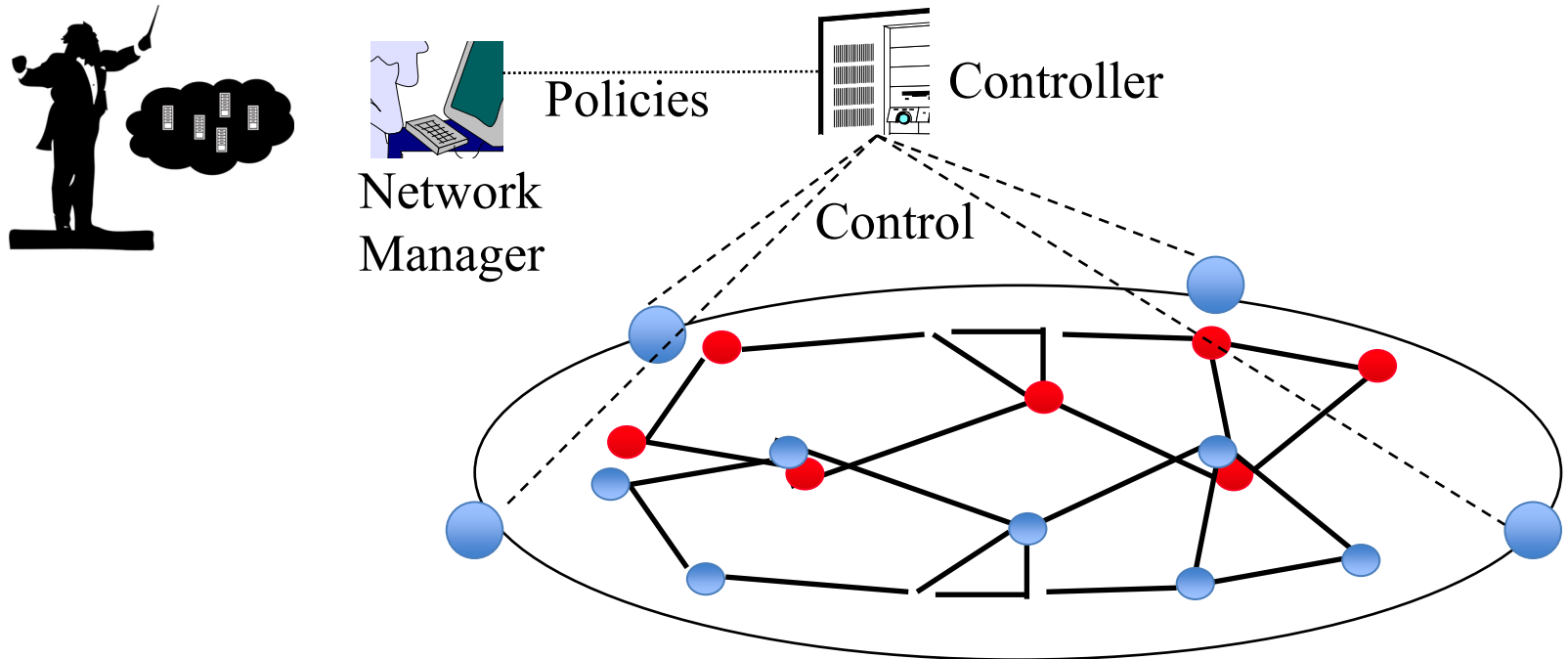
http://www.es.net/assets/pubs_presos/Monga-WAN-Switch-SC12SRS.pdf

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-18/>

©2018 Raj Jain

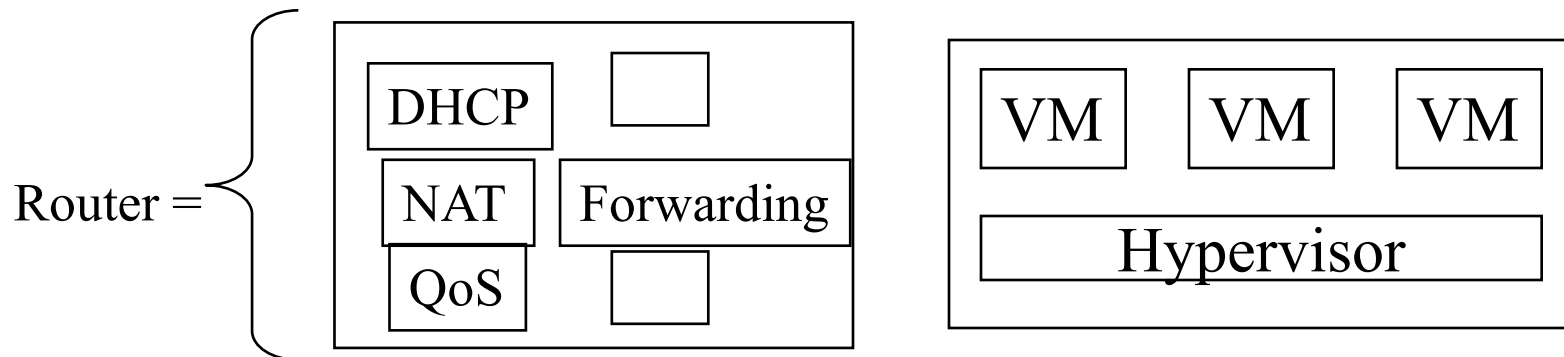
Software Defined Networking



- ❑ Centralized Programmable Control Plane
- ❑ Allows automated orchestration (provisioning) of a large number of virtual resources (machines, networks, storage)
- ❑ Large Hadoop topologies can be created on demand

Network Function Virtualization (NFV)

- ❑ Fast standard hardware \Rightarrow Software based Devices
Virtual networking modules (DHCP, Firewall, DNS, ...) running on standard processors
- ❑ Modules can be combined to create any combination of function for data privacy, access control, ...
- ❑ Virtual Machine implementation \Rightarrow Quick provisioning
- ❑ Standard Application Programming Interfaces (APIs)
 \Rightarrow Networking App Market
 \Rightarrow Privacy and Security for Big data in the multi-tenant clouds



Big Data for Networking

- Today's data center:
 - Tens of tenants
 - Hundreds of switches and routers
 - Thousands of servers
 - Hundreds of administrators
- Tomorrow:
 - 1k of clients
 - 10k of pSwitches
⇒ 100k of vSwitches
 - 1M of VMs
 - Tens of Administrators
- Need to monitor traffic patterns and rearrange virtual networks connecting millions of VMs in real-time
⇒ Managing clouds is a real-time big data problem.
- Internet of things ⇒ Big Data generation and analytics



Summary

1. I/O virtualization allows all storage in the rack to appear local to any VM in that rack \Rightarrow Solves the co-location problem of MapReduce
2. Network virtualization allows storage anywhere in the data center or even other data centers to appear local
3. Software defined networking allows orchestration of a large number of resources \Rightarrow Dynamic creation of Hadoop clusters
4. Network function virtualization will allow these clusters to have special functions and security in multi-tenant clouds.

Acronyms

- ❑ ADCOM Advanced Computing and Communications
- ❑ API Application programming interface,
- ❑ CapEx Capital Expenditure
- ❑ DARPA Defense Advanced Project Research Agency
- ❑ DHCP Dynamic Host Control Protocol
- ❑ DN Data Node
- ❑ DNS Domain Name System
- ❑ DoD Department of Defense
- ❑ DOE Department of Energy
- ❑ ESNNet Energy Science Network
- ❑ GDP Gross Domestic Production
- ❑ GRE Generic Routing Encapsulation
- ❑ HDFS Hadoop Distributed File System
- ❑ IEEE Institution of Electrical and Electronic Engineers
- ❑ IOV I/O Virtualization
- ❑ IP Internet Protocol

Acronyms (Cont)

- ❑ LAN Local Area Network
- ❑ MR-IOV Multi-root I/O Virtualization
- ❑ NAT Network Address Translation
- ❑ NFV Network Function Virtualization
- ❑ NN Name Node
- ❑ NSA National Security Agency
- ❑ OpEx Operational Expenses
- ❑ PB Provider Bridging
- ❑ PBB Provider Backbone Bridging
- ❑ PCI-SIG PCI Special Interest Group
- ❑ PCI Peripheral Computer Interface
- ❑ PCIe PCI Express
- ❑ pM Physical Machine
- ❑ pSwitches Physical Switch
- ❑ QoS Quality of Service
- ❑ RFC Request for Comments

Acronyms (Cont)

- ❑ SDN .Software Defined Networking
- ❑ SR-IOV Single Root I/O Virtualization
- ❑ TRILL Transparent Interconnection of Lots of Link
- ❑ TT Task Tracker
- ❑ USGS United States Geological Survey
- ❑ VBE Virtual Bridgeport Extension
- ❑ VM Virtual Machine
- ❑ vSwitch Virtual Switch
- ❑ WAN Wide-Area Network

Scan This to Download These Slides



Raj Jain

<http://rajjain.com>

Related Modules



CSE567M: Computer Systems Analysis (Spring 2013),

https://www.youtube.com/playlist?list=PLjGG94etKypJEKjNAa1n_1X0bWWNyZcof

CSE473S: Introduction to Computer Networks (Fall 2011),

https://www.youtube.com/playlist?list=PLjGG94etKypJWOSPMh8Azcg5e_10TiDw



Wireless and Mobile Networking (Spring 2016),

https://www.youtube.com/playlist?list=PLjGG94etKypKeb0nzyN9tSs_HCd5c4wXF

CSE571S: Network Security (Fall 2011),

<https://www.youtube.com/playlist?list=PLjGG94etKypKvzfVtutHcPFJXumyyg93u>



Video Podcasts of Prof. Raj Jain's Lectures,

<https://www.youtube.com/channel/UCN4-5wzNP9-ruOzQMs-8NUw>