BROCADE

DATA CENTER BANDWIDTH SCENARIOS

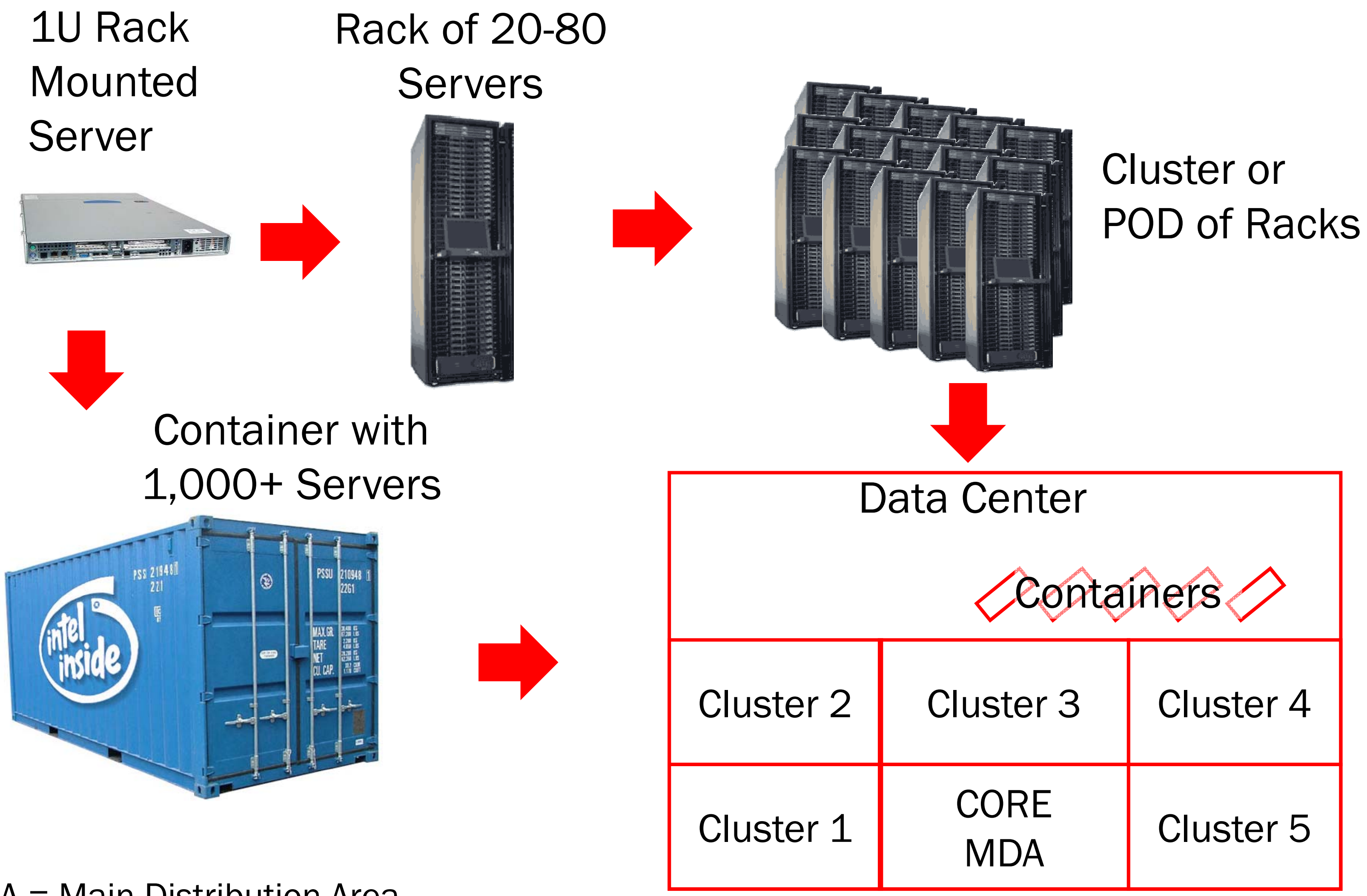
Scott Kipp – Brocade

May 18, 2011

Goal of This Presentation

- This presentation shows how a rack of servers can create hundreds of Gbps of bandwidth and can be oversubscribed at the rack
- This presentation shows how racks of servers are consolidated into pods or containers to produce Tbps of bandwidth and are oversubscribed at the pod
- This presentation shows how pods are aggregated in a core of the data center and are oversubscribed to the MAN/WAN to produce Gbps or 100s of Gbps

Data Center Summary



1U Rack Mounted Server

Rack of 20-80 Servers

Cluster or POD of Racks

Container with 1,000+ Servers

Data Center

Containers

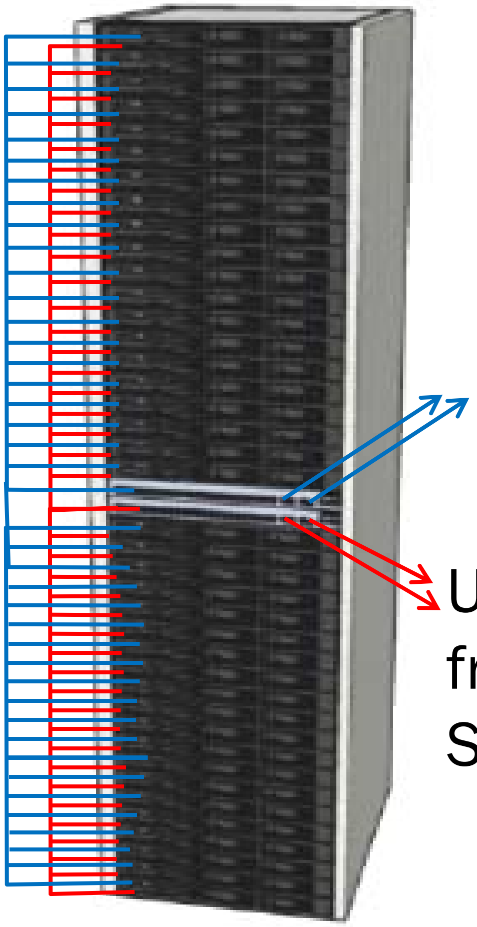
Cluster 2	Cluster 3	Cluster 4
Cluster 1	CORE MDA	Cluster 5

MDA = Main Distribution Area

Servers per Rack

Let’s study 40 Servers / Rack

10- 80 1U servers / Rack



Copper Links to Middle of Rack Switch

Uplinks from Rack Switch A

Uplinks from Rack Switch B

64 Blade servers / Rack

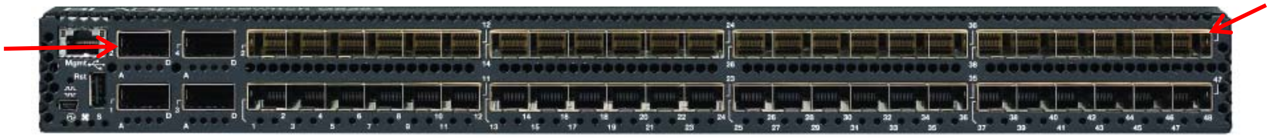
10U Blade Server chassis holds 16 Blade Servers

Bandwidth to/from Rack Switches

Let’s assume 400 Gbps / Rack in 2015

I/O per Server (Gbps)	5	10	20	40	80
Servers / Rack	40	40	40	40	40
Bandwidth / Rack (Gbps)	200	400	800	1600	3200
10GbE Uplinks with 1:1 Subscription	20	40	80	160	320
40GbE Uplinks with 1:1 Subscription	5	10	20	40	80
100GbE Uplinks with 1:1 Subscription	2	4	8	16	32
10GbE Uplinks with 4:1 Subscription	5	10	20	40	80
40GbE Uplinks with 4:1 Subscription	1.25	2.5	5	10	20
100GbE Uplinks with 4:1 Subscription	0.5	1	2	4	8

Some current ToR Switches support this



4 40GbE QSFP Ports

48 10GbE SFP+ Ports

Cluster Bandwidth Requirements

- Clusters or PODs are groups of racks of servers or a container of servers
- If each rack needs 400 Gbps, then it's pretty easy to calculate cluster bandwidth based on the number of racks.
- 100 Racks deliver 40 Tbps
  - With several switches being sold with 5Tbps of bandwidth, 8 switches would be needed in a cluster

4x4 Cluster with 16 racks

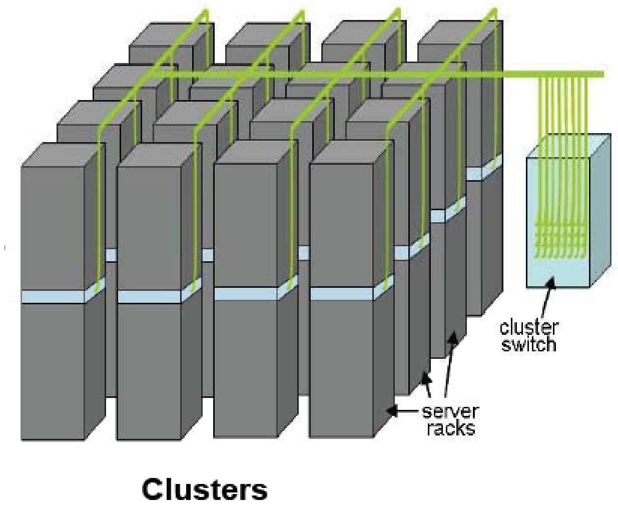
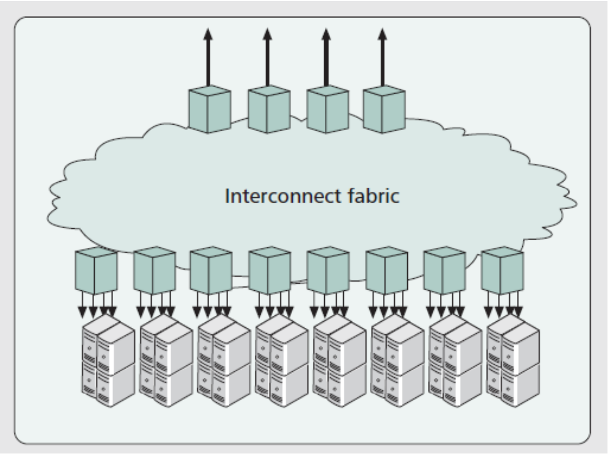


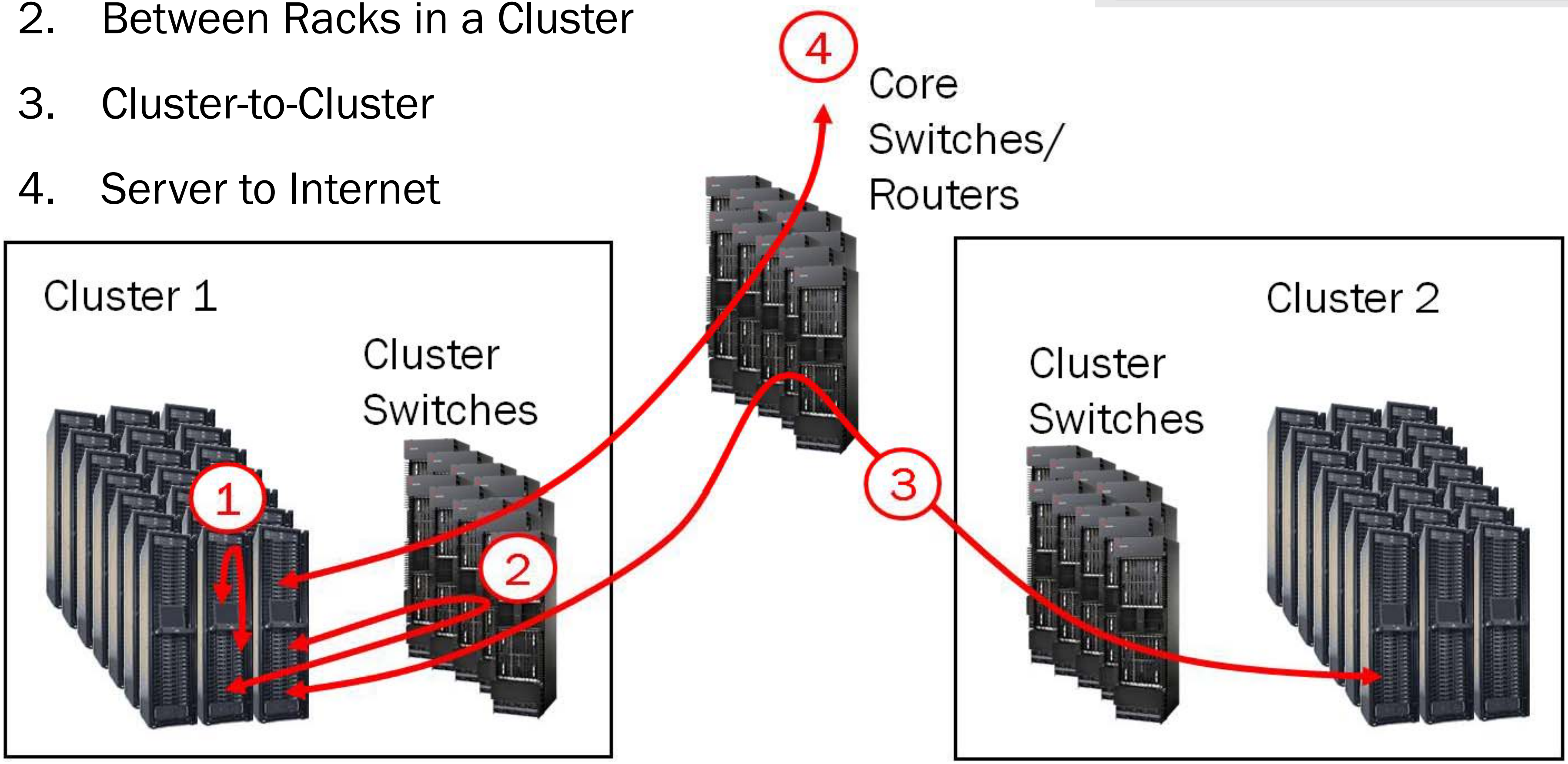
Illustration courtesy of Google

Intra/Inter-Cluster Traffic

Listed in order of over-subscription



1. Within a Rack
2. Between Racks in a Cluster
3. Cluster-to-Cluster
4. Server to Internet



Core Switches/ Routers

Cluster 1

Cluster Switches

Cluster 2

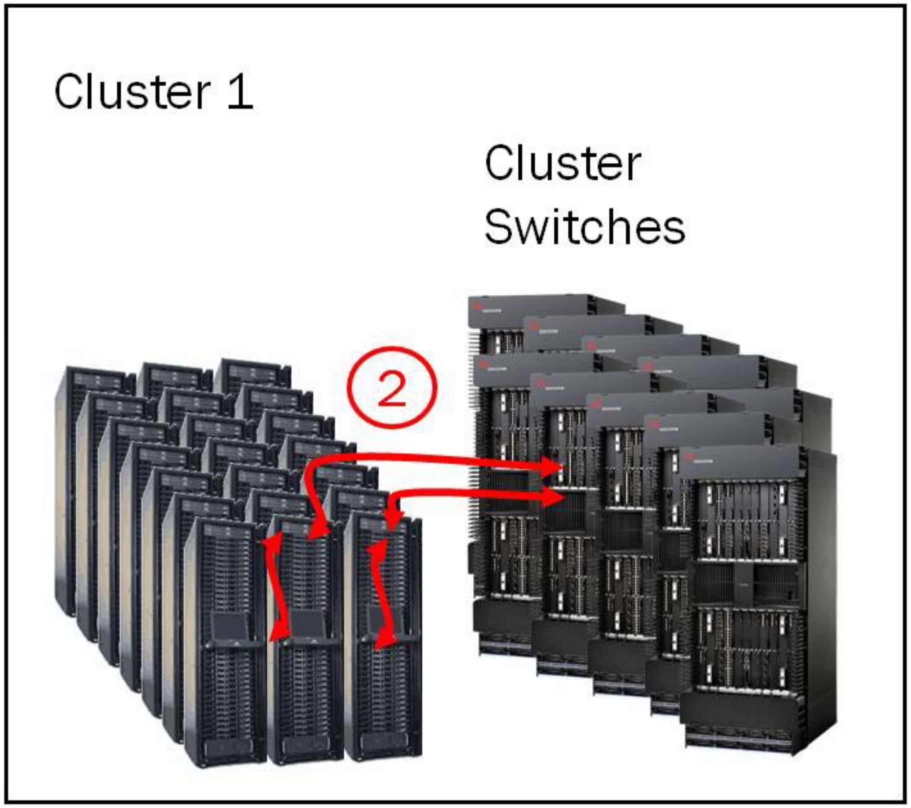
Cluster Switches

Within a Rack Communication

- Only 1 switch involved
- 1 Switch latency
- 2 links
- A few meters distance

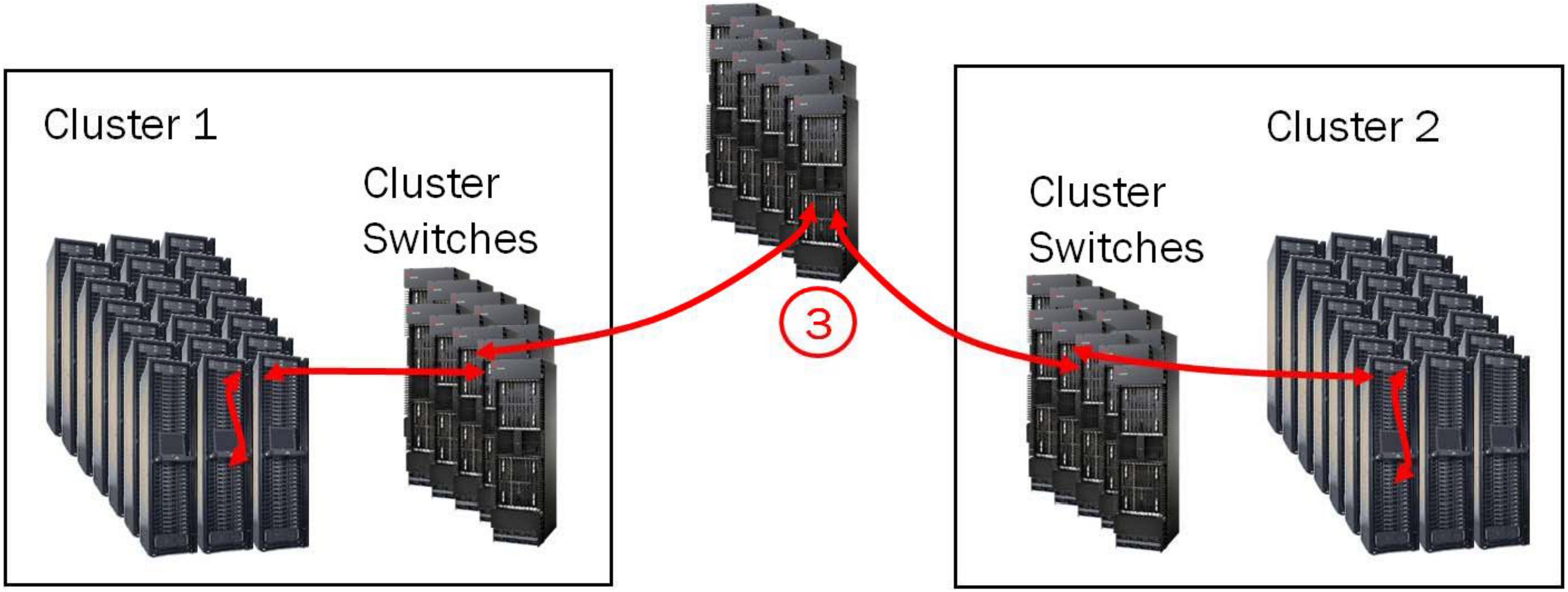
Between Racks within a Cluster

- 3 switches involved
- 3 Switch latencies
- 4 links
- <100 meter distance



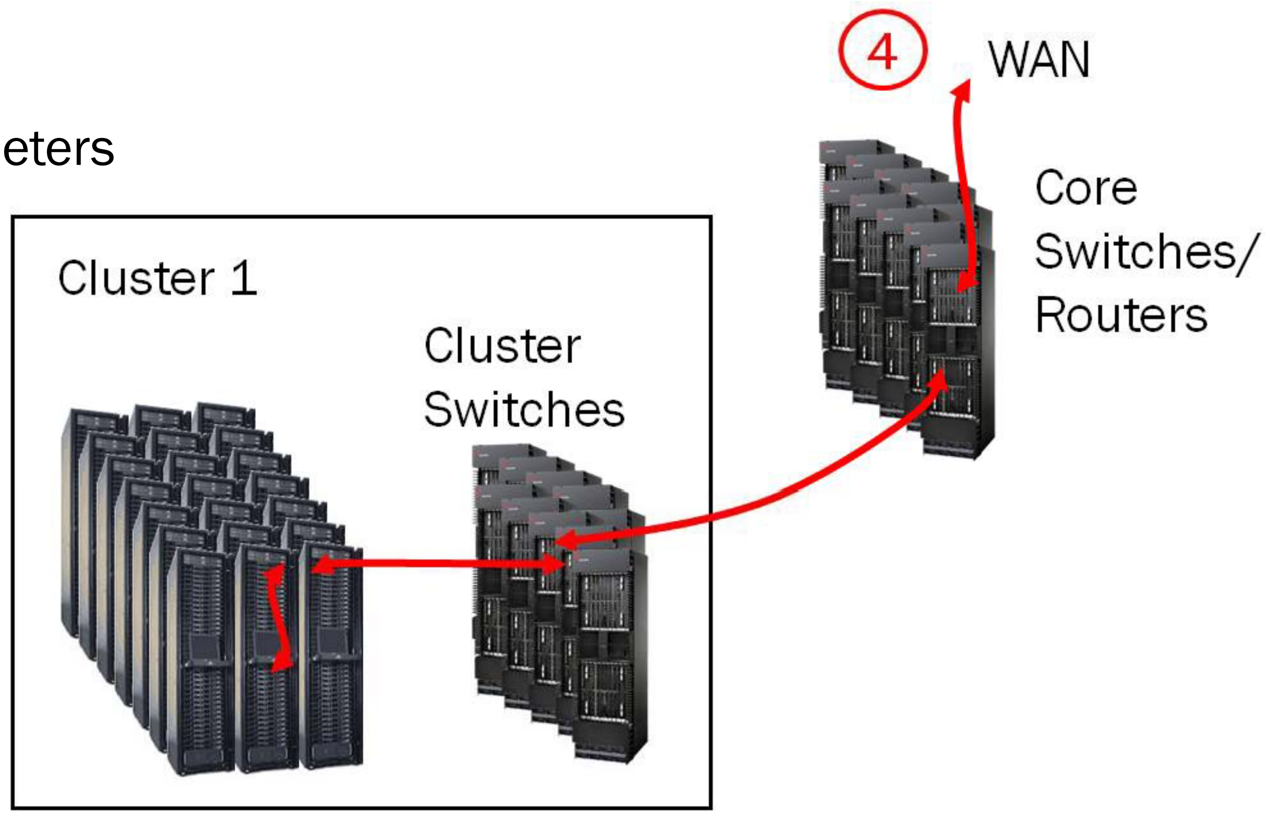
Cluster to Cluster Traffic

1. 5+ Switches involved – more can be in the core
2. 5+ switch latencies
3. 6 links
4. Hundreds of meters



Server to Internet

1. 3+ Switches involved – more can be in the core
2. 3+ switch latencies + Router Latency
3. 4+ links
4. Hundreds of meters



Cluster Summary

- Large oversubscription between:
  - Server and Rack
  - Rack and Cluster Switch
  - Cluster Switch and Core

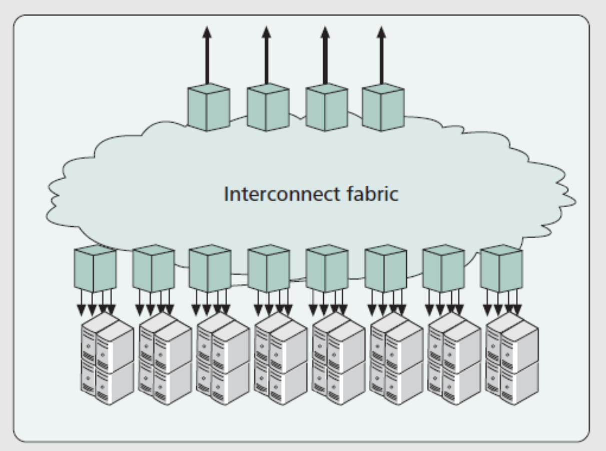
Each server producing 10-80Gbps

Each 40 server rack producing 0.4-3.2 Tbps

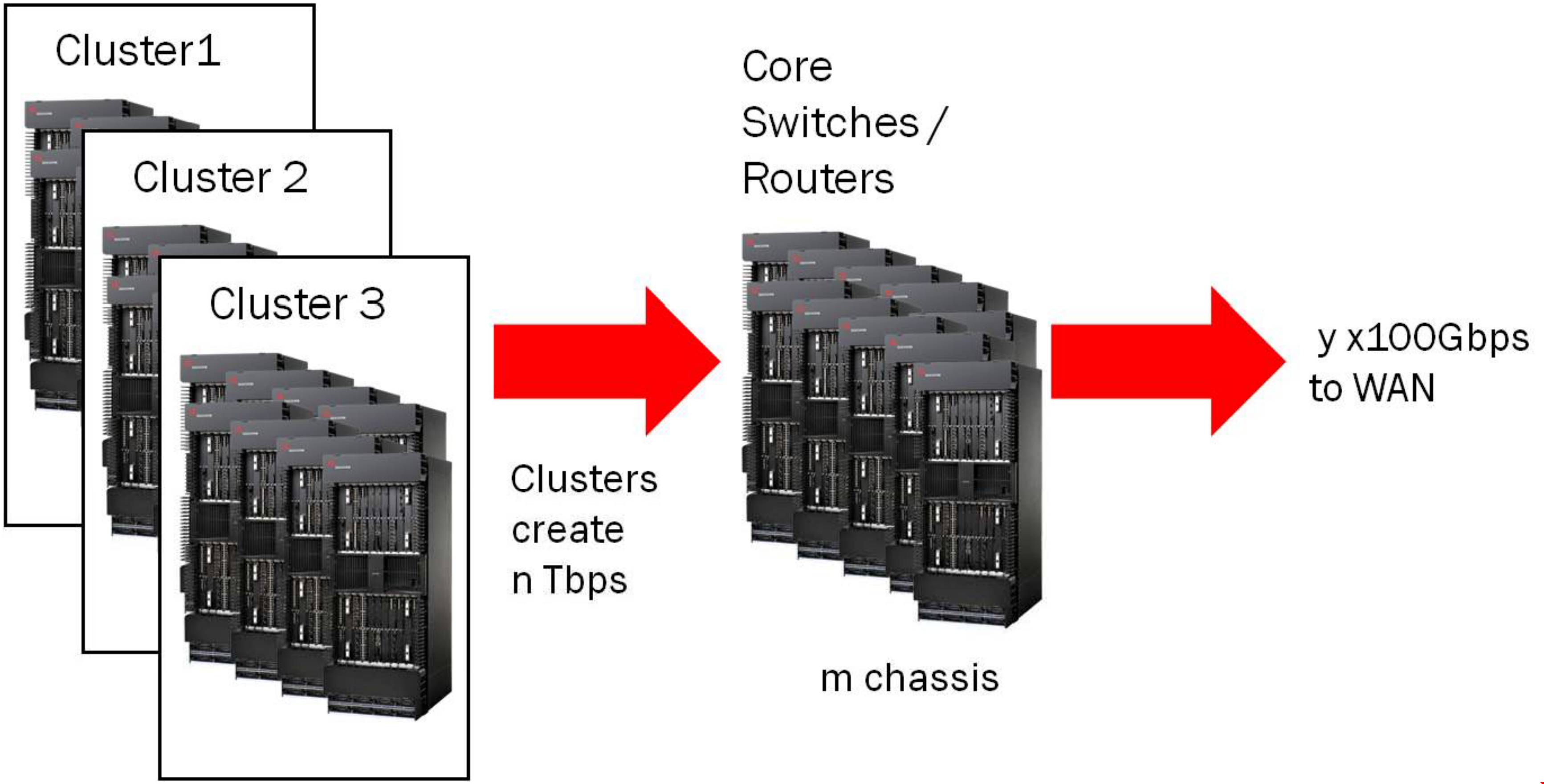
Each 1,000 server cluster producing 10-80 Tbps

Each 1,000 server cluster sends fraction of possible bandwidth to Interconnect Fabric

Data Center Summary



- Large oversubscription in Core to WAN



Cluster1

Cluster 2

Cluster 3

Clusters create n Tbps

Core Switches / Routers

m chassis

y x100Gbps to WAN

Oversubscription to WAN

Usually very high oversubscriptions to WAN

Clusters	10	10	10	10
Bandwidth / Cluster to Core (Tbps)	0.4	1	2	4
Bandwidth to Core (Tbps)	4	10	20	40
Bandwidth to WAN (Gbps)	20	40	200	400
Oversubscription to WAN	200	250	100	100

Summary

- 1,000 server clusters can produce a Tbps of bandwidth at 1 Gbps/server
- Oversubscription occurs at several levels before the data reaches the WAN
- Oversubscriptions occur because of users don’t perceive the need for 1:1 subscription and won’t pay for it
- The bandwidth demand is there, but the cost is prohibitive