Routing Domains in Data Centre Networks

University of Sussex

Morteza Kheirkhah

Informatics Department

University of Sussex

Multi-Service Networks

July 2011

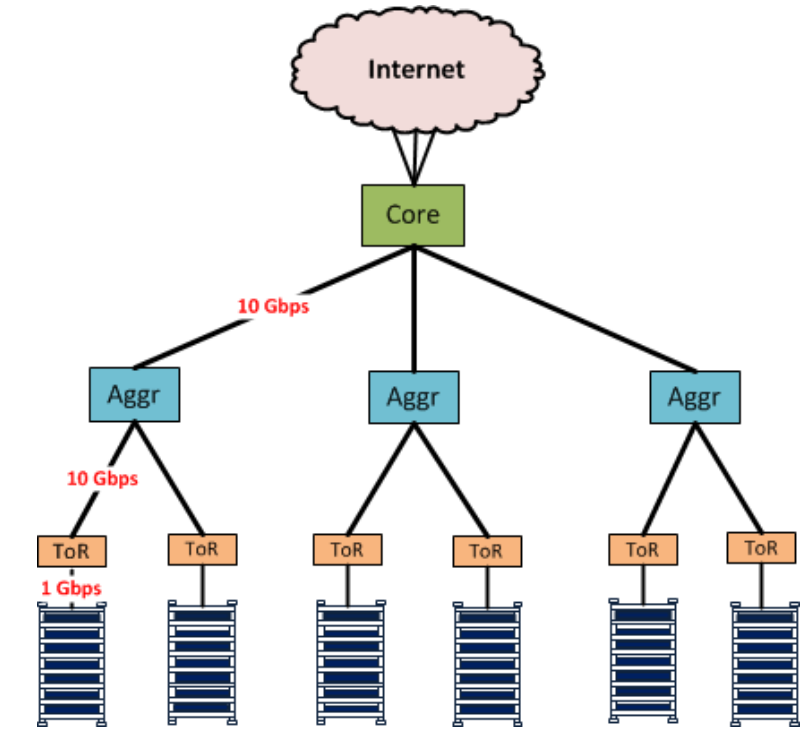
What is a Data Centre?

University of Sussex

- Large-scale Data Centres (DC) consist of tens of thousands of networked computers that provide services to cloud applications.
- Cloud applications have become an important part of our day-to-day activities.
- Examples of such cloud applications are Gmail, Google Search, DropBox, Apple iCloud and Facebook.
- Examples of such distributed services are MapReduce, Hadoop, Google File System, Google Bigtable , Amazon Dynamo and Microsoft Dryad.

Conventional Data Centre Network Architecture

University of Sussex



- Typically 20-40 servers are located in each rack, and are connected to a Top-of-Rack (ToR) switch via 1Gbps links.
- ToRs are connected to Aggregation (Aggr) switches via 10Gbps links.
- Aggregation switches are aggregated further up and connected to Core switches via 10Gbps links.

Data Centre Traffic Patterns

University of Sussex

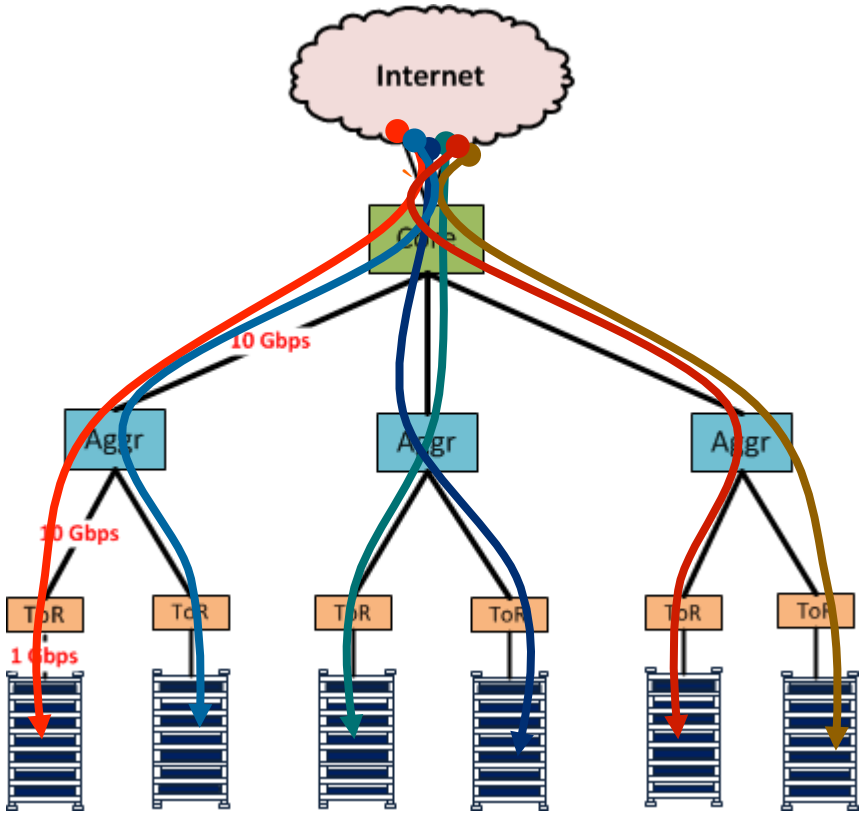
- We can categorise DC's traffic in two distinct groups.

1. **External Traffic:** Traffic flowing between external end systems (from Internet) and internal DC's servers.

2. **Internal Traffic:** Traffic flowing between internal DC's servers.

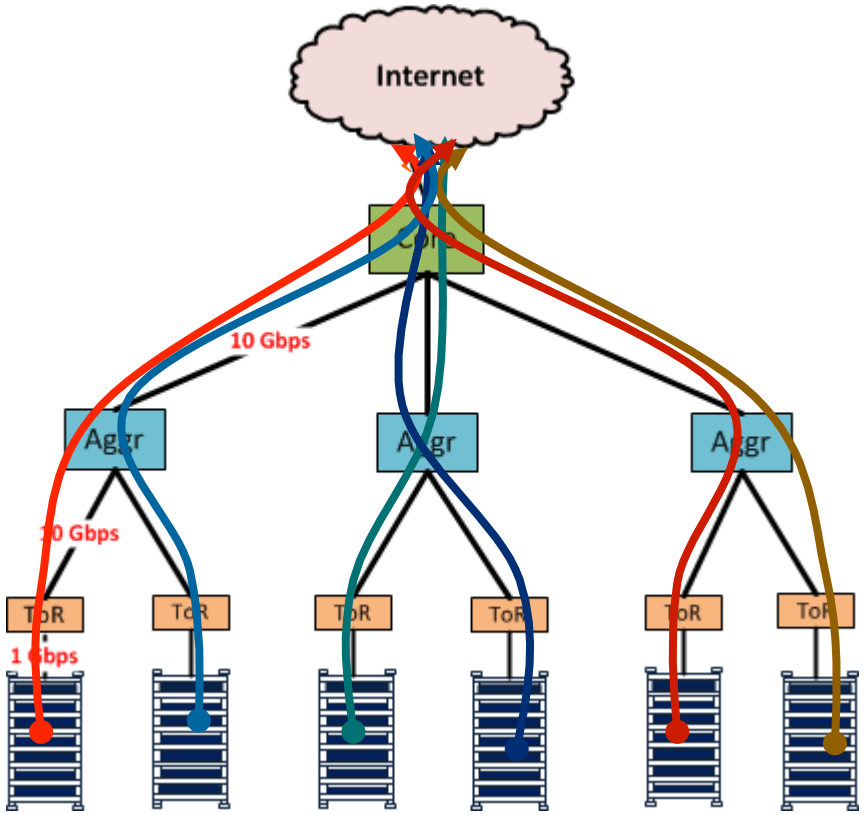
1. External Traffic

University of Sussex



1. External Traffic

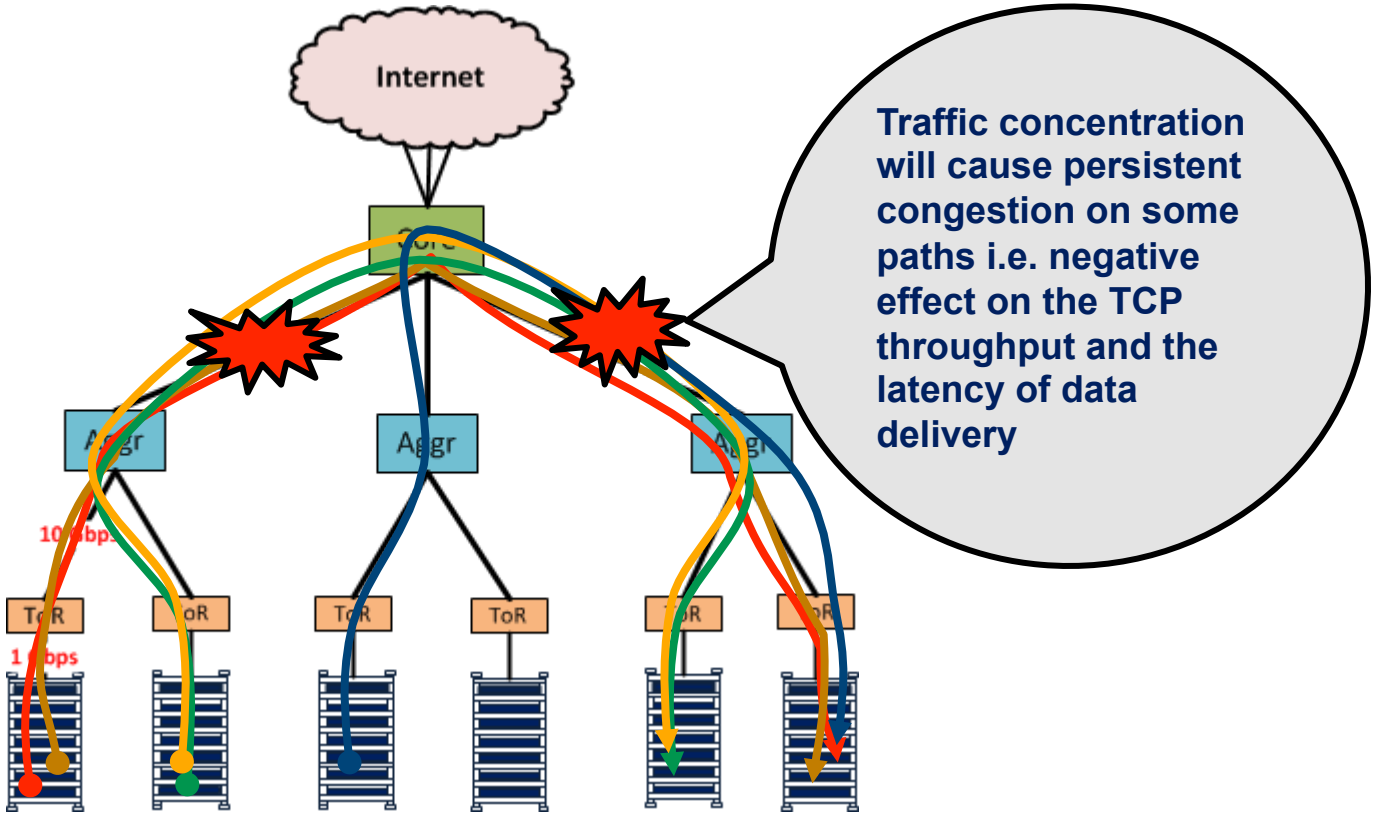
University of Sussex



- This type of traffic pattern seems to be handled well by a conventional DC topology because the load balancers can evenly distribute traffic as even as possible between servers.

2.Internal Traffic

University of Sussex



Scalable	No	Congestion Free	No
Fault Tolerant	No	Efficient use of network resources	No
Flexible	No	Energy Efficient	No

Solutions for Traffic Concentration Problem

University of Sussex

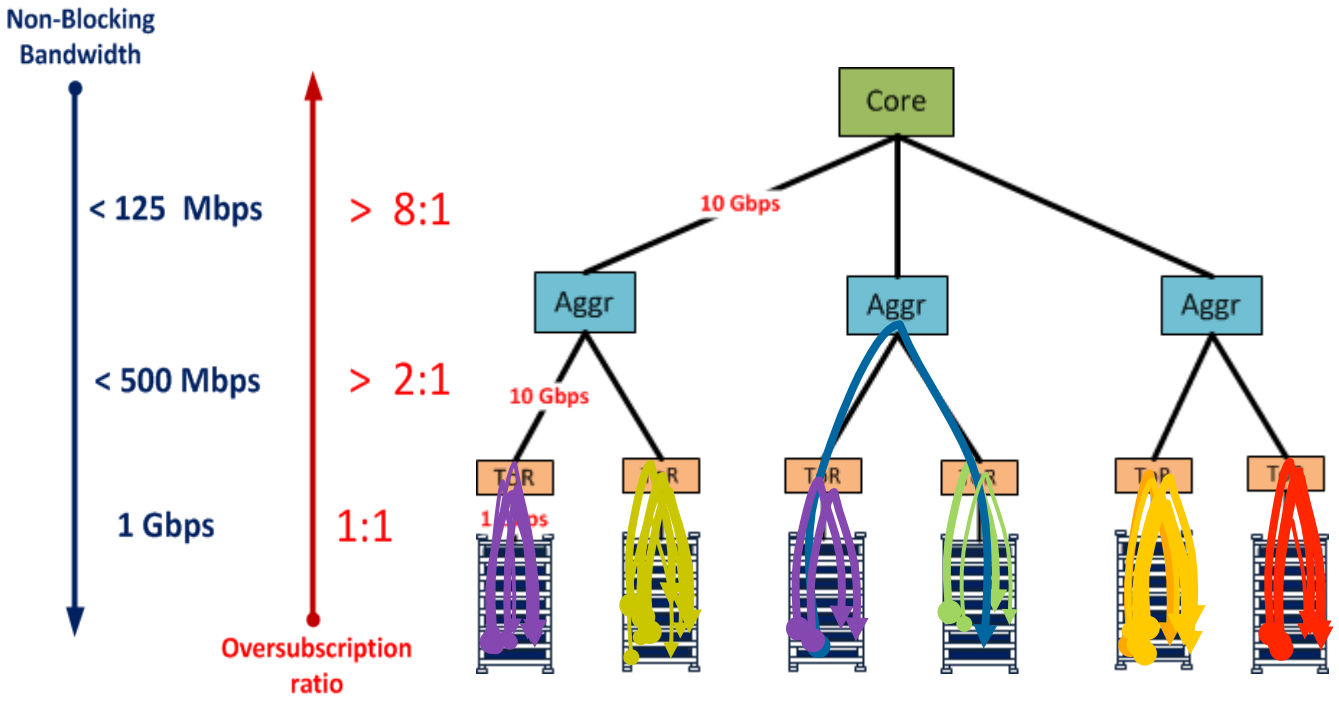
1. Localising traffic within racks

2. Providing full bisection bandwidth between all pairs of servers (VL2, FatTree and BCude)

3. Providing extra capacity on-demand when needed (Flyways and c-Through).

1.Localising Traffic to the racks

University of Sussex



- Extra servers need to be reserved for each service.
- Hard to relocate servers from one VLAN to another.

Solutions for Traffic Concentration Problem

University of Sussex

1. Localising traffic within racks

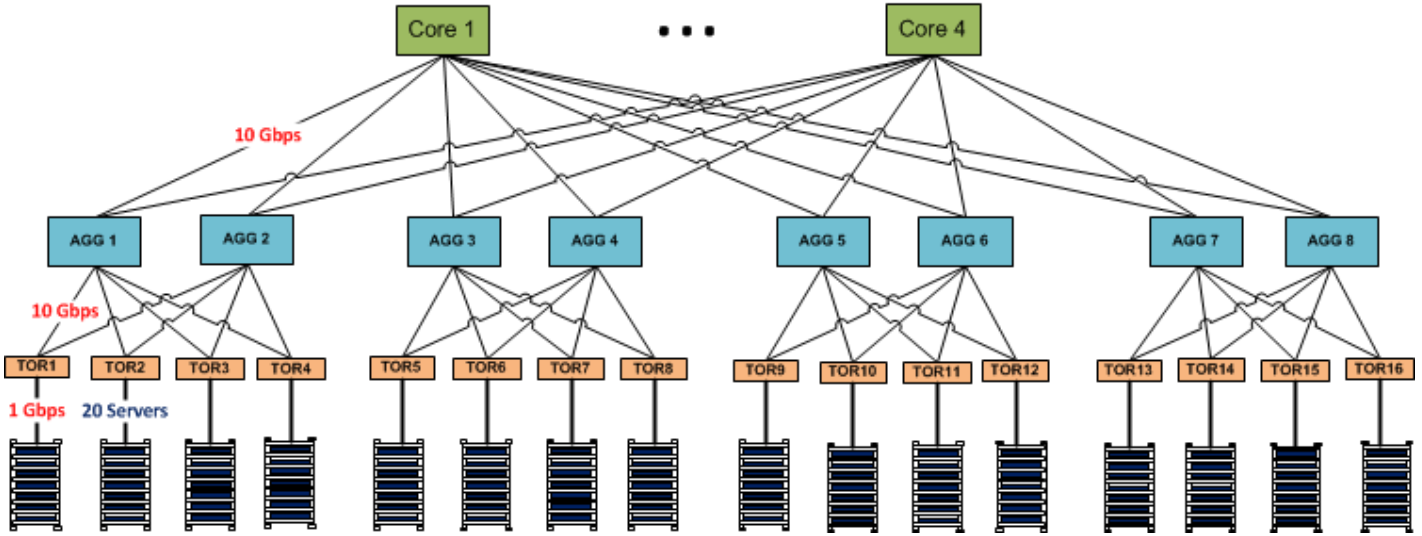
2. Providing full bisection bandwidth between all pairs of servers (VL2, FatTree and BCude)

3. Providing extra capacity on-demand when needed (Flyways and c-Through).

2.Full Bisection Bandwidth

VL2 Topology

University of Sussex

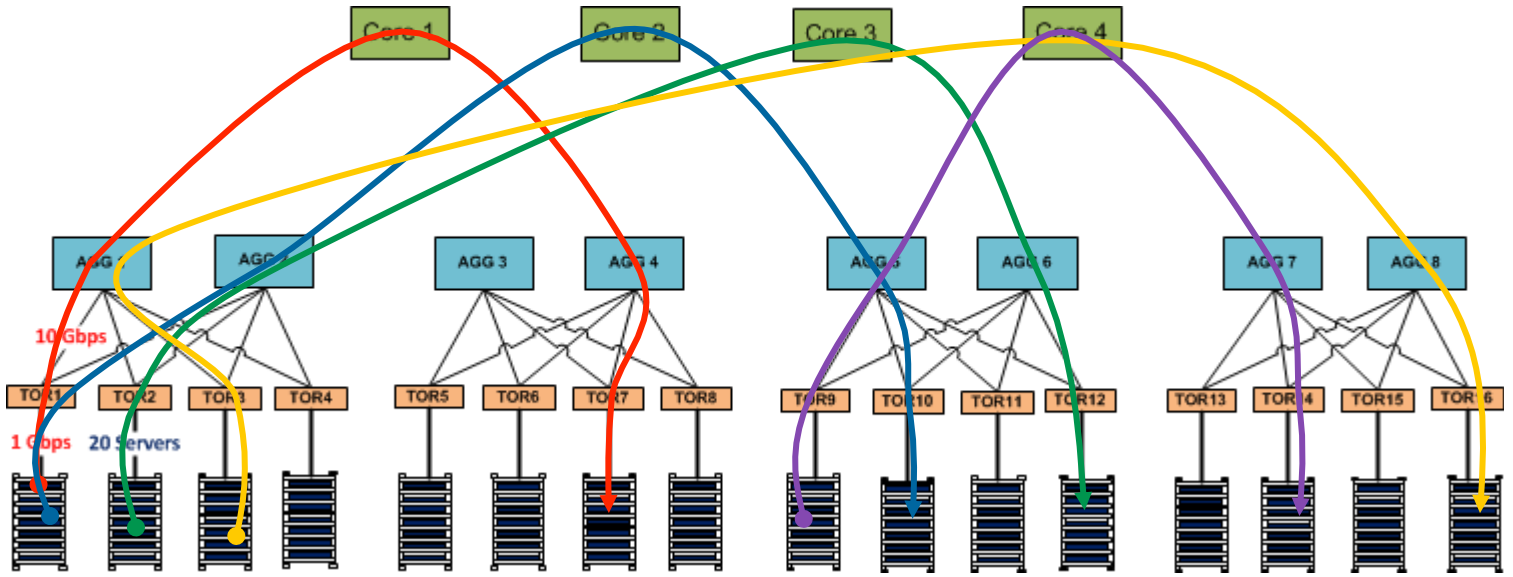


- Providing full bisection bandwidth between all pairs of servers.
- Servers can be located anywhere in the network by using flat addressing scheme.
- Using Random Load Balancing (RLB), such as Valiant Load Balancing (VLB) and Equal-Cost Multi-Path (ECMP) techniques, to exploit parallel paths in the network.

2.Full Bisection Bandwidth (cont)

VL2 Solved the Traffic Concentration Problem…

University of Sussex



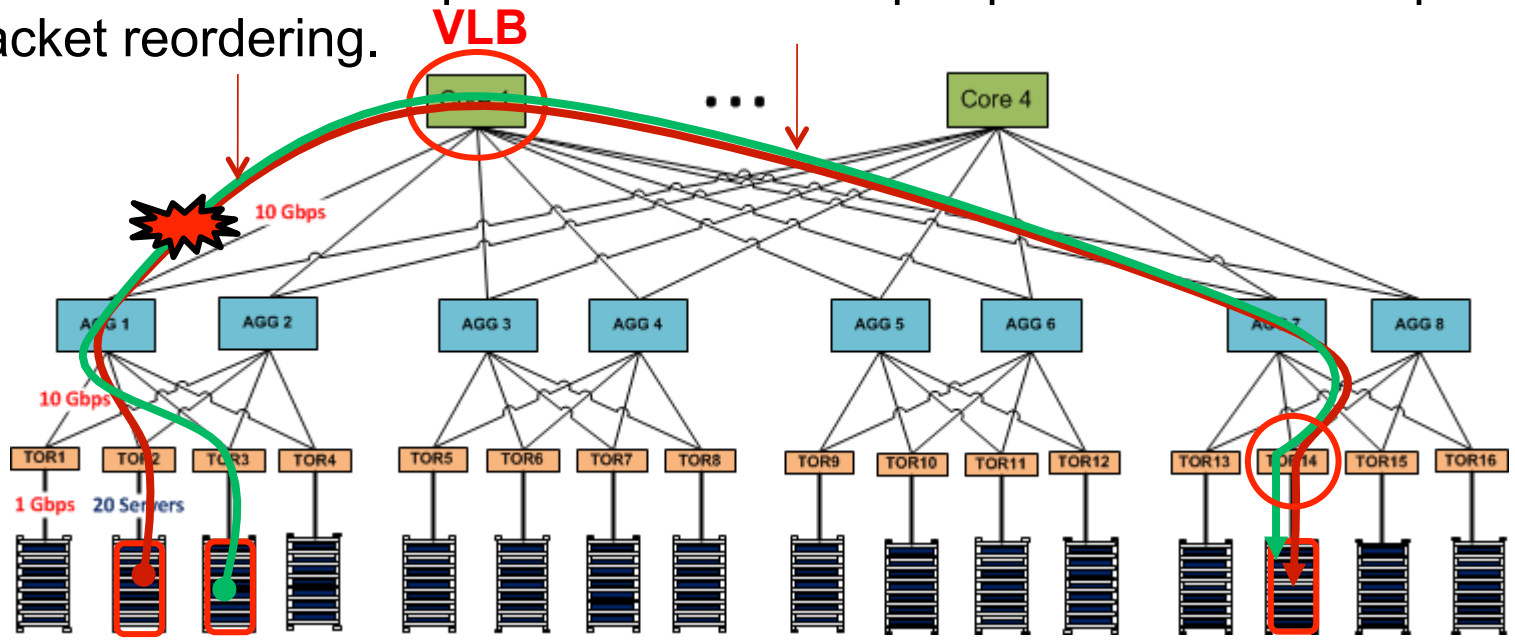
Scalable	Yes	Congestion Free	Not completely
Fault Tolerant	Yes. ToRs?	Efficient use of resources	No
Flexible	Yes	Energy Efficient	No

2.Full Bisection Bandwidth (cont)

VL2 Routing (OSPF-ECMP and VLB)

University of Sussex

- Hash of the standard five tuple **MOD** the number of equal-cost paths to the next hop.
- ECMP need to work per-flow instead of per-packet in order to prevent packet reordering.



A key limitation of ECMP is that two or more long flows can collide on their hashes and end up on the same output port (i.e. same links), creating an avoidable congestion since there are unused capacity elsewhere in the network.

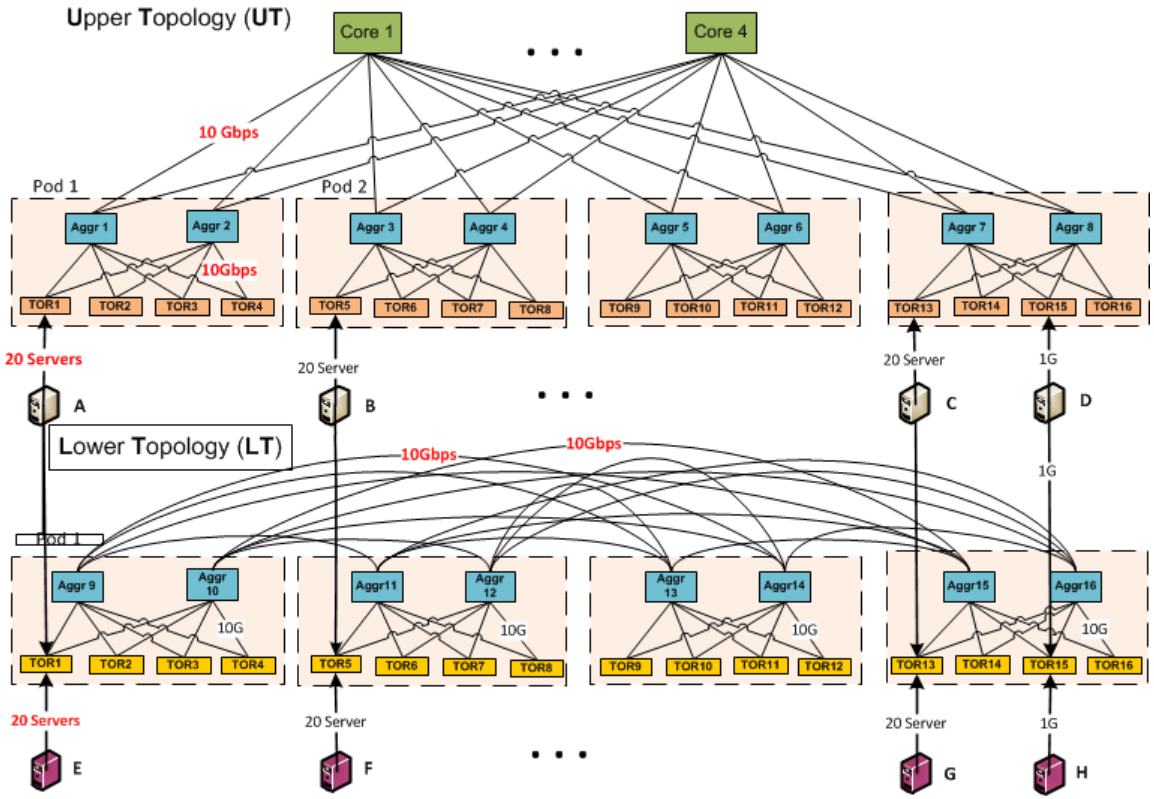
What is our Intuition and Approach?

University of Sussex

- Non-uniform network topology seems a good solution for non-uniform traffic matrices.
- Non-uniform network topology, RLB and MPTCP could be a good match to cope with traffic concentrations and to use network resource efficiently.

MKDC Topology Construction
Our Strawman Proposal

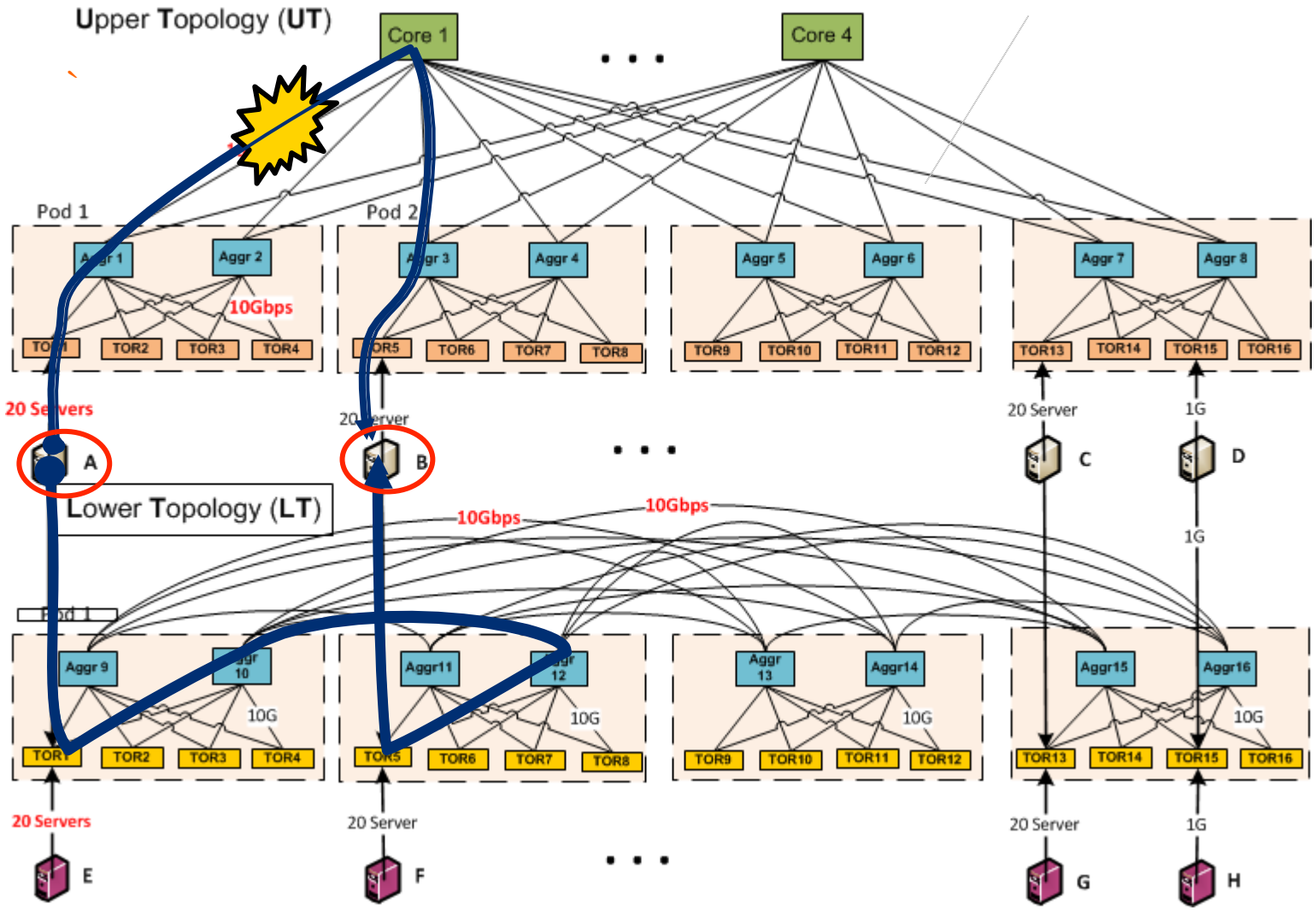
University of Sussex



- Upper Topology (UT) construction is identical to the VL2 topology with an exception that each server in the MKDC topology has two network interfaces.
- Lower Topology (LT) does not have any core switches. Aggregation switches interconnect to each other randomly. The number of ports in ToRs at the LT has doubled compared to ToRs at the UT.

MKDC Data Delivery

University of Sussex



MKDC Key Benefits

University of Sussex

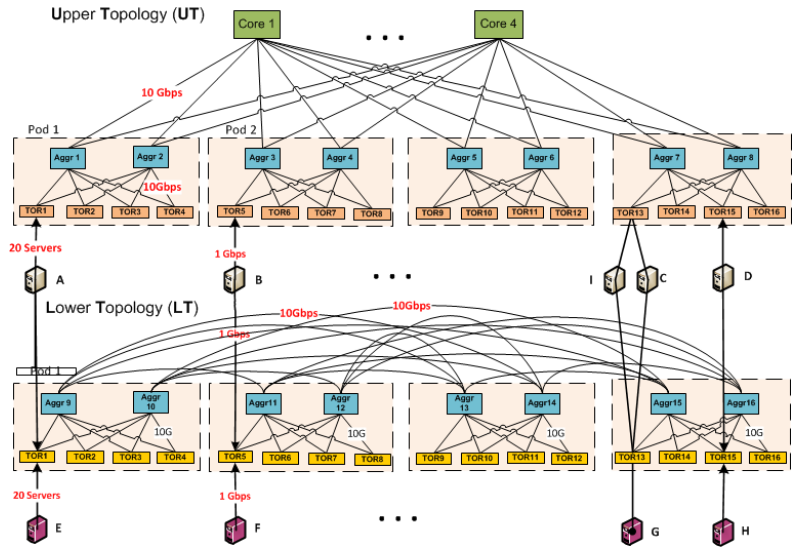
1. It can handle more traffic in/out of a data centre since most of the intra-traffic at the UT can be handled by the LT, i.e. efficient use of network resources in both UT and LT.

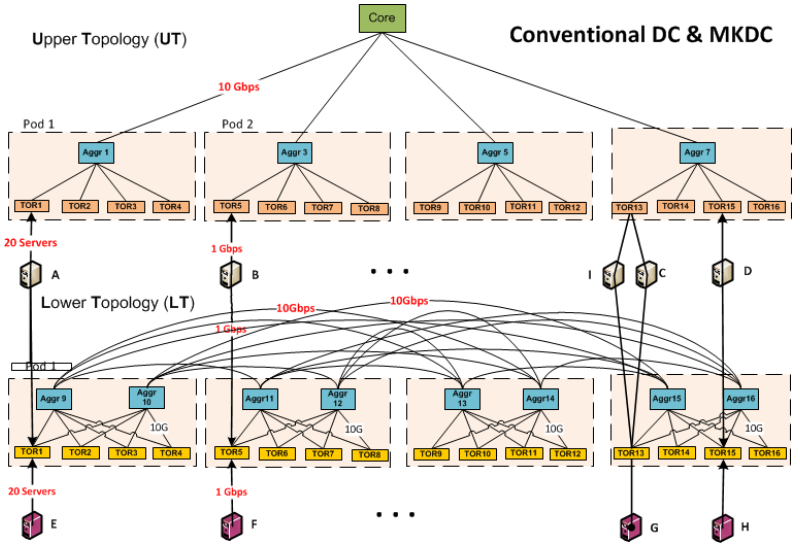
2. All traffic matrices can be handled via the LT, e.g. latency-sensitive, high bandwidth or both without any delays.

3. There are savings in energy consumption due to the elimination of cores in the LT.

4. It is not limited to the recently proposed DC topologies; it can solve traffic concentration problem in the conventional DC.

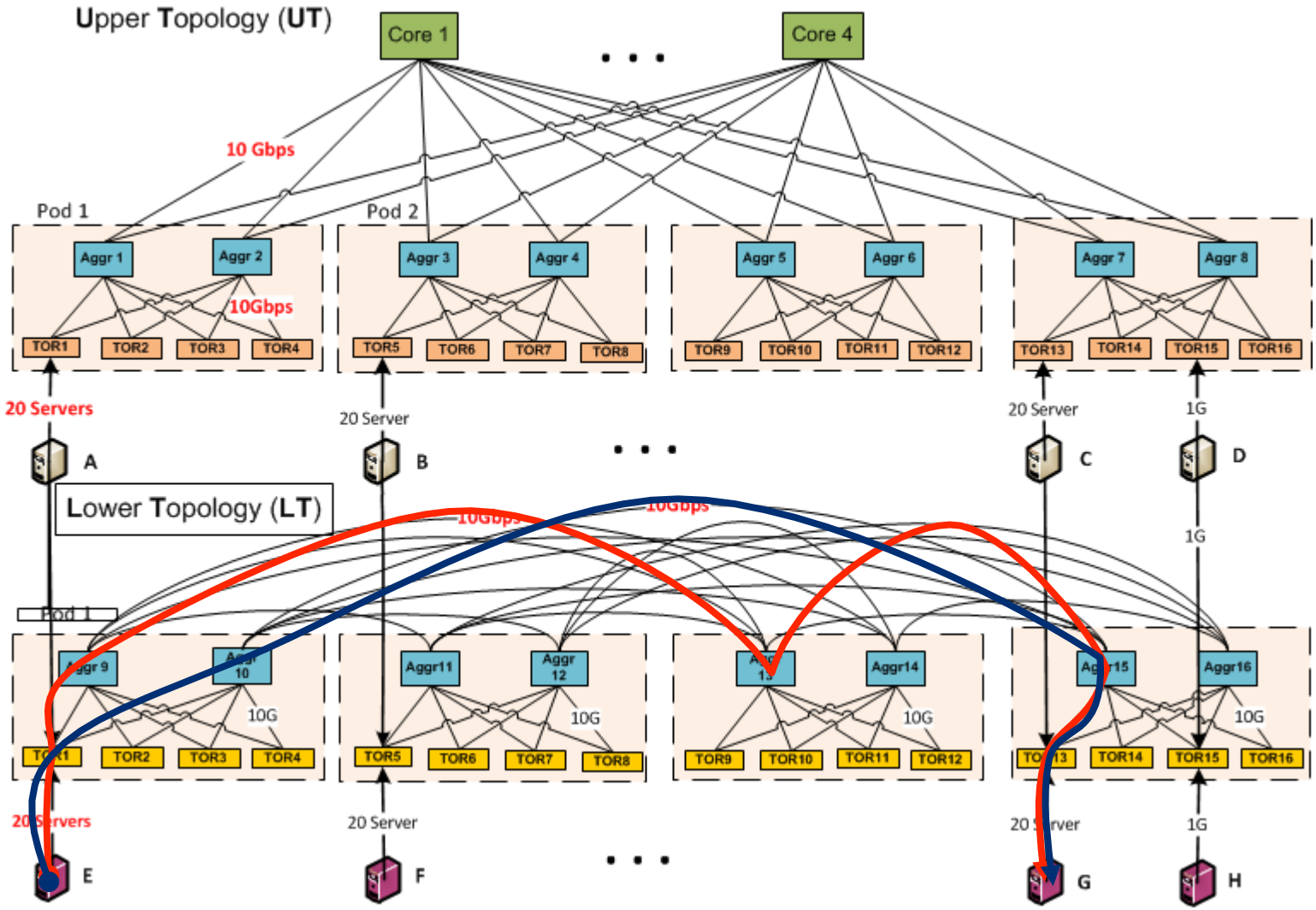
5. It Improves fault tolerance and scalability.





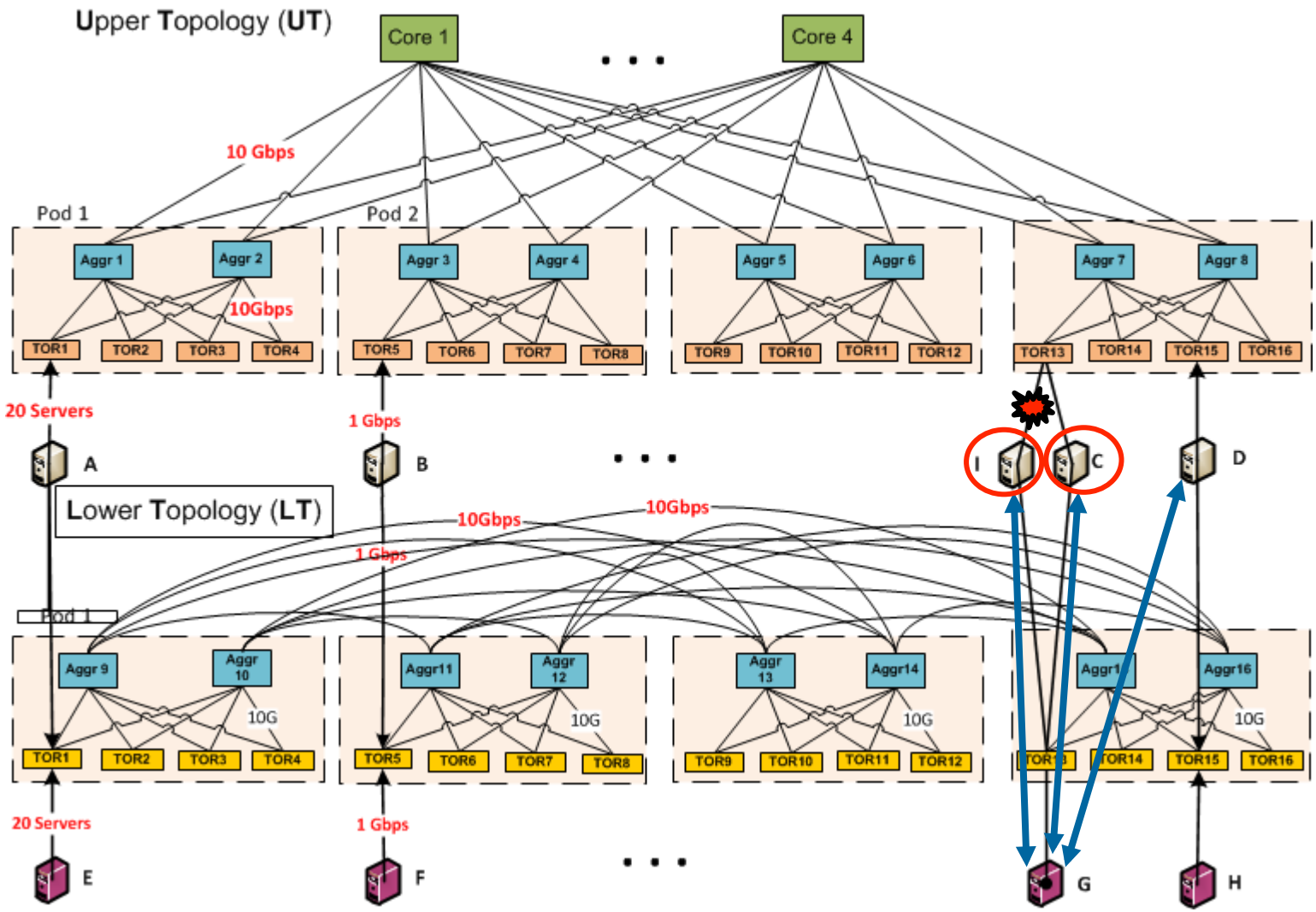
MKDC Challenges
LT's Routing

University of Sussex



MKDC Challenges (cont)
LT's External Traffic

University of Sussex



Thank you!

Question?

Solutions for Traffic Concentration Problem

University of Sussex

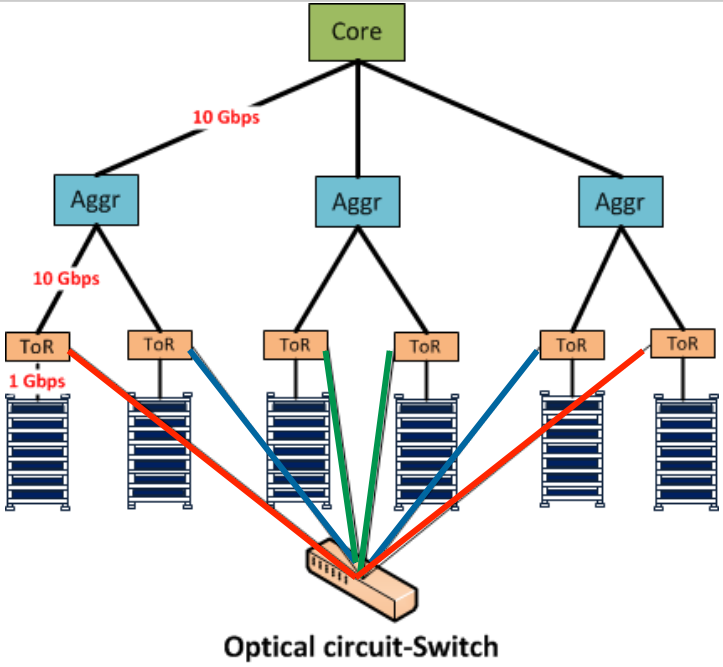
1. Localising traffic within racks

2. Providing full bisection bandwidth between all pairs of servers (VL2, FatTree and BCude)

3. Providing extra capacity on-demand when needed (Flyways and c-Through).

3.c-Through

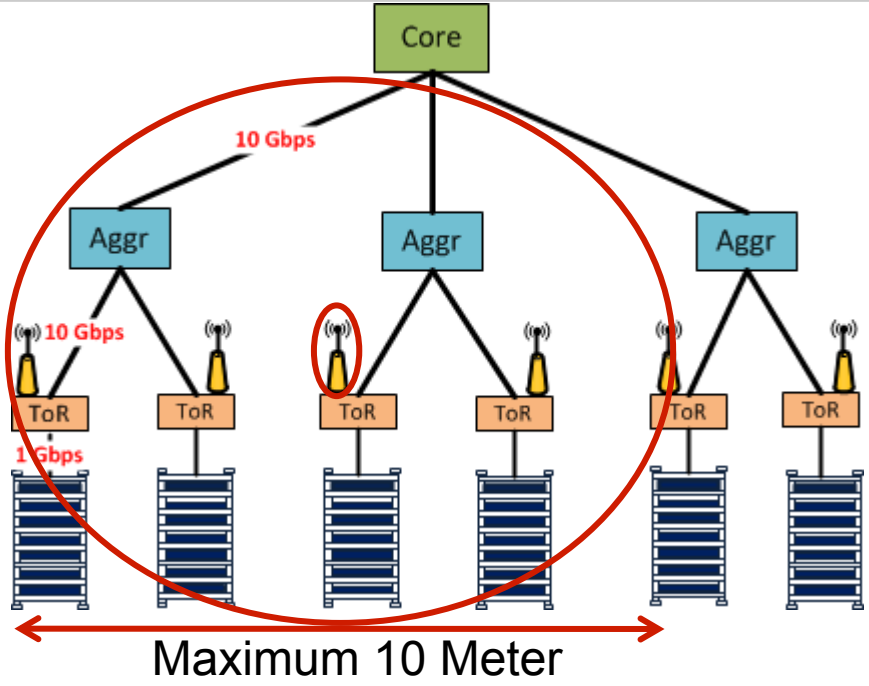
University of Sussex



- The number of ToRs may be far larger than number of available ports on an optical switch. It is thus well-suited only to a small number of traffic matrices (it only works for long-lived flows which are not latency sensitive).
- It can only provide one circuit at a time for connecting two racks. A new connection requires reconfiguration, which will introduce overhead and delay.

3.Flyways

University of Sussex



- Radio coverage: ~max 10 meter (very short range).
- Maximum Capacity: ~1 Gbps (60 GHz).
- Improves performance only for slightly oversubscribed network.
- Suboptimal: the number of ToR switches that can communicate with one another is merely depended on the number of ports in each switch.
- Slow Reaction to congestion: central controller requires to detect where these links are needed and then actives them on-demand i.e. it works with some traffic matrices.