

Big Data Infrastructure

where *speed* matters

Dr. Posco Tso

Senior Lecturer
Department of Computer Science

About Me

- PhD, City University of Hong Kong (QS 57th worldwide)
 - ★ 1 US Patent and 1 Start-up
- SICSA Next Generation Internet Fellow
 - ★ Built the cloud and big data testbed
- Senior Lecturer
 - ★ Wants to do better!

Big Data Infrastructure

- How Big is “Big Data”?
 - ✦ > 1 TB
 - ✦ Simple C/C++ code with legacy database will beat “Big Data Analytics” systems in speed.
 - ✦ Not able to leverage parallelism
- Components for Big Data infrastructure
 - ✦ Compute cluster(s); Data analytics tools; File systems/databases

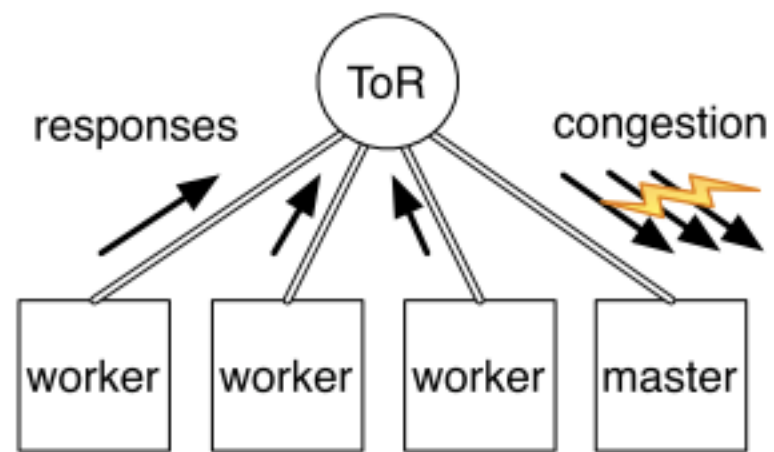
Big Data Infrastructure

Compute Clusters	Data Analytics Tools	File Systems/ Databases
(Virtual) Machine cluster(s); High performance computing (HPC) clusters;	Hadoop (framework); Spark; Storm; MapR; Pig; ...	HDFS; S3; GFS; Cassandra; HBase; BigTable; MongoDB; ...

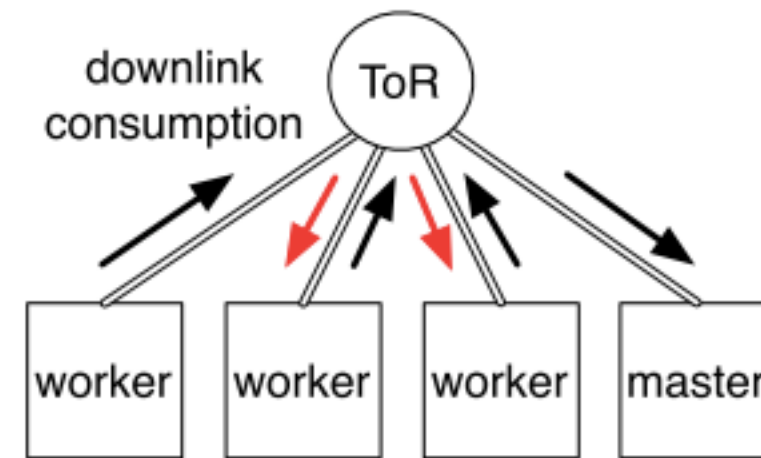
Big Data Infrastructure

Compute Clusters	Data Analytics Tools	File Systems/ Databases
(Virtual) Machine cluster(s); High performance computing (HPC) clusters; My research interests	Hadoop (framework); Spark; Storm; MapR; Pig; ...	HDFS; S3; GFS; Cassandra; HBase; BigTable; MongoDB; ...

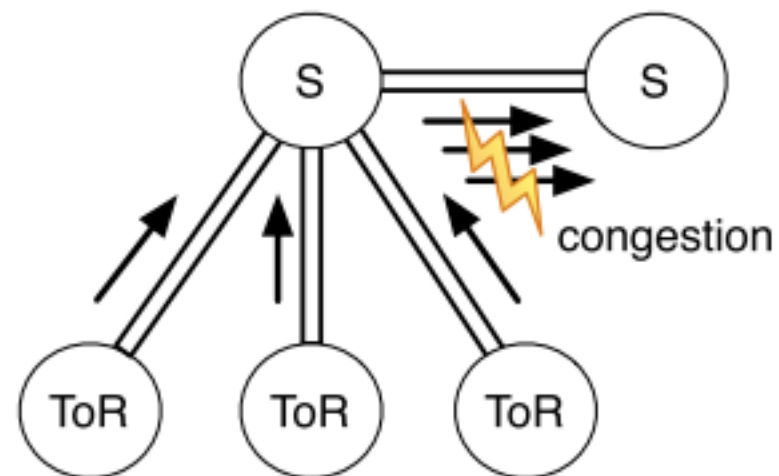
Example 1



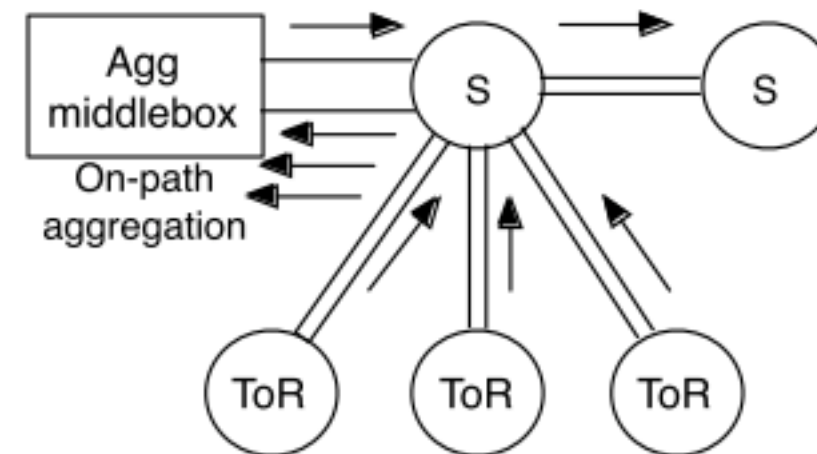
(a) Rack-level aggregation



(b) Chain aggregation



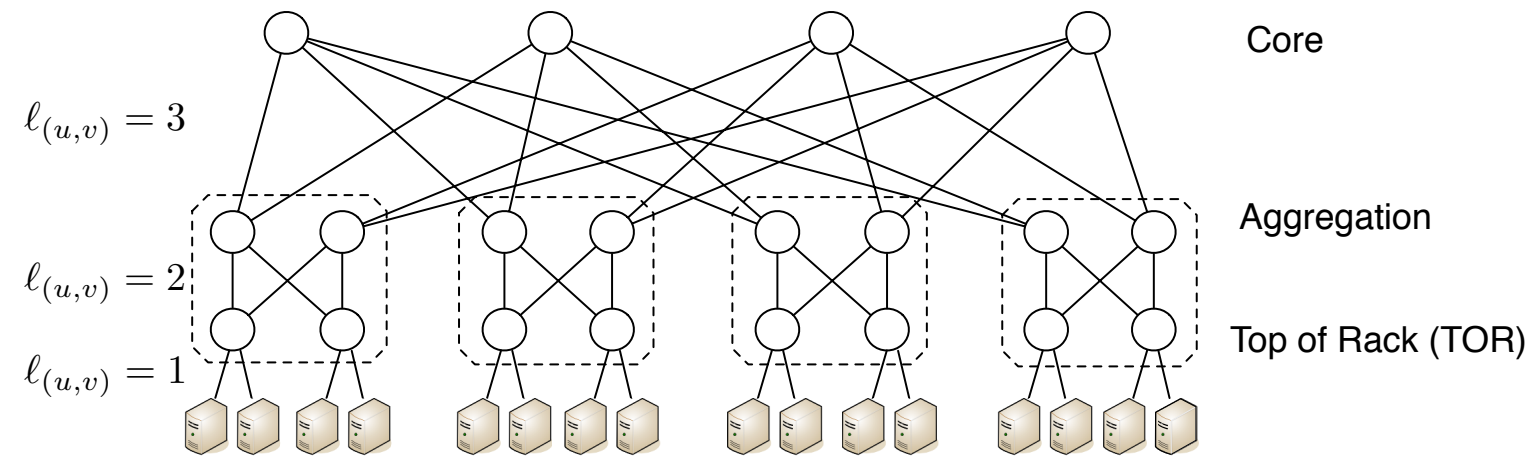
(c) Cross-rack aggregation



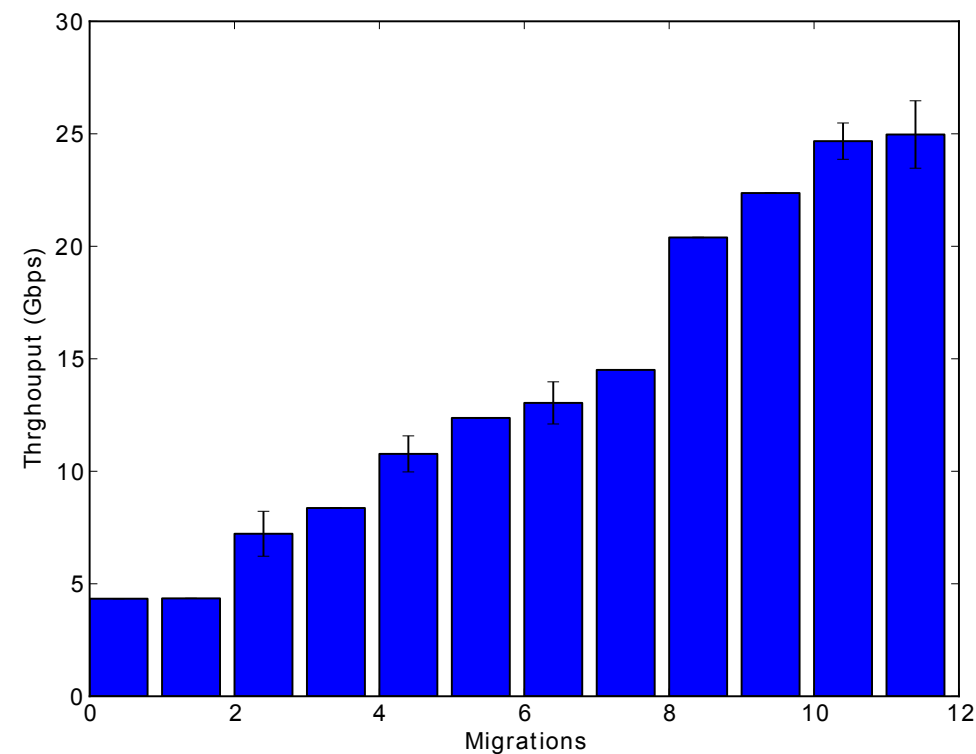
(d) On-path aggregation

On-path data aggregation

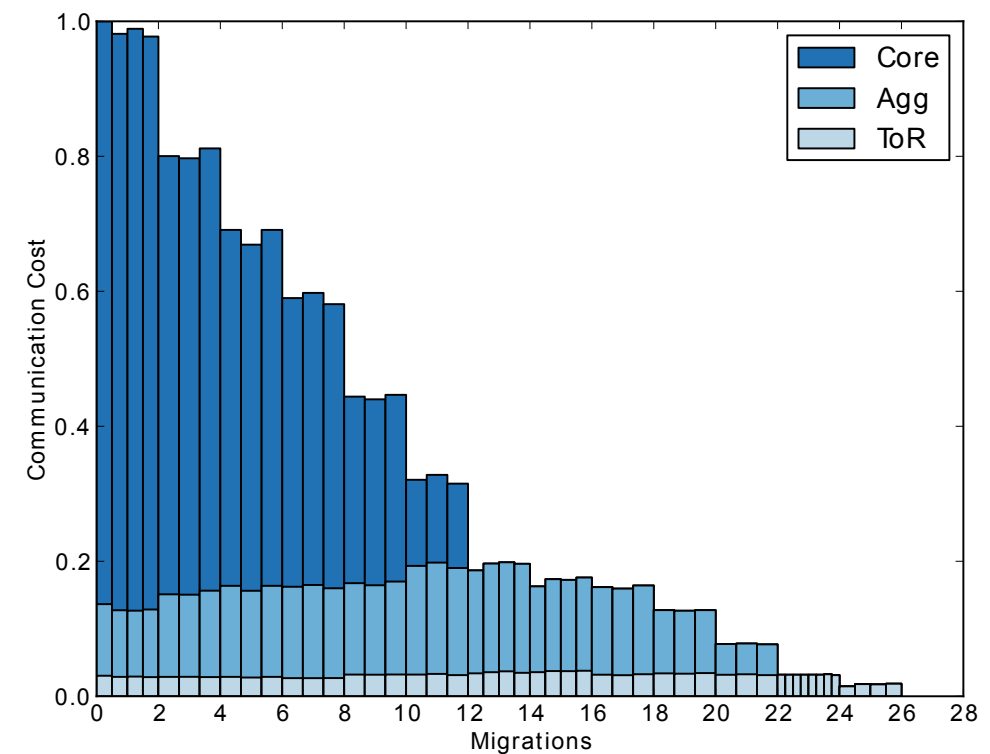
Example 2



(a) Network topology



(b) Aggregate throughput



(c) Link utilisation at different layers

Traffic-aware virtual machine migration



p.tso@ljmu.ac.uk
@drscake