Zurich Hotspot Benchmark

Output Generated Single Hotspot in Multi-hop Topologies

Mitch Gusat and Cyriel Minkenberg
IBM Research
ZRL, Switzerland

Why Output Generated in Multi-Hop?

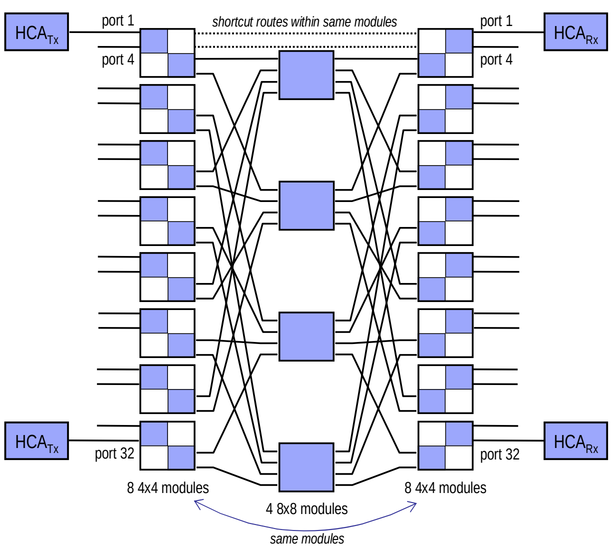
- OG is a stress test
  - quadratic effect on G_d of hotspot degree (HSD) and bottleneck's service rate μ_{HS} => $G_d \sim (\mu_{HS}/HSD)^{-2}$

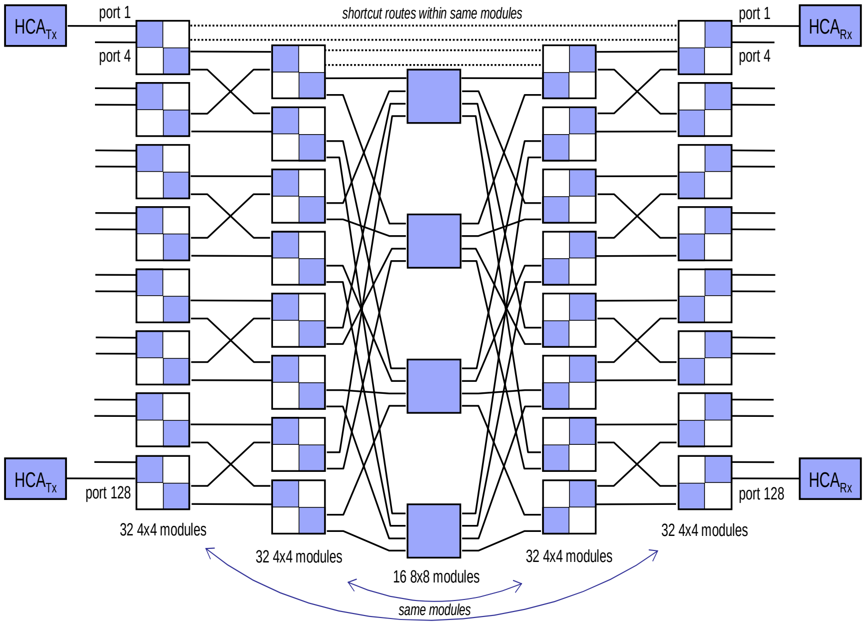
- Single stage case was covered by IBM and Cisco during the Nov.-December sim calls

- Next phase: Study OG-HS in multistage fabrics

- What fabric topology should we choose?

A. Baseline Multistage Topology: Bidir Fat Trees (FT)





- 2-level / 3-stage bidir MIN
- Simulate: 8 – 32 nodes
- Time per run: < 1.5 hrs.

- 3-level / 5-stage bidir MIN
- Simulate: 128 – 2K nodes
- Time per run: TBD

Fat-trees: Scalable, w/ excellent routing and performance properties. Optimum performance/cost with current trends in technology. Can emulate <u>any</u> k-ary n-fly *and* n-cube topology. Large body of knowledge.

Issues Related to Fat-Tree Topology Simulation

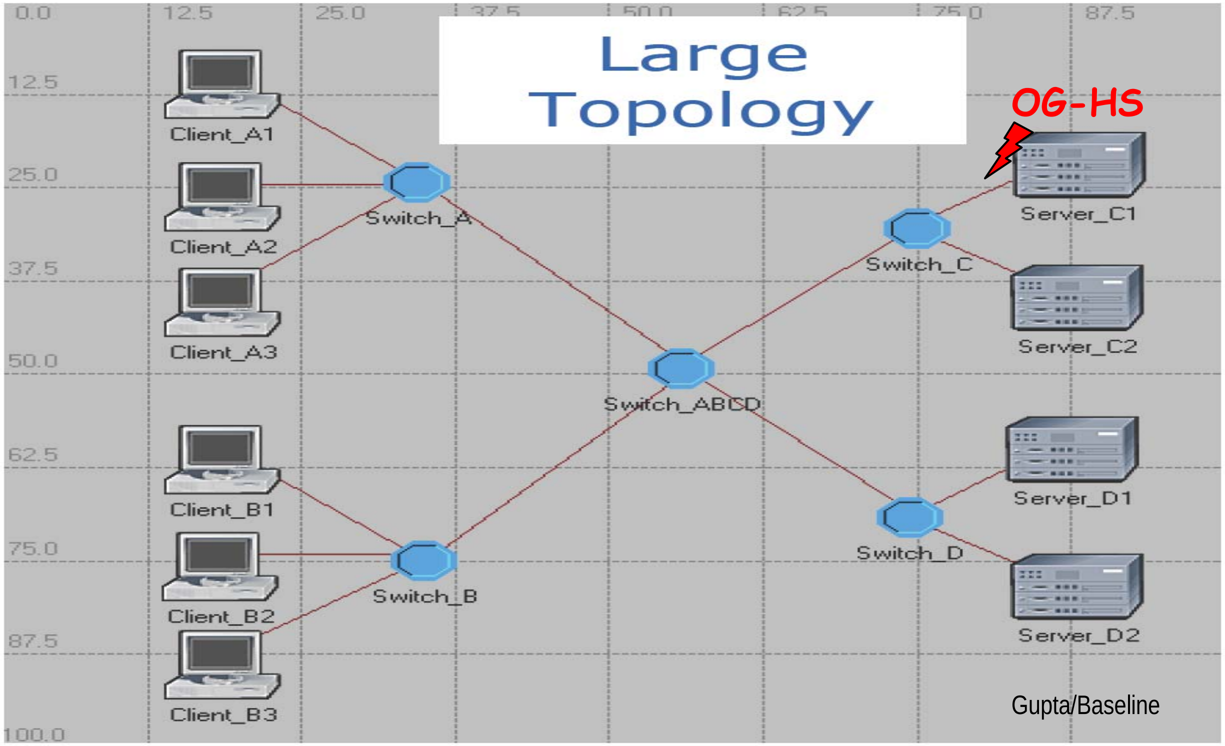
- While desirable, FTs are not an easy target…

  - MUCH larger simulation model => memory and runtime
    - may elicit an upgrade of the simulation environment, even new tool

  - Specific problems to be solved
    - deadlocks of 2 types:
      - circular dependency
      - routing loops
    - routing
    - RLT – to – CPID association
    - …

- Interim step to FTs: 'lumped' trees

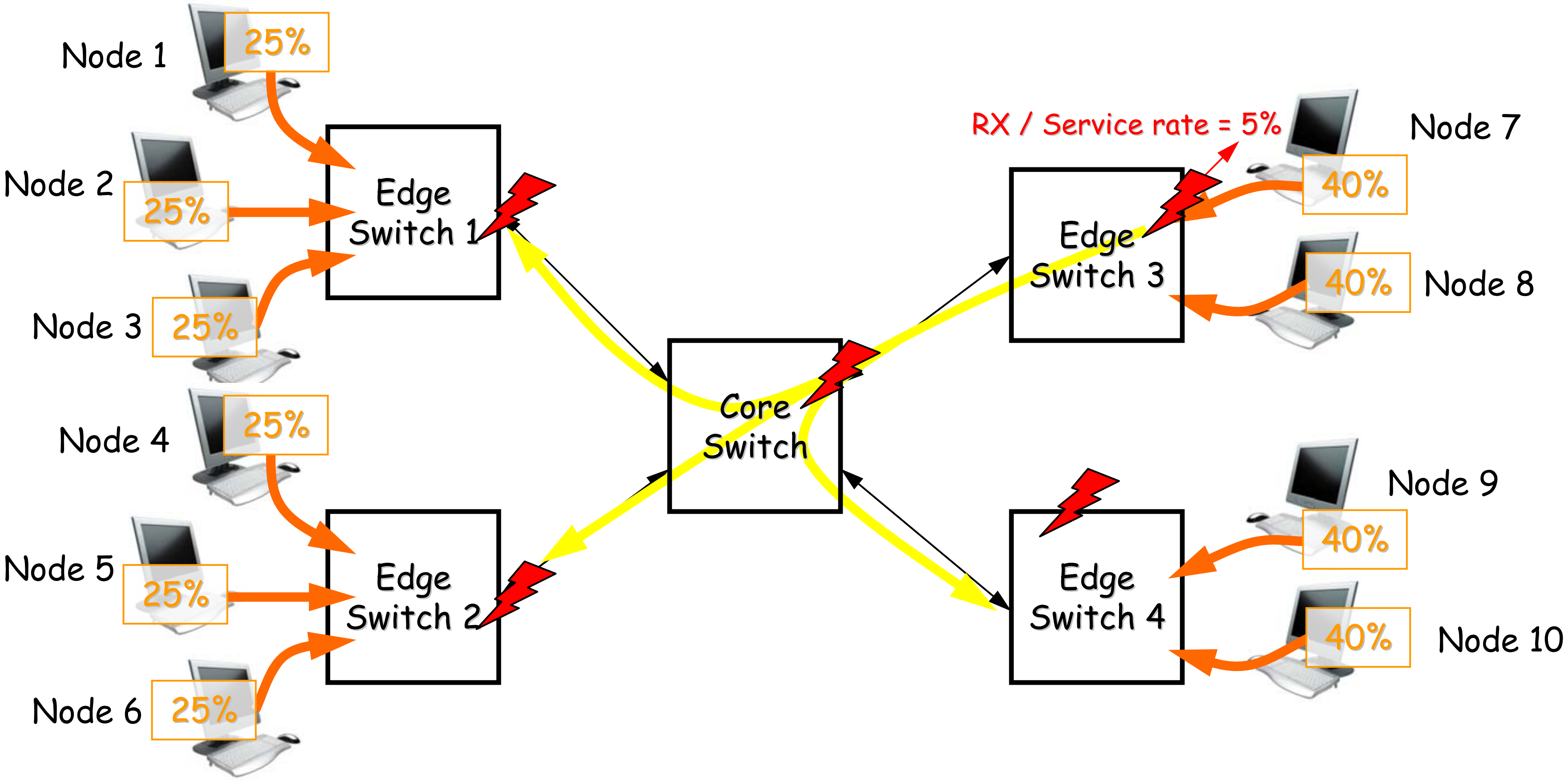
B. 3-Stage Lumped Tree with OG Hotspot



1. edge nodes are bidir
   1. dual function: each client and server will simultaneously source and sink traffic
   2. hence 10 sources and 10 sinks
2. 3-hop network

- Hotspot period: from t_i to t_f, the sink at Server_C1 slows down its RX rate (as in ZRL's tree-based study)
- Victim traffic
  1. Non-selective: Constant background load, uniform distribution
  2. Selective: Elect designated victim flows

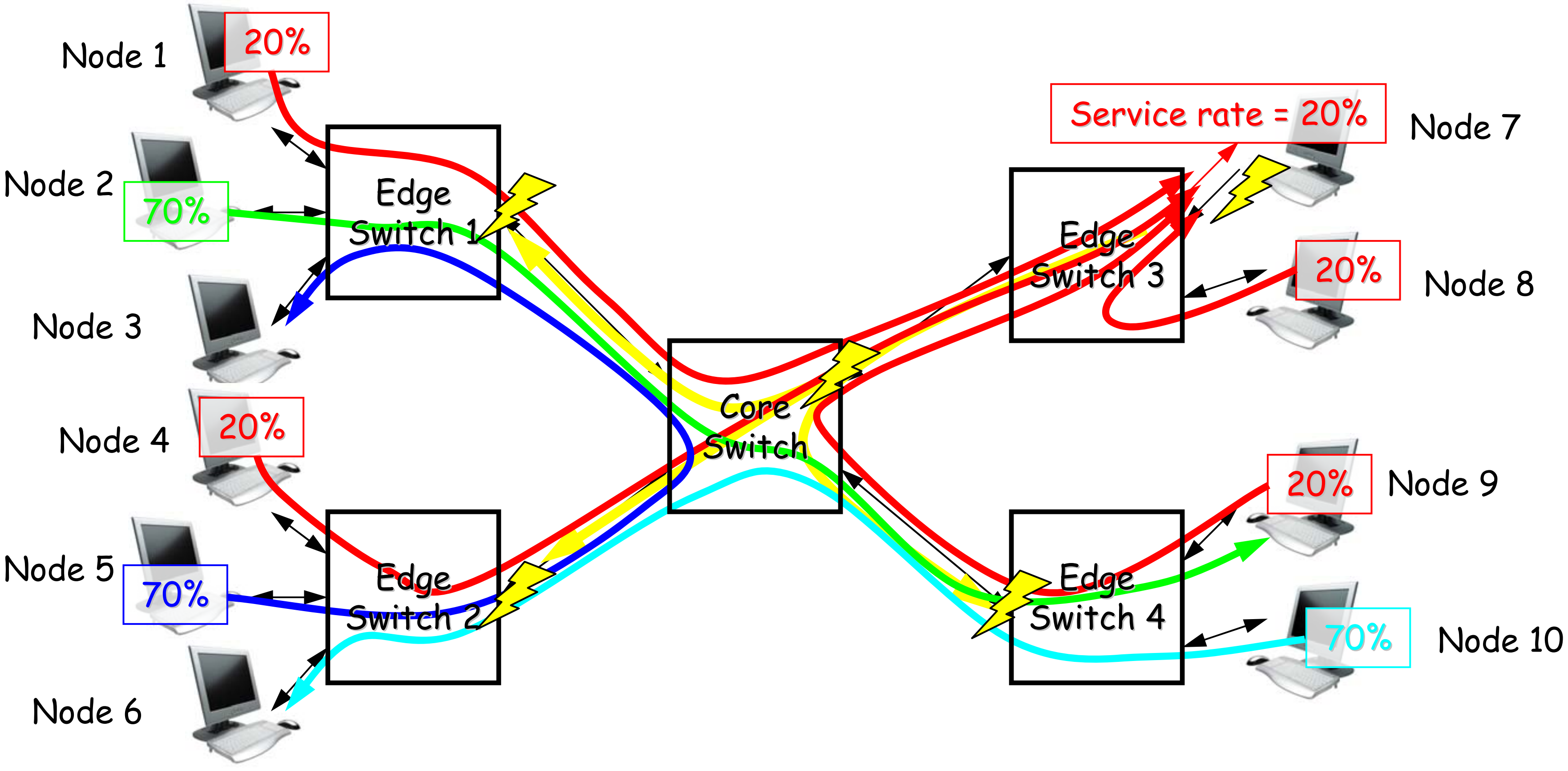
OG Hotspot Example #1: Background “Victims” (non selective)



- Offered load definition: Mean aggregate load = (6*.25+4*.4)/10 = 31% (3.1 Gb/s)
  - All: Uniform destination distribution (background traffic)
  - nodes 1-6 = 25% (2.5 Gb/s)
  - nodes 7-10 = 40% (4 Gb/s)
- Primary hotspot: node 7 service rate = 5% (RX only)
  - if saturation tree backspreads => 5 secondary congestion points (induced hotspots)

Obs. All switches and all flows affected.

OG Hotspot Example #2: Selected “Victims”



- Four culprit flows of 2 Gb/s each from nodes 1, 4, 8, 9 to node 7 (hotspot)
- Three victim flows of 7 Gb/s each: node 2 to 9, node 5 to 3, node 10 to 6
- Node 7 service rate = 20%
- Five congestion points
  - All switches and all flows affected
  - Fair allocation provides 0.5 Gb/s to all culprits and 7 Gb/s to all victims

Obs. All switches and all flows affected.