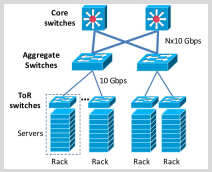
Research 2 (series): Performance Enhancement in Overloaded Networks

Motivation

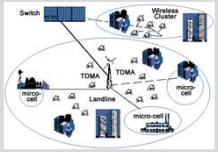
- Extensive analysis over underloaded systems

- However, overloaded situation becomes more frequent in IoT but under unsystematic study & results

Server Farm



Datacenter



Mobile System

Communication infrastructure in Smart Grid; Cloud; HPC; Edge computing, etc.

Contributions

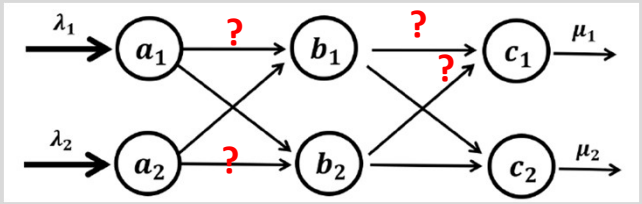
1) Model queue dynamics by flow, which generalizes different network settings.
- overload/underload, shared/split buffer, etc.

2) Propose network policies that optimize QoS metrics: latency, fairness, throughput, under network overload.

Two papers are published, and one is submitted.

(1) Latency

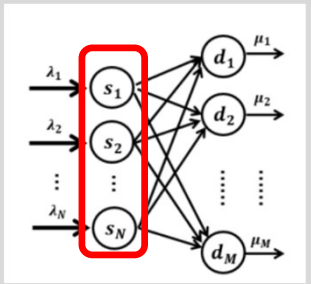
Set service rates to minimize queueing latency when network is overloaded:



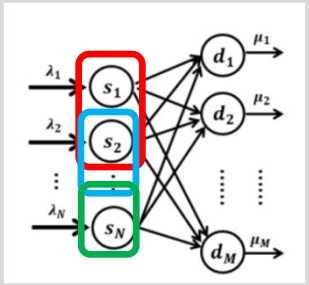
Prove & evaluate that
- Setting max rates on all links is generally **NOT** optimal.
- Properly setting smaller rates reduces latency & saves energy.
- 10% ↓ in avg. & 50% ↓ in max latency.

(2) Fairness

Balancing input loads when egress buffer is bounded:



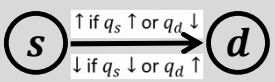
Centralized



Distributed

(3) Stability

A queue-based policy design criterion to stabilize the networks that generalizes a set of policies:

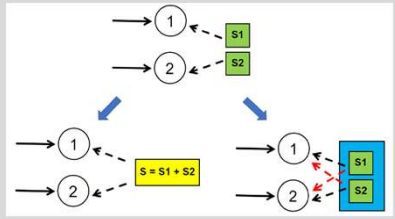


$$\frac{\partial g_{ij}(q_i, q_j)}{\partial q_i} \geq 0, \quad \frac{\partial g_{ij}(q_i, q_j)}{\partial q_j} \leq 0$$

Ongoing

- Study the risk of resource pooling failures on network performance, where the above results may be applied.

Buffer sharing; Power supply



Cloud server cluster; Power sharing

Connection to Industry Research:
- **Optimization** & **algorithm design** on network infrastructure: load balancing in VMs, job scheduling for ML tasks, resource allocation in cloud platform, etc.