CloudCom 2014
Dec 18, 2014

Virtual Network Allocation for Fault Tolerance with Bandwidth Efficiency in a Multi-Tenant Data Center

Yukio Ogawa
Hitachi, Ltd., Japan

Go Hasegawa and Masayuki Murata
Osaka University, Japan

Contents

Research background

• A data center (DC) for the IaaS cloud computing
– serves virtual DC for multiple client organizations, i.e. tenants
– needs to host business-critical and mission-critical applications

• The virtual network (VN) for a tenant's virtual DC
– is an overlay network built by connecting VMs, based on VXLAN, etc
– has a topology independent of the physical substrate network (SN)
– should be appropriately assigned to the SN
to share the SN's resources effectively and tolerate SN failures

• Goal:
ensuring high availability for the VN so that mission critical applications can be hosted on it

Research objectives

• Mapping VNs to the shared physical SN is a kind of the *Virtual Network Embedding* problem

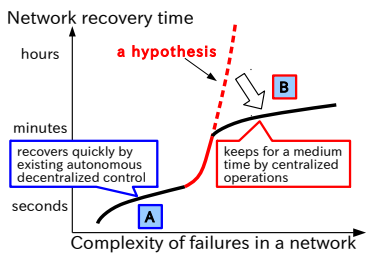
• Problems:
in a multi-tenant data center,
– nodes and links of VNs share a single component of the SN
– a failure of a single SN component can cause multiple simultaneous failures in a VN
– significantly disrupts the services offered on the VN, as compared to a traditional network

• Research objectives:
clarifying how the fault tolerance of a VN is affected by a SN failure, from the perspective of VN allocation

A hypothesis on the failure recovery time in a single VN

• A hypothesis: multiple simultaneous failures can lead to a longer recovery time in physical and virtual networks

• Proposal: switching from hot- to cold-standby recovery with reference to the failure complexity

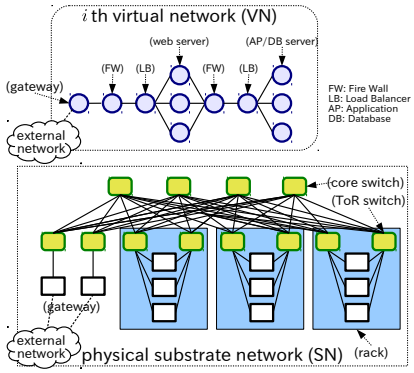


A Low complexity
• One or a few simultaneous failures
▷ VN recovers after a few seconds by switching to hot-standby nodes and links (VRRP, VMware FT, etc)

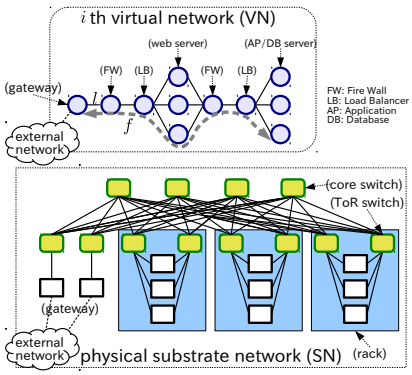
B High complexity
• Many simultaneous failures
▷ A centralized control force the failed nodes to be terminated and cold-standby nodes are alternatively booted (VMware HAT, etc)

Network model for a multi-tenant data center

Mapping a VN onto the SN



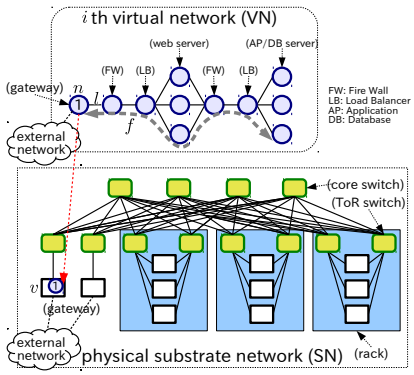
Network model for a multi-tenant data center



Mapping a VN onto the SN

• Traffic flow f is assigned to logical link l

Network model for a multi-tenant data center

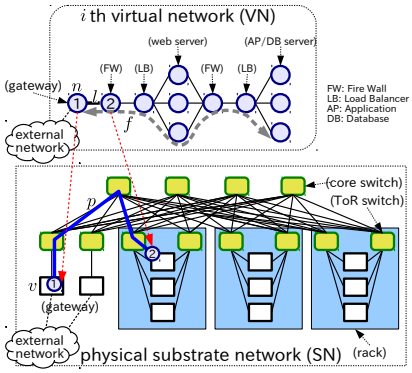


Mapping a VN onto the SN

• Traffic flow f is assigned to logical link l

• Logical node n is mapped onto physical server v

Network model for a multi-tenant data center



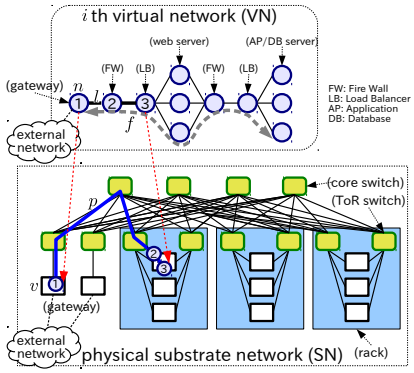
Mapping a VN onto the SN

• Traffic flow f is assigned to logical link l

• Logical node n is mapped onto physical server v

• Logical Link l is mapped onto physical path p

Network model for a multi-tenant data center



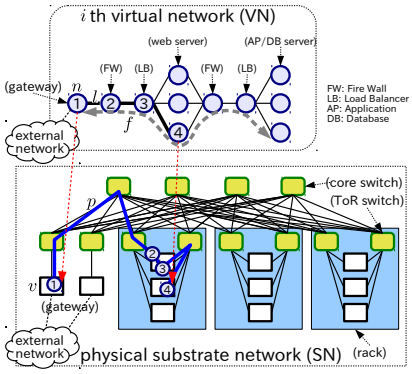
Mapping a VN onto the SN

• Traffic flow f is assigned to logical link l

• Logical node n is mapped onto physical server v

• Logical Link l is mapped onto physical path p

Network model for a multi-tenant data center



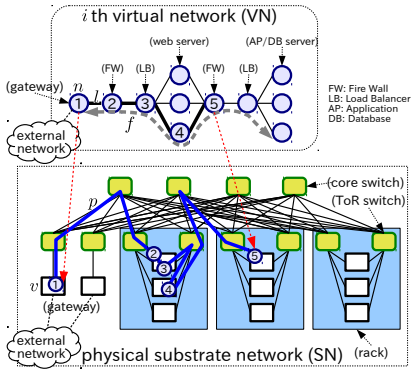
Mapping a VN onto the SN

• Traffic flow f is assigned to logical link l

• Logical node n is mapped onto physical server v

• Logical Link l is mapped onto physical path p

Network model for a multi-tenant data center



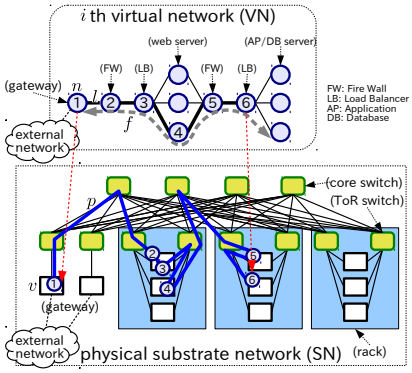
Mapping a VN onto the SN

• Traffic flow f is assigned to logical link l

• Logical node n is mapped onto physical server v

• Logical Link l is mapped onto physical path p

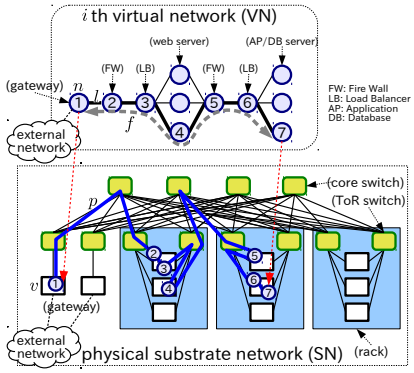
Network model for a multi-tenant data center



Mapping a VN onto the SN

- Traffic flow f is assigned to logical link l
- Logical node n is mapped onto physical server v
- Logical Link l is mapped onto physical path p

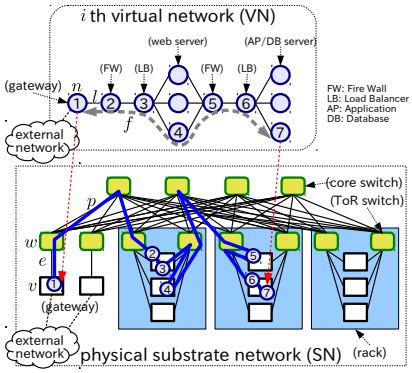
Network model for a multi-tenant data center



Mapping a VN onto the SN

- Traffic flow f is assigned to logical link l
- Logical node n is mapped onto physical server v
- Logical Link l is mapped onto physical path p

Network model for a multi-tenant data center



Mapping a VN onto the SN

- Traffic flow f is assigned to logical link l
- Logical node n is mapped onto physical server v
- Logical Link l is mapped onto physical path p
- Physical path p is mapped onto physical server v, physical switch w, and physical link e

Objective

Goal of VN allocation: minimizing the bandwidth loss when a failure happens in the SN

bandwidth loss of i th VN

that for a failure of physical server v

that for a failure of physical switch w

that for a failure of physical link e

Objective: minimize

$$\sum_{i\in I} B_i = \sum_{i\in I}\left(\sum_{v\in V} B_v^i + \sum_{w\in W} B_w^i + \sum_{e\in E} B_e^i\right)$$

failure rate of physical server v

ith VN's recovery time after a failure of physical server v

bandwidth of traffic flow f

$$B_v^i = D\,T_v^i \sum_{f\in F_i} X_{fv}^i c_f^i$$

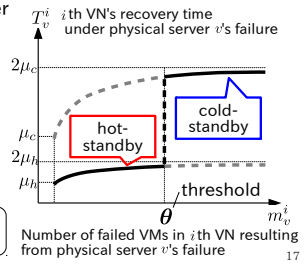
1: i th VN's traffic flow f is mapped to physical server v
0: otherwise

Recovery time model of a single VN

- A failure of a single physical switch/link
  - ▷ Recovery of the physical switch/link leads to recovery of the VN links
  - ▷ Recovery time of the VN does not influenced by how the VN is embedded in the physical switch/link
- A failure of a single physical server
  - ▷ VN should recover the VMs by utilizing its own failure-recovery mechanism
  - ▷ Recovery time of the VN depends on how complicated the VN becomes
    ≒ the number of multiple VMs failing simultaneously
    = the number of VMs assigned to the physical server
  - ▷ **Subject to:**

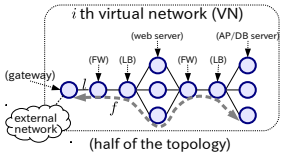
$$\sum_{n\in N_i} x_{nv}^i \le \theta$$

prohibiting assigning more than θ VMs for ensuring the VMs' hot-standby recovery



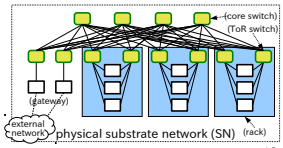
Data center network for evaluation

- A single VN
  - ▷ three-tier web serving architecture
  - ▷ 5.8 web and AP/DB servers, a total of 15.7 VMs on average.
  - ▷ CPU cores per VM: 1
  - ▷ average bandwidth demand from an external network: 1.7×10^8 bit/s
  - ▷ recovery time of a VM
    – hot-standby: 4 s, cold-standby: 60 s



(half of the topology)

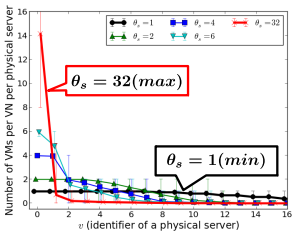
- The SN
  - ▷ two-level fat-tree topology
  - ▷ max configuration: 8 core switches, 16 ToR switches, and 120 physical servers
  - ▷ CPU cores/physical server: 32, bandwidth of each link: 1×10^{10} bit/s
  - ▷ available CPU cores: 3,360
  - ▷ failure rates– physical server: 4/year, physical link/switch: 0.05/year (neglected)



Overview of a single VN mapping

- VN embedding problem is *NP-hard* :
  initially – *Greedy Algorithm* , refined - *Tabu search*
- VN recovery time depends on θ (threshold for switching hot- to cold-standby), which can not be defined in advance
  - ▷ θ_s (a setting value of θ) is initially chosen
  - ▷ VN is allocated by using θ_s, then evaluated for various values of θ

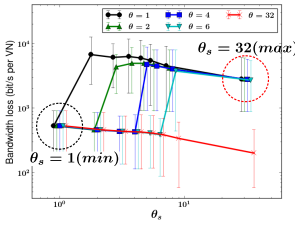
- θ_s determines the *shape* of the VN
  - ▷ $\theta_s = 1(min)$
  – The VMs and logical links are scattered across many physical servers and links
  - ▷ $\theta_s = 32(max)$
  – All the VMs and links are consolidated in a few physical servers

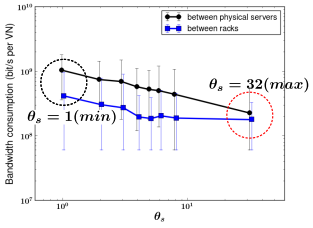


Trade-off between fault tolerance and physical bandwidth consumption

- $\theta_s = 1(min)$: one VM to one physical server mapping
  - ▷ The bandwidth loss is nearly the minimum for hot-standby recovery
  - ▷ The consumed bandwidth between servers/racks reaches the maximum

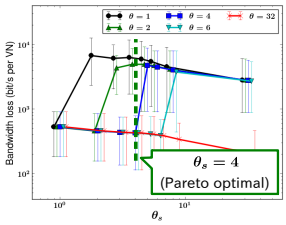
- $\theta_s = 32(max)$: many VMs to one physical server mapping
  - ▷ The bandwidth loss is nearly the maximum for cold-standby recovery
  - ▷ The consumed bandwidth between servers/racks becomes the minimum

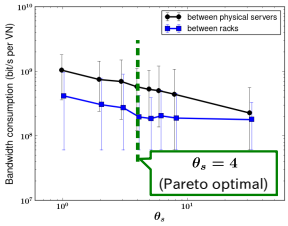




VN Allocation Policy Derived from the Results

- Minimizing the bandwidth loss of the VN while avoiding holding too many redundant core switches
- Pareto optimality: $\theta_s = 4$
  - ▷ Almost of the logical links were mapped onto the physical links between the physical servers and ToR switches.
  - ▷ The VN had almost no inter-rack traffic flows other than the one coming through the gateway





Conclusion

- The fault tolerance of each VN in an IaaS data center
  - ▷ Focusing on the situation of multiple simultaneous failures in each VN caused by a single physical failure
  - ▷ The trade-off between the bandwidth loss and the required bandwidth between physical servers
  - ▷ Balancing by assigning every four VMs to a physical server,
    – the required bandwidth of the outside racks was minimized

- Future work
  - ▷ Investigation of resource allocation over WANs, i.e., in a hybrid cloud environment