Virtual Network Embeddings in Data Center Networks

Soroush Haeri and Ljiljana Trajković

Communication Networks Laboratory
http://www.ensc.sfu.ca/~ljilja/cnl/
Simon Fraser University
Vancouver, British Columbia, Canada

Roadmap

- Network virtualization
- Virtual network embedding (VNE)
- Virtual network embedding, software defined networks, and data centers
- Data center topologies
- Simulation results
- Conclusions and references

Network virtualization

- Enables coexistence of multiple virtual networks on a physical infrastructure
- Virtualized network model divides the role of Internet Service Providers (ISPs) into:
  - Infrastructure Providers (InPs)
    - manage the physical infrastructure
  - Service Providers (SPs)
    - aggregate resources from multiple InP into multiple Virtual Networks (VNs)

Substrate network vs. virtual network

- InPs operate physical substrate networks (SNs)
- SN components:
  - physical nodes (substrate nodes)
  - physical links (substrate links)
- Substrate nodes and links are:
  - interconnected using arbitrary topology
  - used to host various virtualized networks with arbitrary topologies
- Virtual networks are embedded into a substrate network

Virtual network embedding

- Virtual Network Embedding (VNE) allocates SN resources to VNs
- InP's revenue depends on VNE efficiency
- VNE problem may be reduced to the multi-way separator:
  - NP-hard
  - optimal solution may only be obtained for small instances

- M. Yu, Y. Yi, J. Rexford, and M. Chiang, "Rethinking virtual network embedding: substrate support for path splitting and migration," *Comput. Commun. Rev.*, vol. 38, no. 2, pp. 19–29, Mar. 2008.

VNE solution

- Two subproblems:
  - Virtual Node Mapping (VNoM): maps virtual nodes to substrate nodes
  - Virtual Link Mapping (VLiM): maps virtual links to substrate paths
- VNE algorithms address the VNoM while solving the VLiM using:
  - Shortest-Path (SP) algorithms

  or
  - Multicommodity Flow (MCF) algorithm

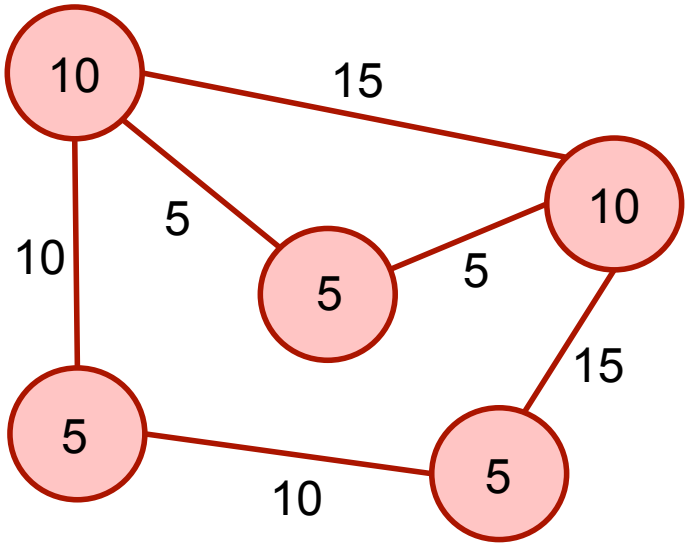
VNE solution: VLiM and path splitting

- The shortest-path algorithms do not permit path splitting:
  - stricter than the MCF algorithm
- MCF enables path splitting:
  - a flow may be divided into multiple flows with lower capacity
  - flows are routed through various paths

- D. G. Andersen, “Theoretical approaches to node assignment,” Dec. 2002, Unpublished Manuscript. [Online]. Available: http://repository.cmu.edu/compsci/86/.

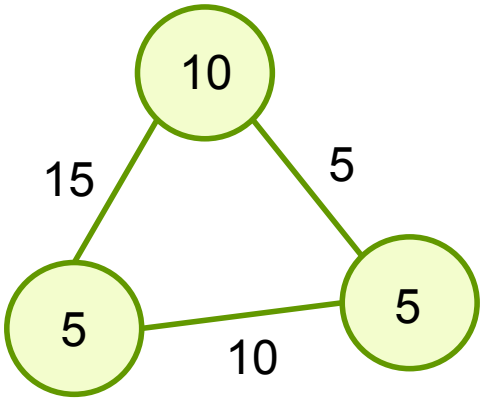
VNE formulation: constrains

- Substrate network graph: $G^s(N^s, E^s)$
- Resources:
  - substrate nodes: CPU capacity $\mathcal{C}(n^s)$
  - substrate links: bandwidth $\mathcal{B}(e^s)$



VNE formulation: constrains

- Virtual network graph: $G^{\Psi_i}(N^{\Psi_i}, E^{\Psi_i})$
- Resources:
  - virtual nodes: CPU capacity $\mathcal{C}(n^{\Psi_i})$
  - virtual links: bandwidth $\mathcal{B}(e^{\Psi_i})$



VNE objective

- Maximize the profit of InPs
- Contributing factors to the generated profit:
  - embedding revenue
  - embedding cost
  - acceptance ratio

- M. Chowdhury, M. R. Rahman, and R. Boutaba, "ViNEYard: Virtual network embedding algorithms with coordinated node and link mapping," *IEEE/ACM Trans. Netw.*, vol. 20, no. 1, pp. 206–219, Feb. 2012.
- L. Gong, Y. Wen, Z. Zhu, and T. Lee, "Toward profit-seeking virtual network embedding algorithm via global resource capacity," in *Proc. IEEE INFOCOM*, Toronto, ON, Canada, Apr. 2014, pp. 1–9.

VNE objective: revenue

- Maximize revenue:

$$\mathbf{R}(G^{\Psi_i}) = w_c \sum_{n^{\Psi_i} \in N^{\Psi_i}} \mathcal{C}(n^{\Psi_i}) + w_b \sum_{e^{\Psi_i} \in E^{\Psi_i}} \mathcal{B}(e^{\Psi_i})$$

  - w_c : weights for CPU requirements
  - w_b : weight for bandwidth requirements
  - general assumption: $w_c = w_b = 1$

VNE objective: revenue

- Generated revenue is not a function of the embedding configuration:
  - $\mathbf{R}(G^{\Psi_i})$ is constant regardless of the embedding configuration

VNE objective: cost

- Minimize the cost:

$$\mathbf{C}(G^{\Psi_i}) = \sum_{n^{\Psi_i} \in N^{\Psi_i}} \mathcal{C}(n^{\Psi_i}) + \sum_{e^{\Psi_i} \in E^{\Psi_i}} \sum_{e^s \in E^s} f_{e^s}^{e^{\Psi_i}}$$

  - $f_{e^s}^{e^{\Psi_i}}$: total allocated bandwidth of the substrate link e^s for virtual link e^{Ψ_i}
- $\mathbf{C}(G^{\Psi_i})$ depends on the embedding configuration

VNE objective: acceptance ratio

- Maximize acceptance ratio:

$$p_a^{\tau} = \frac{|\Psi^a(\tau)|}{|\Psi(\tau)|}$$

  - $|\Psi^a(\tau)|$: number of accepted Virtual Network Requests (VNRs) in a given time interval τ
  - $|\Psi(\tau)|$: number of all arrived VNRs in τ

VNE objective function

- Objective of embedding a VNR is to maximize:

$$\mathcal{F}(\Psi_i) = \begin{cases} \mathbf{R}(G^{\Psi_i}) - \mathbf{C}(G^{\Psi_i}) & \text{successful embeddings} \\ \Gamma & \text{otherwise} \end{cases}$$

  - Γ : large negative penalty for unsuccessful embedding
- The upper bound:

$$\mathcal{F}(\Psi_i) \leq 0$$

VNE algorithms: R-Vine and D-Vine

- Formulate VNE problem as a Mixed Integer Program (MIP)
- Their objective is to minimize the cost of accommodating the VNRs
- Use a rounding-based approach to obtain a linear programming relaxation of the relevant MIP
- Use Multicommodity Flow algorithm for solving VLiM

- M. Chowdhury, M. R. Rahman, and R. Boutaba, "ViNEYard: Virtual network embedding algorithms with coordinated node and link mapping," *IEEE/ACM Trans. Netw.*, vol. 20, no. 1, pp. 206–219, Feb. 2012.

VNE algorithms: Global Resource Capacity (GRC)

- Node-ranking-based algorithm:
  - computes a score/rank for substrate and virtual nodes
  - employs a large-to-large and small-to-small mapping scheme to map the virtual nodes to substrate nodes
- Employs the Shortest-Path algorithm to solve VLiM
- Outperforms earlier similar proposals

- L. Gong, Y. Wen, Z. Zhu, and T. Lee, "Toward profit-seeking virtual network embedding algorithm via global resource capacity," in *Proc. IEEE INFOCOM*, Toronto, ON, Canada, Apr. 2014, pp. 1–9.

VNE algorithms: Global Resource Capacity (GRC)

- Calculates the embedding capacity $r(n_i^s)$ for a substrate node n_i^s:

$$r(n_i^s) = (1-d)\hat{\mathcal{C}}(n_i^s) + d \sum_{n_j^s \in \mathcal{N}(n_i^s)} \frac{\mathcal{B}\left(e^s(n_i^s, n_j^s)\right)}{\sum_{n_k^s \in \mathcal{N}(n_j^s)} \mathcal{B}\left(e^s(n_j^s, n_k^s)\right)}$$

  - $0 < d < 1$: damping factor
  - $e^s(n_i^s, n_j^s)$: substrate link connecting n_i^s and n_j^s
  - $\hat{\mathcal{C}}(n_i^s)$: normalized CPU resource of n_i^s

$$\hat{\mathcal{C}}(n_i^s) = \frac{\mathcal{C}(n_i^s)}{\sum_{n^s \in N^s} \mathcal{C}(n^s)}$$

VNE algorithms: Global Resource Capacity (GRC)

- Matrix form:

$$\mathbf{r} = (1-d)\hat{\mathbf{c}} + d\mathbf{M}\mathbf{r}$$

  - $\hat{\mathbf{c}} = (\hat{\mathcal{C}}(n_1^s), \hat{\mathcal{C}}(n_2^s), \ldots, \hat{\mathcal{C}}(n_j^s))^T$
  - $\mathbf{r} = (r(n_1^s), r(n_2^s), \ldots, r(n_k^s))^T$
  - $\mathbf{M}$ is a k-by-k matrix:

$$m_{ij} = \begin{cases} \dfrac{\mathcal{B}\left(e^s(n_i^s, n_j^s)\right)}{\sum\limits_{n_k^s \in \mathcal{N}(n_j^s)} \mathcal{B}\left(e^s(n_j^s, n_k^s)\right)} & e^s(n_i^s, n_j^s) \in E^s \\ 0 & \text{otherwise} \end{cases}$$

VNE algorithms: Global Resource Capacity (GRC)

- $\mathbf{r}$ is calculated iteratively:

$$\mathbf{r}_{k+1} = (1-d)\mathbf{c} + d\mathbf{M}\mathbf{r}_k$$

- Initially: $\mathbf{r}_0 = \hat{\mathbf{c}}$
- Stop condition: $|\mathbf{r}_{k+1} - \mathbf{r}_k| < \sigma,$
  - $0 < \sigma << 1$

SDN and network virtualization

- Software-Defined Networking (SDN):
  - separates network intelligence from network devices
  - enables central implementation of network control logic
- SDN may enable cloud providers such as the Amazon Web Services to offer network virtualization services:
  - requires embedding of the virtual networks in data center networks

VNE in data center networks

- Data center networks have defined topologies:
  - examples of two recent proposals: BCube and Fat-Tree
- Topological features significantly affect quality of the VNE solution
- Goal: identify the network topology that is better suited for VNE

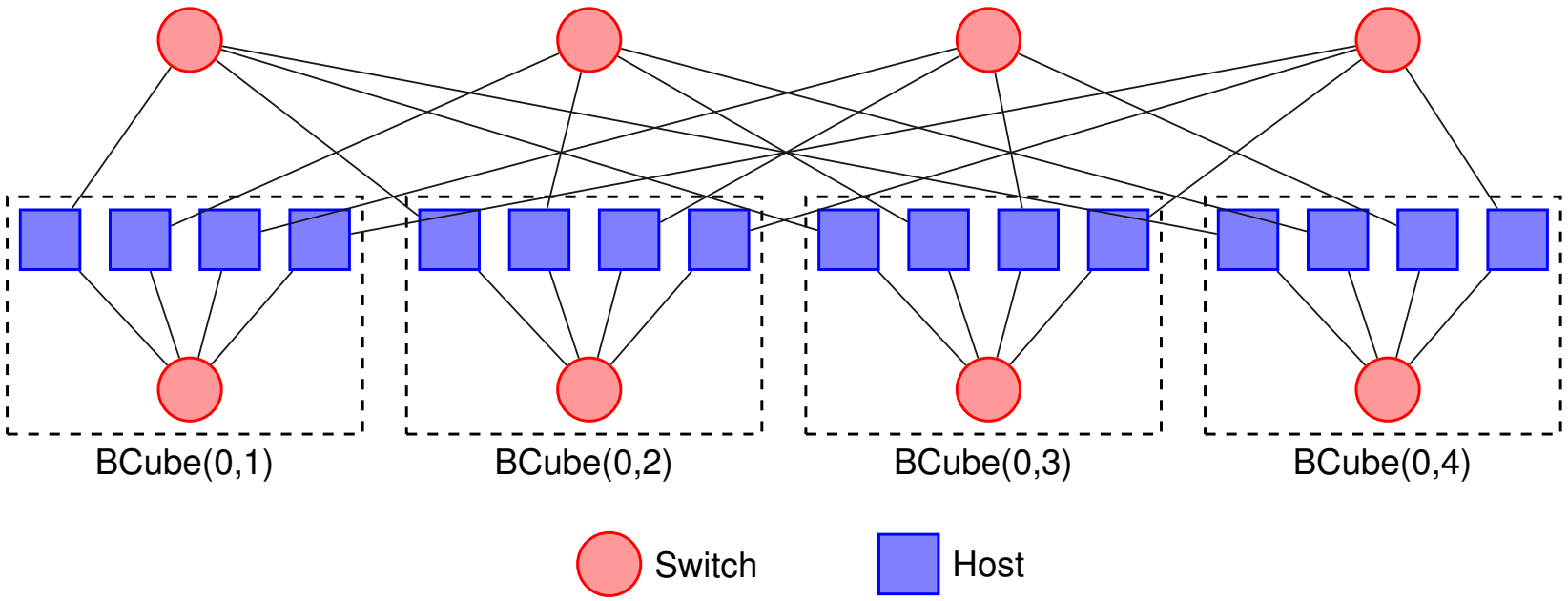
Data center topologies: BCube

- Notation: $\mathrm{BCube}(k, n)$
  - k : BCube level
  - n : number of hosts in the level-0 BCube
- Recursively structured
- Switches are not directly interconnected
- Hosts perform packet forwarding functions

- C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu, "BCube: A high performance, server-centric network architecture for modular data centers," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 63–74, Oct. 2009.

Data center topologies: BCube

- Example: $\mathrm{BCube}(2, 4)$



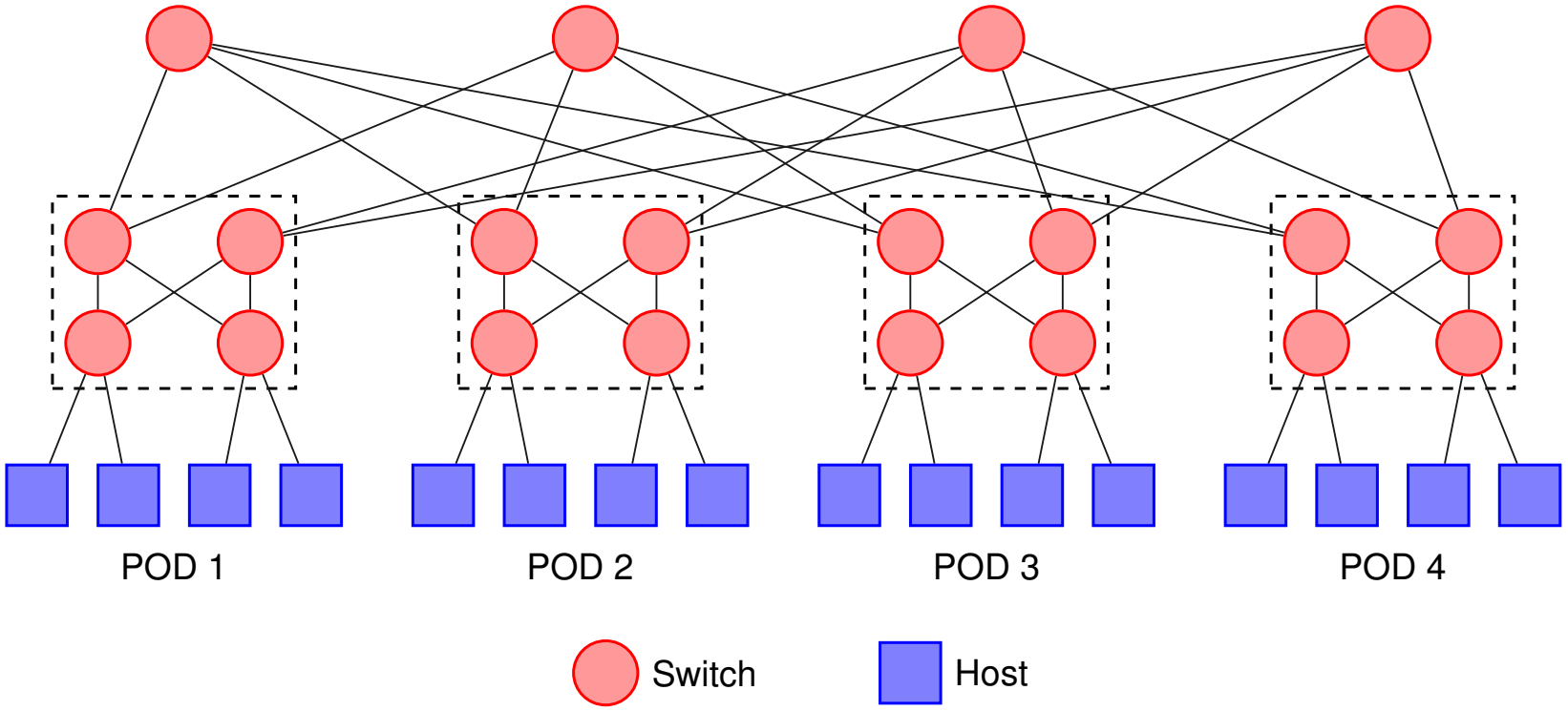
Data center topologies: Fat-Tree

- Notation: Fat-Tree_k
- Special Clos architecture
- Initially proposed to interconnect processors of parallel supercomputers
- $(k/2)^2 + k^2$ k-port switches
- Supports $k^3/4$ hosts

- C. E. Leiserson, "Fat-Trees: universal networks for hardware-efficient supercomputing." *IEEE Trans. Comput.*, vol. 30, no. 10, pp. 892–901, Oct. 1985.
- M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, pp. 63–74, Oct. 2008.

Data center topologies: Fat-Tree

- Example: $\mathrm{Fat\text{-}Tree}_4$



Traffic

BCube:

- hosts are used to forward traffic
  - introduces additional traffic over the links that are connected to the hosts

Fat-Tree:

- traffic forwarding is only performed by the switches

Simulations: substrate networks

- $\mathrm{BCube}(2,4)$: 64 hosts, 48 switches, and 192 link
  - Switch to host ratio: 0.75
- $\mathrm{Fat\text{-}Tree}_6$: 54 hosts, 45 switches, and 162 links
  - Switch to host ratio: 0.84
- CPU resources:
  - Hosts: 100 units
  - Switches: 0 units
- Bandwidth resources: 100 units per link

Simulations: virtual network graphs

- Waxman algorithm used to generate virtual network graphs:
  - $\alpha = 0.5$ and $\beta = 0.2$
  - number of nodes: uniformly distributed between 3 and 10
  - each virtual node: connected to a maximum of 3 virtual nodes

Simulations: virtual network graphs

- CPU requirements:
  - uniformly distributed between 2 and 20 units
- Bandwidth requirements:
  - uniformly distributed between 1 to 10 units
  - illustrates a substrate network with 10 Gbps links and virtual networks with 100 Mbps to 1 Gbps links

Simulations: other parameters

- Poisson distribution for arrivals with implies 1 to 8 units per 100 time units
- Exponentially distributed life-times with mean 1,000 time units
- Traffic loads: 10, 20, 30, 40, 50, 60, 70, and 80 Erlangs
- Total simulation time: 50,000 time units
- Performance metrics:
  - acceptance ratio, revenue to cost ratio, and substrate resource utilization

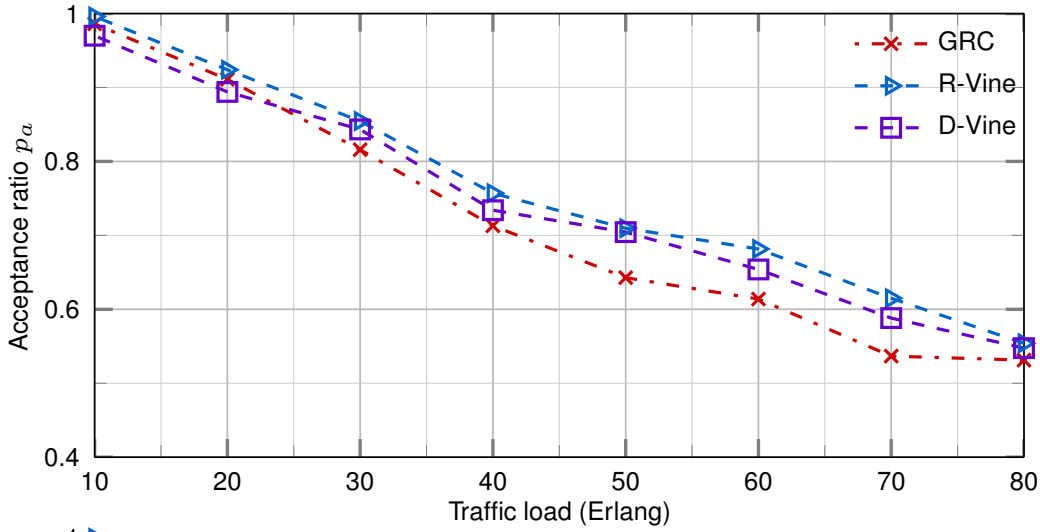
VNE-Sim

- A discrete event VNE simulator written in C++
- Based on the Discrete Event System Specification (DEVS) framework
- Employs the Adevs library

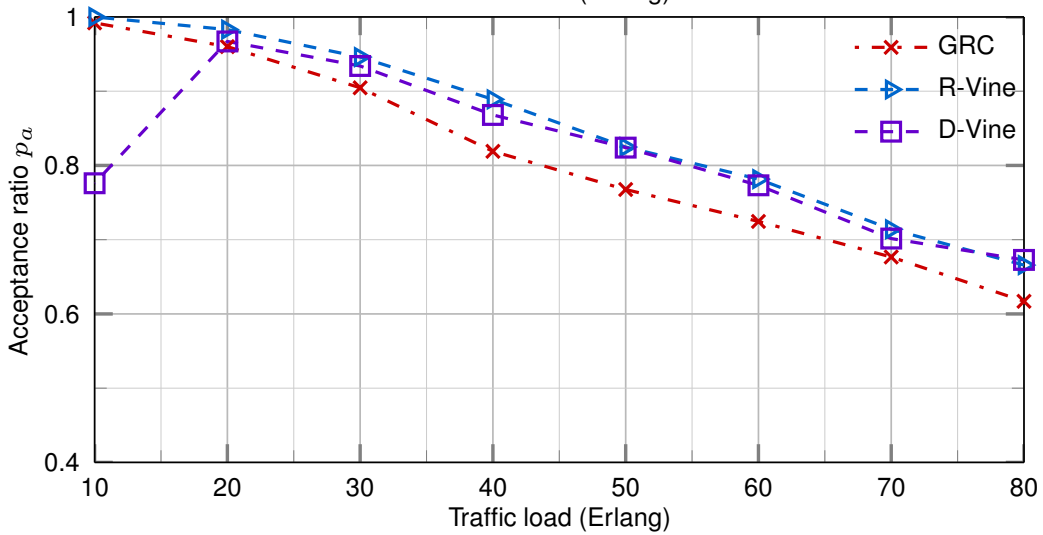
- A. M. Uhrmacher, "Dynamic structures in modeling and simulation: a reflective approach," *ACM Trans. Modeling and Computer Simulation*, vol. 11, no. 2, pp. 206–232, Apr. 2001.
- J. J. Nutaro, *Building Software for Simulation: Theory and Algorithms, with Applications in C++*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2010.

Acceptance ratio

BCube

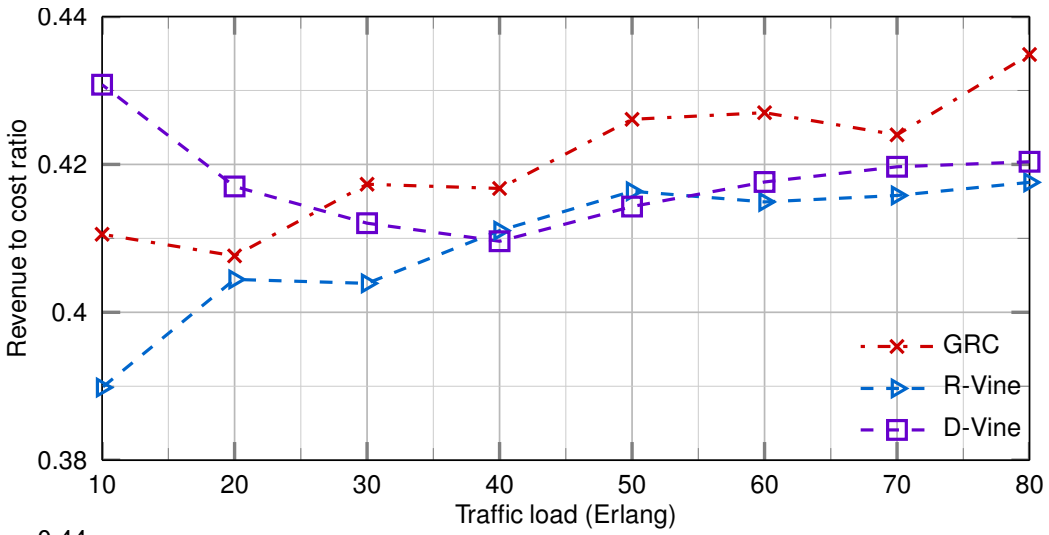


Fat-Tree

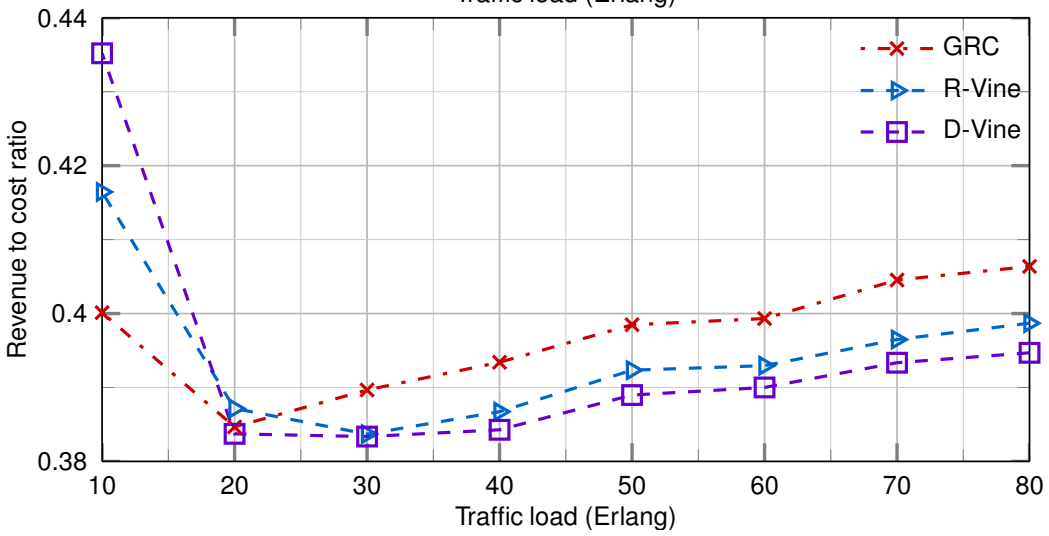


Revenue to cost ratio

BCube

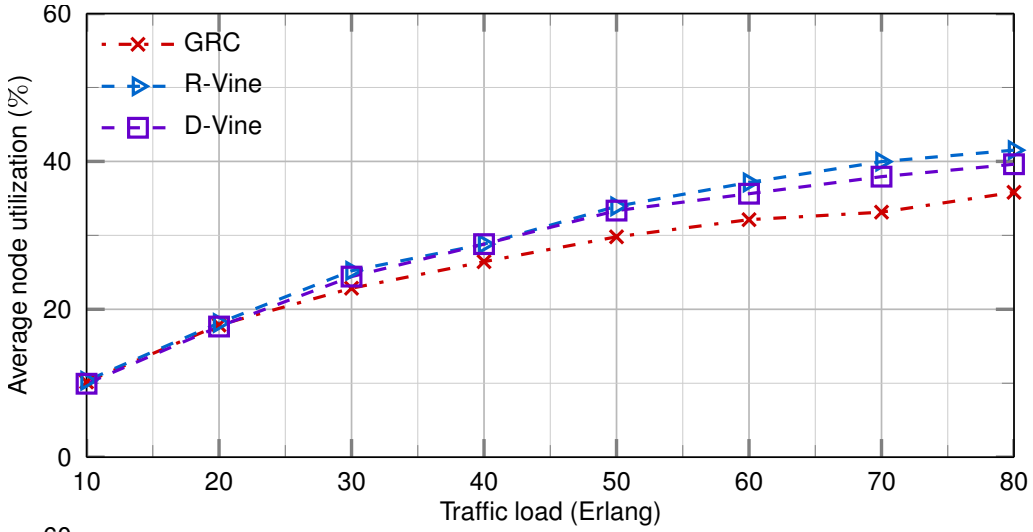


Fat-Tree

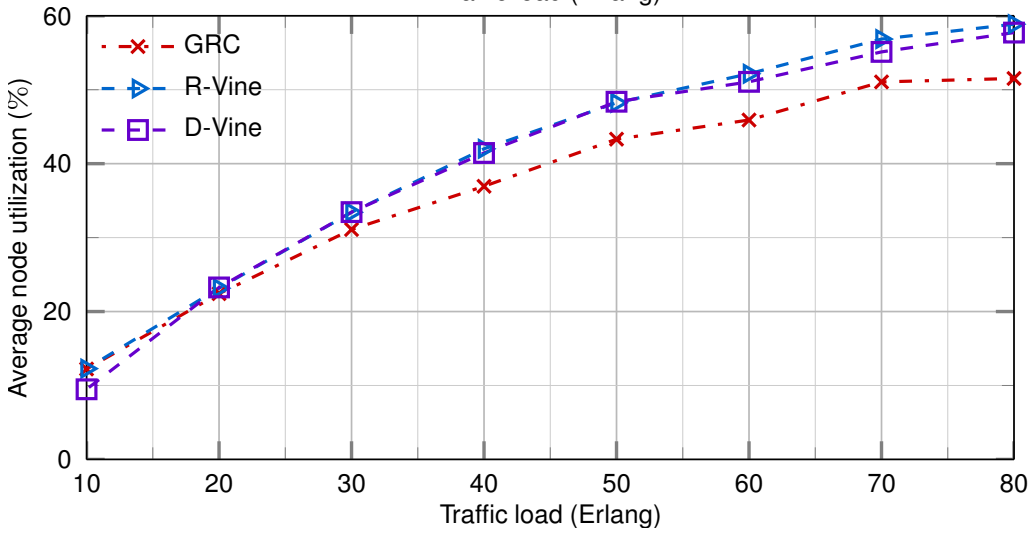


Average node utilization

BCube

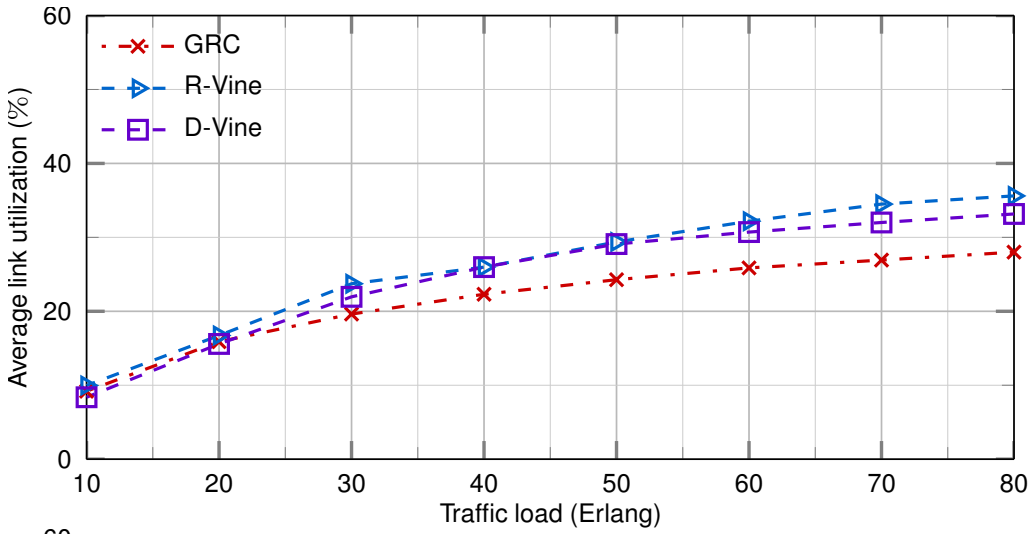


Fat-Tree

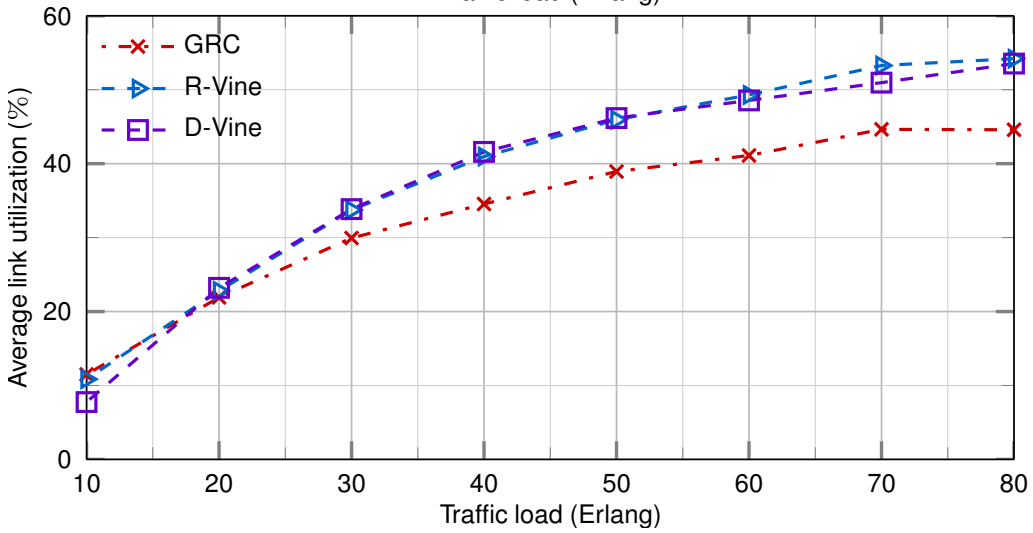


Average link utilization

BCube



Fat-Tree



Simulation results

- Fat-Tree topology offers up to:
  - 10% higher acceptance ratio
  - 20% higher node utilization
  - 10% higher link utilization
- The revenue to cost ratio of the Fat-Tree topology is slightly lower than the BCube topology
- Desirable:
  - high acceptance ratio, high substrate resource utilization, and high revenue to cost ratio

Conclusions

- Links that are connected to the hosts are important for the virtual network embeddings:
  - especially for embedding virtual nodes that require multiple connections to other nodes
- Performing traffic forwarding using only the core switches instead of the hosts may lead to higher VNR acceptance ratio

Conclusions

Simulated Fat-Tree topology:

- Has higher switch to host ratio (0.84) compared to the BCube topology (0.75)
- Additional paths between the hosts enable:
  - higher acceptance ratio
  - higher resource utilization
- Tradeoff:
  - lower revenue to cost ratio

Current project

- Model VNE as a decision-making problem using the Markov Decision Process (MDP) framework
- Solve the MDP to find the best action policy for embedding virtual networks
- Improve performance of GRC by employing MCF instead of the Shortest-Path algorithm to solve VLiM

References

Network Virtualization:

- M. Chowdhury and R. Boutaba, "Network virtualization: state of the art and research challenges," *IEEE Commun. Mag.*, vol. 47, no. 7, pp. 20–26, July 2009.
- N. Feamster, L. Gao, and J. Rexford, "How to lease the Internet in your spare time," *Comput. Commun. Rev.*, vol. 37, no. 1, pp. 61–64, Jan. 2007.

Data Center Networks:

- C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu, "BCube: A high performance, server-centric network architecture for modular data centers," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 63–74, Oct. 2009.
- M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, pp. 63–74, Oct. 2008.

References

- C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, "DCell: a scalable and fault-tolerant network structure for data centers," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, pp. 75–86, Oct. 2008.
- C. Guo, G. Lu, H. J. Wang, S. Yang, C. Kong, P. Sun, W. Wu, and Y. Zhang, "SecondNet: a data center network virtualization architecture with bandwidth guarantees," in *Proc. ACM CoNEXT 2010*, Philadelphia, PA, USA, Dec. 2010, p. 15.
- H. Ballani, P. Costa, T. Karagiannis, and A. Rowstron, "Towards predictable datacenter networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 242–253, Oct. 2011.
- C. E. Leiserson, "Fat-Trees: universal networks for hardware-efficient supercomputing." *IEEE Trans. Comput.*, vol. 30, no. 10, pp. 892–901, Oct. 1985.

References

Virtual Network Embedding Algorithms:

- L. Gong, Y. Wen, Z. Zhu, and T. Lee, "Toward profit-seeking virtual network embedding algorithm via global resource capacity," in *Proc. IEEE INFOCOM*, Toronto, ON, Canada, Apr. 2014, pp. 1–9.
- M. Chowdhury, M. R. Rahman, and R. Boutaba, "ViNEYard: Virtual network embedding algorithms with coordinated node and link mapping," *IEEE/ACM Trans. Netw.*, vol. 20, no. 1, pp. 206–219, Feb. 2012.
- S. Zhang, Y. Qian, J. Wu, and S. Lu, "An opportunistic resource sharing and topology-aware mapping framework for virtual networks," in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 2408–2416.
- X. Cheng, S. Su, Z. Zhang, H. Wang, F. Yang, Y. Luo, and J. Wang, "Virtual network embedding through topology-aware node ranking," *Comput. Commun. Rev.*, vol. 41, pp. 38–47, Apr. 2011.
- M. Yu, Y. Yi, J. Rexford, and M. Chiang, "Rethinking virtual network embedding: substrate support for path splitting and migration," *SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 19–29, Mar. 2008.