Capacity of Inter-Cloud Layer-2 Virtual Networking

Yufeng Xin, Ilya Baldin, Chris Heermann, Anirban Mandal, and Paul Ruth

Renci, University of North Carolina at Chapel Hill, NC, USA

yxin@renci.org

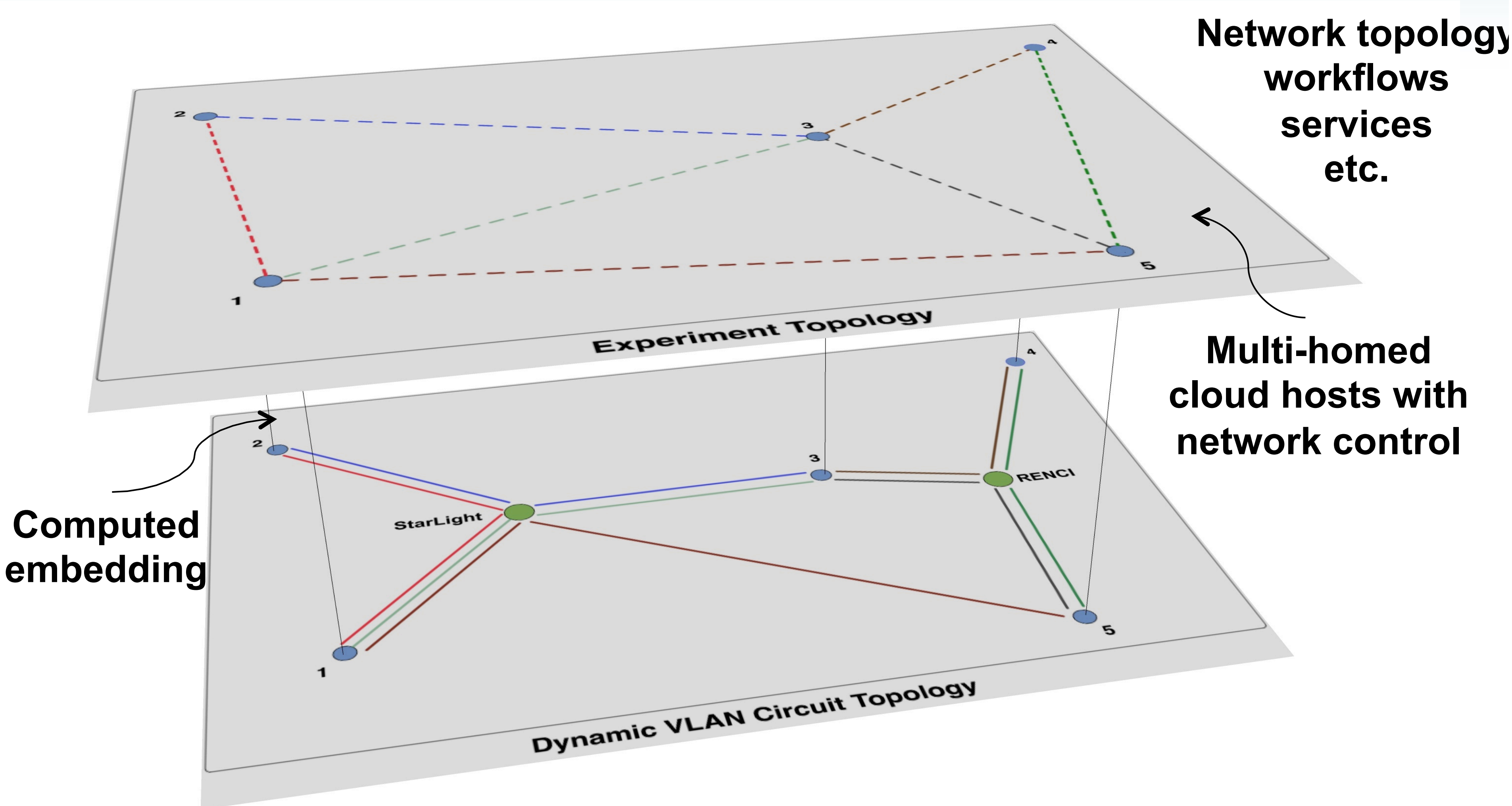
renci

RESEARCH \ ENGAGEMENT \ INNOVATION

Overview

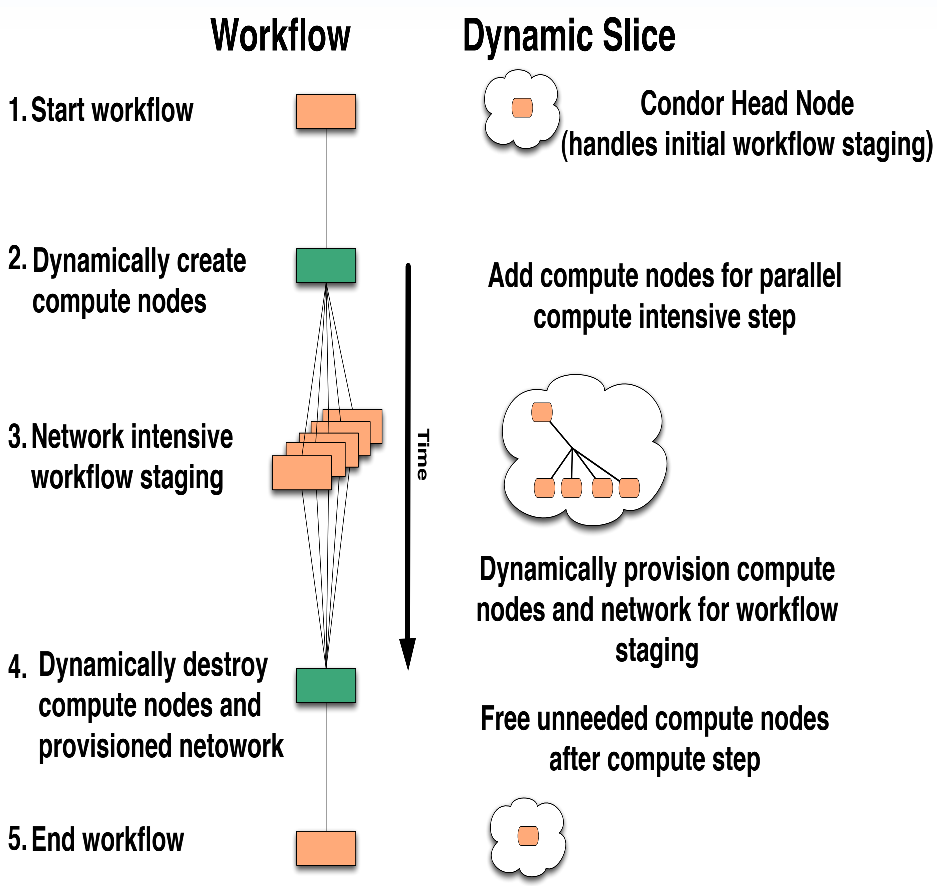
- Introduction and motivation
  - Distributed Cloud IaaS : Economy of Scale
  - Applications: high-end, HPC
  - Inter-Cloud Virtual Networking : Multi-domain, wide-area
- Inter-cloud layer-2 networking
  - Inter-domain VLAN connection
  - Point-to-point and multi-point connections
- Capacity Model
  - Maximal Number of connections
  - Model: complete multipartite graph
  - Static and dynamic capacity
- Conclusion

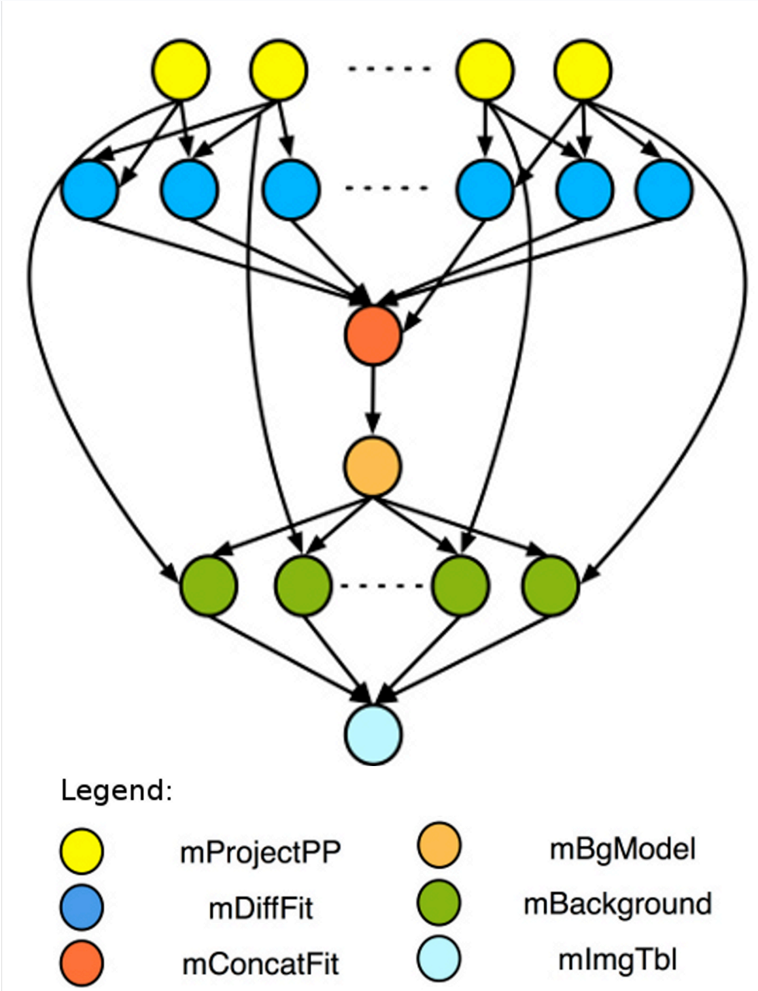
Virtual System Embedding



renci

Virtual HPC, Condor, Workflow, etc





Montage workflow

Virtual Networking (1)

- Multiple VM interfaces
  - Management plane: Internet for reachability
  - Data plane: virtual system networking -> isolation, QoS
- VM and data center networking
  - Layer 3 tunneling: GRE
  - Layer 2 emulation: VXLAN
  - Layer 2 VLAN
- Wide area networks connecting distributed clouds: multi-domain network environment
  - IP tunneling: low performance
  - MPLS: complex and expensive
  - VLAN connections
  - Layer-1 optical path

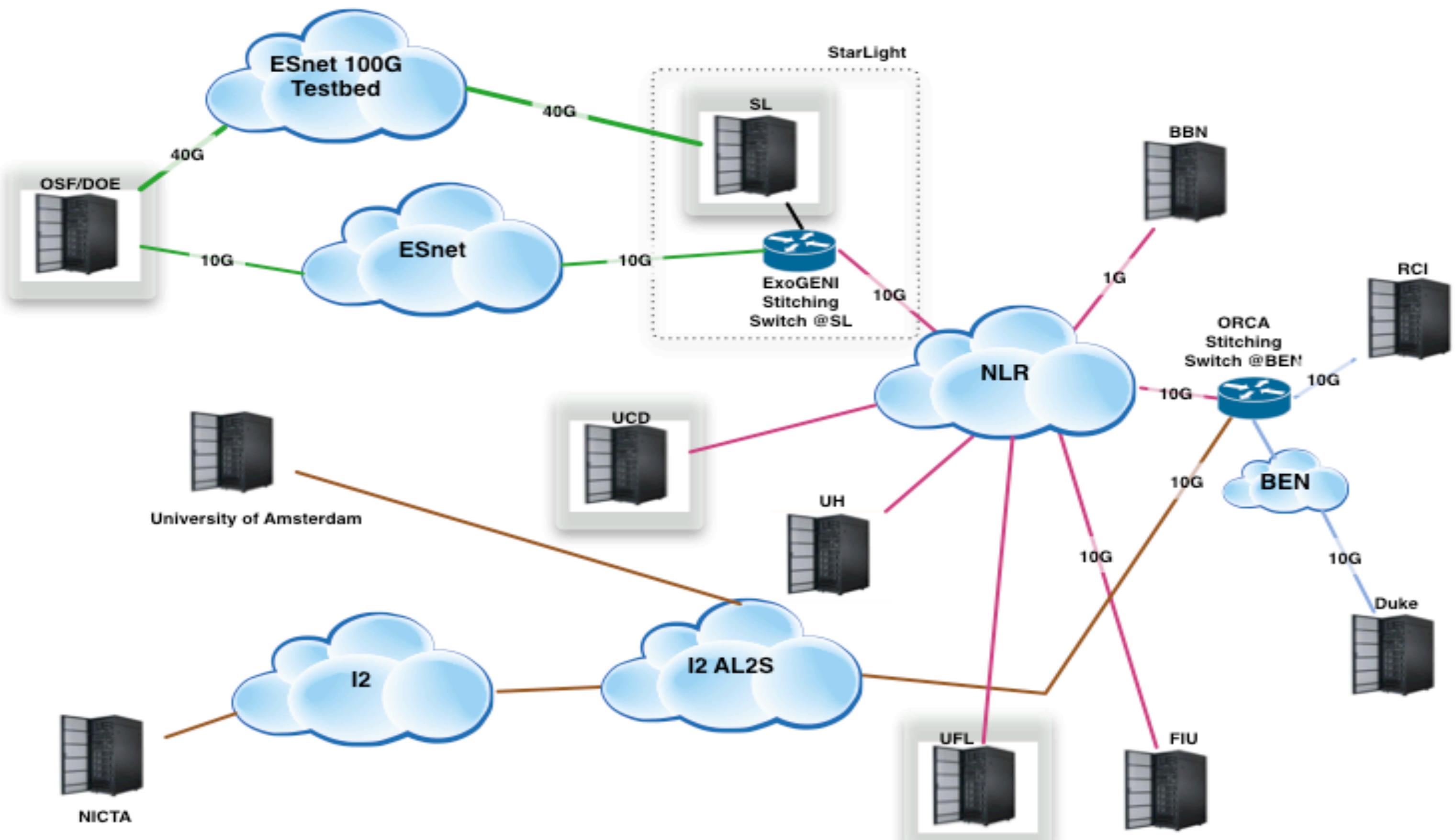
renci

Virtual Networking (2)

- Mechanism
  - Label (tag) for communications channel isolation and identity : IP address, MPLS labels, vlan, lambda,
  - Bandwidth control: orthogonal to label control
- Layer-2:
  - Cheap, QoS, everywhere
  - Carrier Ethernet
  - Dynamic circuits : PNNI, GMPLS, OSCARS, NSI, Stitching
  - Does it scale??

renci

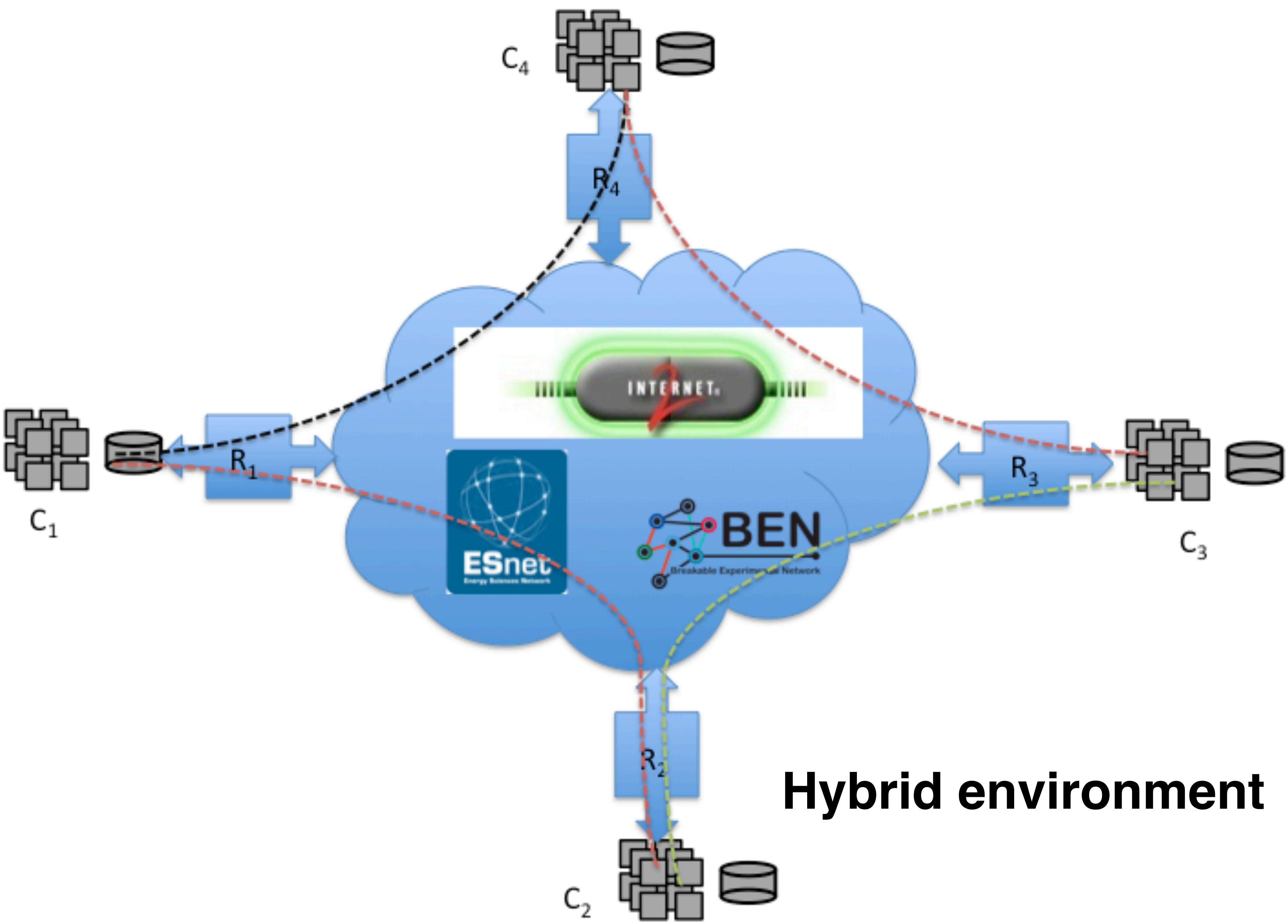
Laye-2 based Distributed Cloud: a rosy picture



renci

The reality : constraints

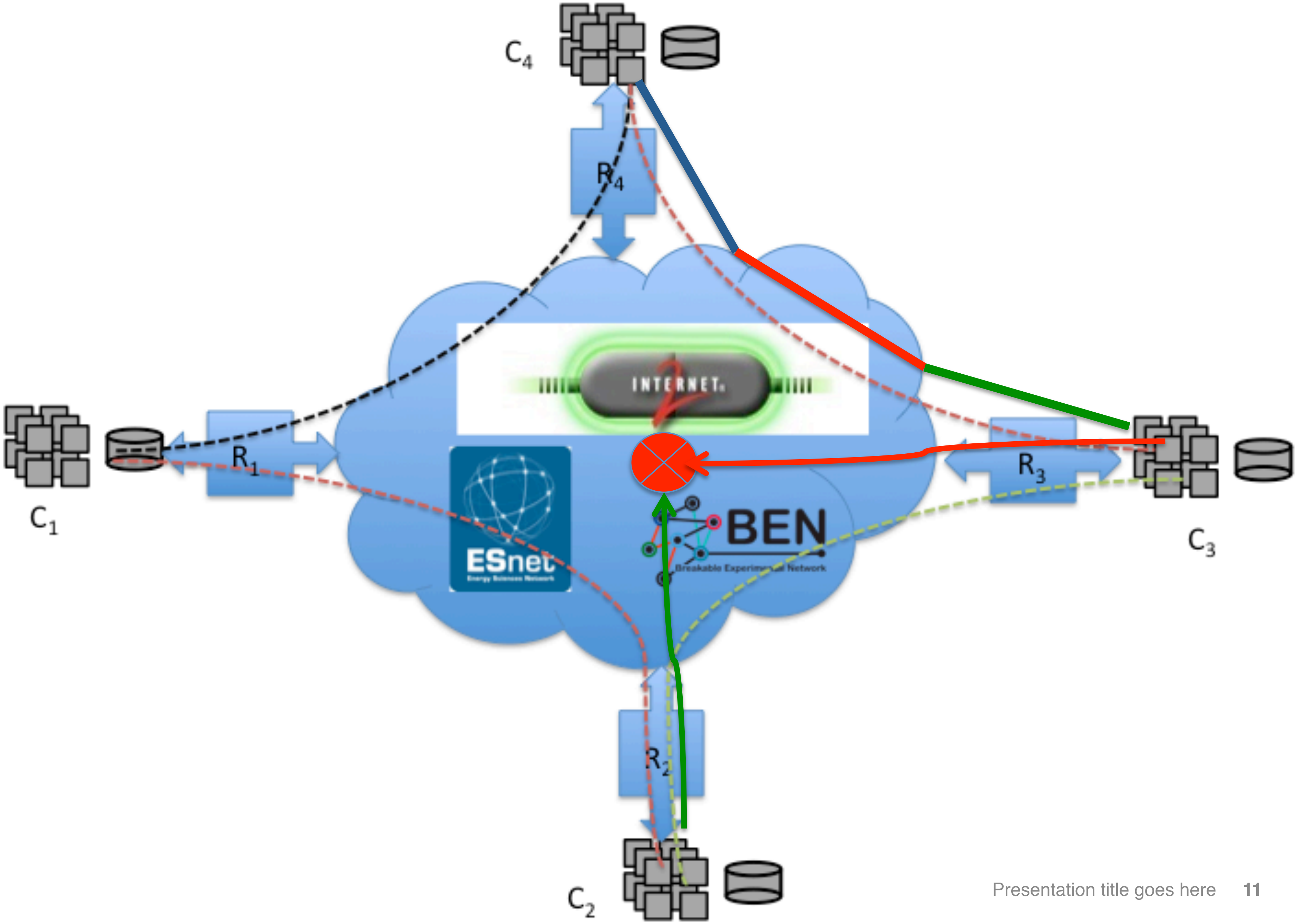
- Label continuity: label locality *vs* global
- Limited label space : 4096 vlans
- Dynamic label path provisioning is not widely deployed : End-to-end automation is difficult
  - ESNet and I2 (OSCARs)
  - NSI (GLIF)
- No multi-point connection



Hybrid environment

The reality : it is hard and not efficient

- Challenge:
  - Static routing and tag assignment with tag continuity constraint is NP-Hard
  - Tag continuity causes low utilization
  - Provisioning process is painful and could be long
- Solutions : dynamic stitching
  - Label translation
  - Label tunneling
  - Label exchange
  - End point location neutrality : virtual system


C4
R4
R1
C1
INTERNET2
ESnet
Energy Sciences Network
BEN
Breakable Experimental Network
R3
C3
R2
C2

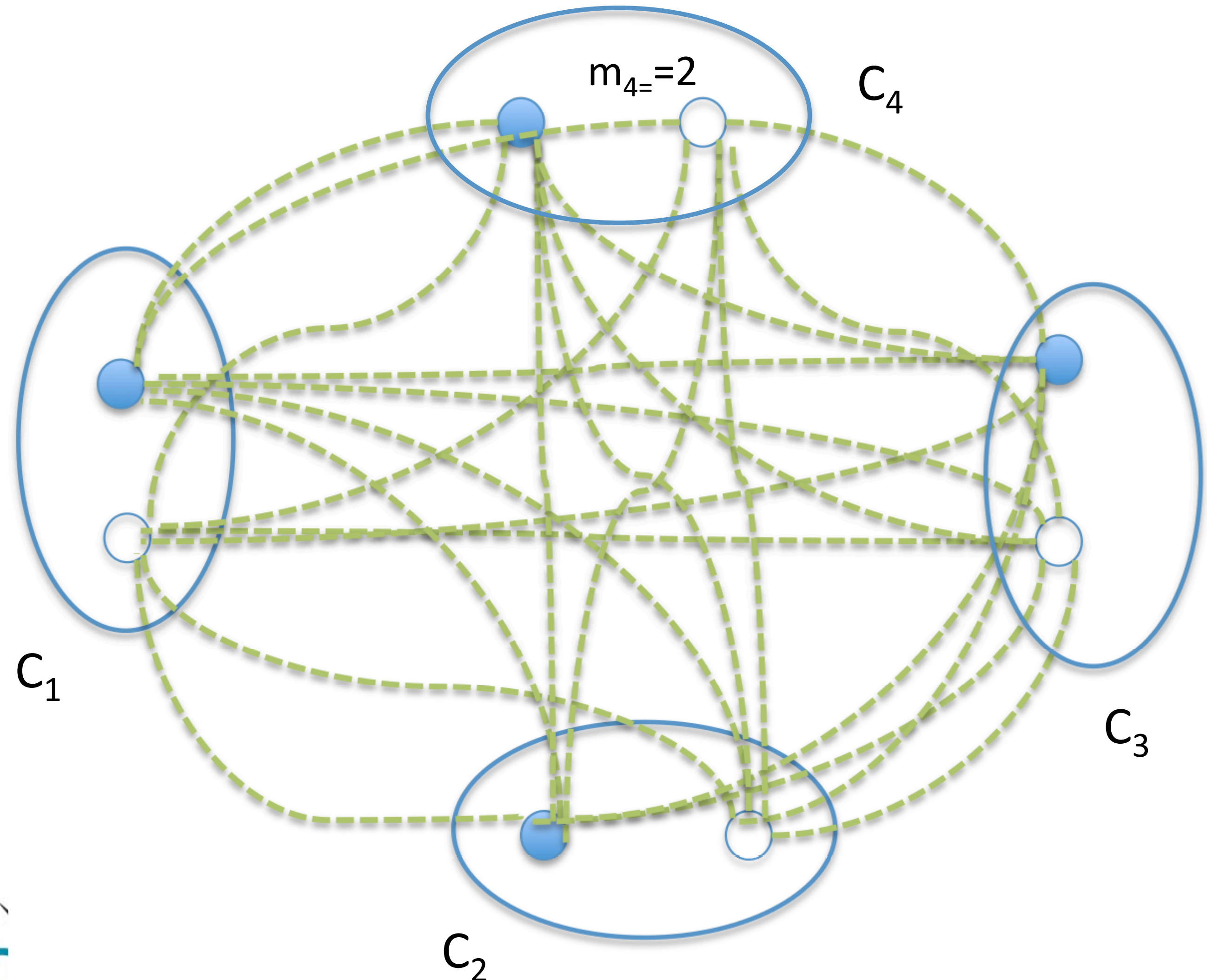
Static Capacity

- Still vlan tags are scarce commodity in many networks : 10 vlans out of most Exogeni rack sites now
- often the vlan tags are exhausted before the bandwidth is consumed
- Inter-cloud network capacity (Static)
  – maximum number of concurrent inter-cloud connections in the system

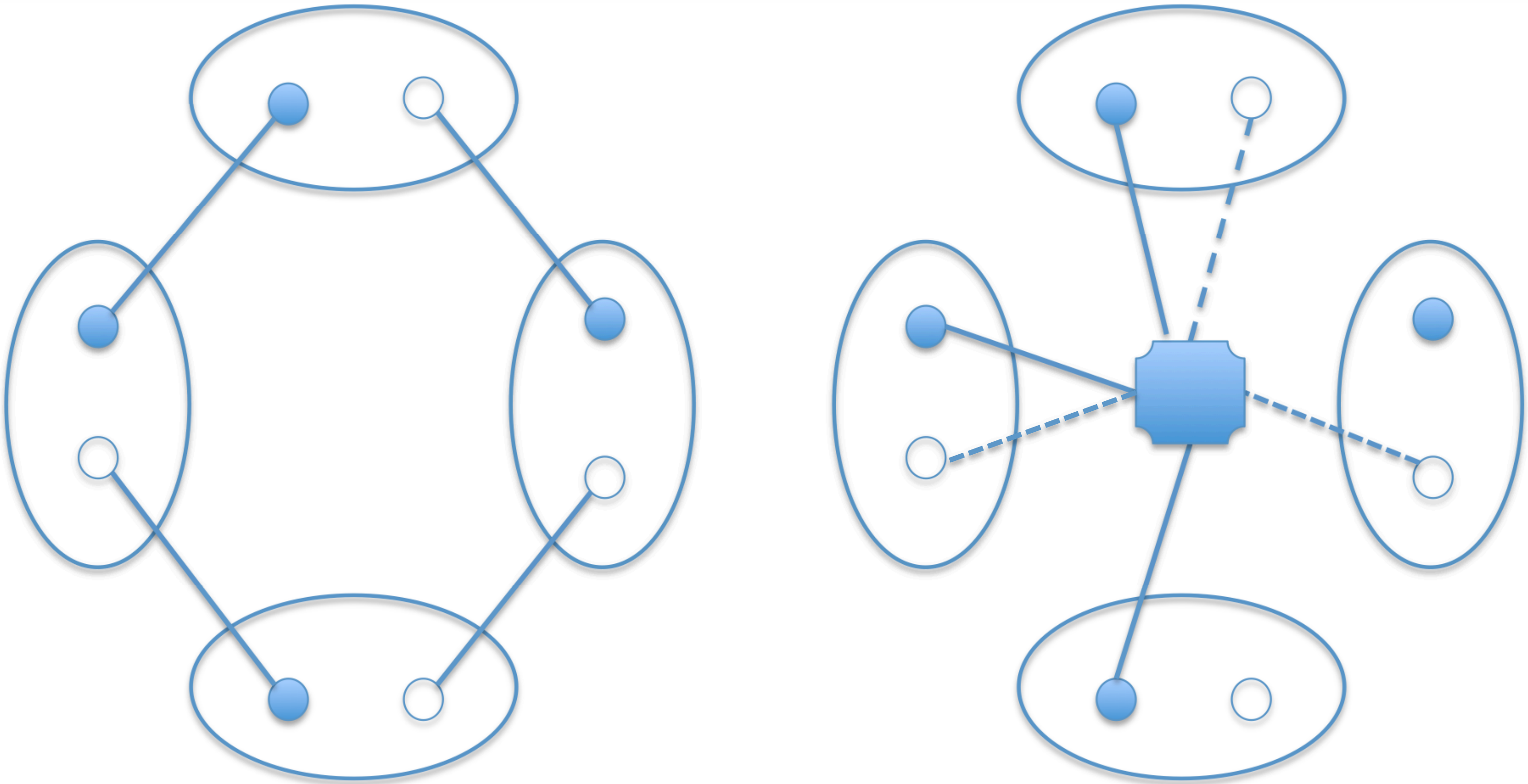
Capacity graph model

- Complete n-partite graph.
  - n cloud sites
  - site C_i , its regional network R_i, M_i pre-provioned vlan, $i \in \{1 \dots n\}$, connects to the backbone networks
  - Backbone networks have “unlimited” vlans
  - Edge $e=(v_x,v_j) \in E$, $v_x \in M_i, v_y \in M_j, i \neq j, \forall i, j$

Complete multipartite graph



Maximum Matching: set of pairwise vertex disjoint edges



(a) Point-to-Point Connection (b) 3-Point Connection

renci

Point-to-point connections

- **Theorem 1** The maximum number of inter-cloud point- to-point connections equals to the maximum matching in complete multipartite graph.

$$M_{max} = min\{\sum_{i=1}^{n-1} m_i, \lfloor \frac{1}{2} \sum_{i=1}^{n} m_i \rfloor\} \qquad (1)$$

The size would be equal to the first value if $m_n \leq \sum_{i=1}^{n-1} m_i$

- **Proof:** Construction Algorithm

renci

Multi-point connection

- **Theorem 2**. The maximum number of inter-cloud K- point broadcasting connections is equivalent to the maximum K-dimensional matching in a complete multipartite hyper- graph.
  - A hypergraph H = (V,E) consists of a set of vertices V and a family E of subsets of V, where each $e \in E$ is called a hyperedge. K-uniform if every hyperedge has exactly K vertices
  - K-point connection : complete K-uniform n-partite hypergraph
- **Proof : Construction**

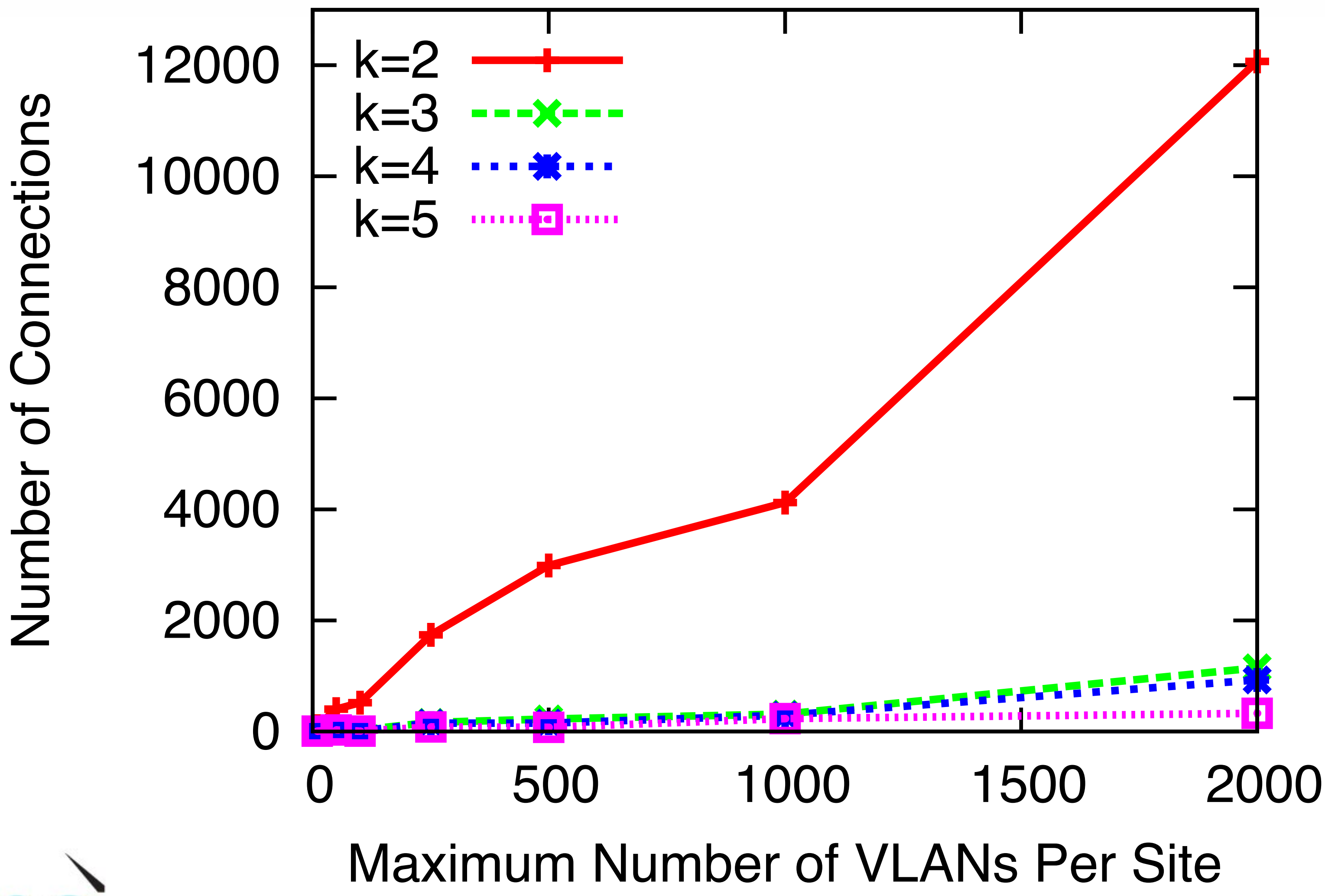
renci

Evaluation

- ExoGeni testbed: 14 rack sites
- Random #valns per site: maximum tag number: 10, 50, 100, 250, 500, 1000, 2000

m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	m_{11}	m_{12}	m_{13}	m_{14}
5	5	5	6	6	6	7	7	8	9	10	10	10	10
11	13	17	27	30	30	32	35	38	42	43	44	44	47
5	5	7	18	18	18	46	49	59	65	72	72	85	87
17	62	71	106	109	139	150	159	166	181	183	196	205	244
17	56	78	100	178	193	226	228	353	357	391	403	408	496
103	131	138	143	189	244	259	300	321	321	342	729	904	972
62	268	597	658	876	952	1143	1161	1191	1230	1259	1300	1372	1392

Result



k=2
k=3
k=4
k=5
Number of Connections
0
2000
4000
6000
8000
10000
12000
0
500
1000
1500
2000
Maximum Number of VLANs Per Site

Discussion

- Point-to-point connection capacity scales well with number of sites and available tags per sites
- Multi-point connection capacity scales much lower
- Results can be useful for backbone network dimensioning design

Further discussion

- Models and results can be generalized to other network layers
- The graph model can be used to develop new topology embedding algorithms
- Dynamic capacity: blocking performance
  - Maximum connections -> Erlang-B formula
  - Scheduling with small look-ahead window to archive low blocking performance and high system utilization

Acknowledge

GENI, **NSF** SDCI, and **DOE** ASCR Support