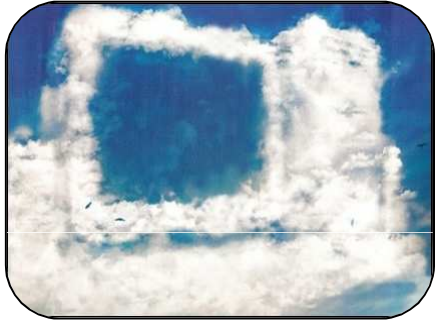


Trends and impacts of new generation data center networking

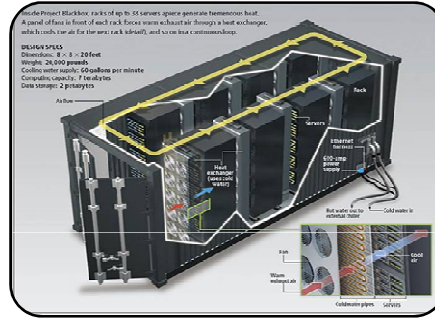
First CPqD International Workshop on New Architectures for Future Internet

Christian Esteve Rothenberg, 24/09/2006
University of Campinas

Agenda



Motivation



New Data Center Designs



Cost & Control

Requirements



Features



Green Internetworking



Inter-Cloud



Unicamp



Windows® Azure™



Next-generation DCN design drivers

- **Application needs**
 - Cloud services drive creation of huge DC designs
- **Technology trends**
 - Commodity servers + Virtualization (host + network)
- **Deployment constraints**
 - Space, location, resources
- **Operational requirements**
 - Auto-configuration, energy concerns, DC modularity
- **Cost-driven design**
 - Design for failure, 1:N resilience at data center level

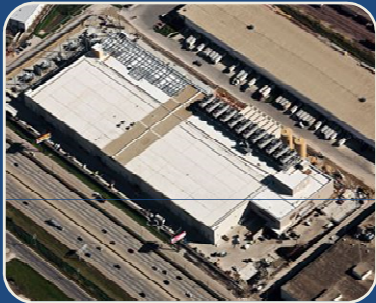
Today: How to interconnect servers in a data center?

- Network should not be bottleneck for DC applications

Future: Connectivity to/between data centers?

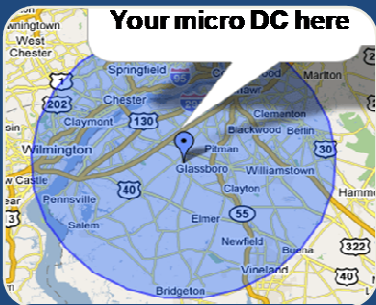
- Emergence of the Inter-Cloud

Types of cloud service data centers



Macro Data Center

- Specially dedicated facilities
- 100.000 or more servers and 10s of Mega-Watts of power at peak
- Computation in the cloud
(e.g., Amazon EC2, Windows Azure, Google AppEngine)



Micro Data Center

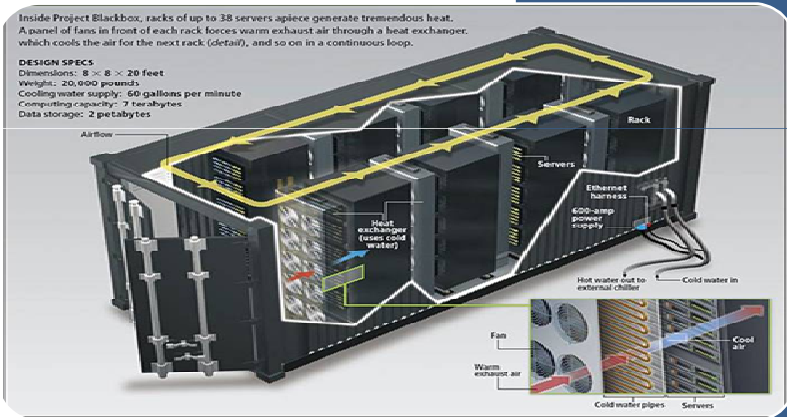
- Geo-diverse placed close to major population centers (e.g. CDN nodes)
- 1000s of servers and 100s of kilowatts
- Higher degree of independence between physical data center outages
- Opportunity to economically reach data center customers with low latency (e.g., front-end cloud apps)



Nano Data Center

- Located in the customer premises equipment (e.g., set-top-box)
- "Why don't we try to take the functionality that we have now in the data center, and distribute it across hundreds of thousands of set top boxes so that we have these 'Nano Data Centers" [EU FP7 NADA]
- P2P-like resource management. Low latency. Low cost.

Data center in a box



Container-based modular DC

- Efficient way to deliver computing and storage services
- 1000-2000 servers in a single container
- Sun Project Black Box (242 systems in 20')



Rackable Systems Container
2800 servers in 40'

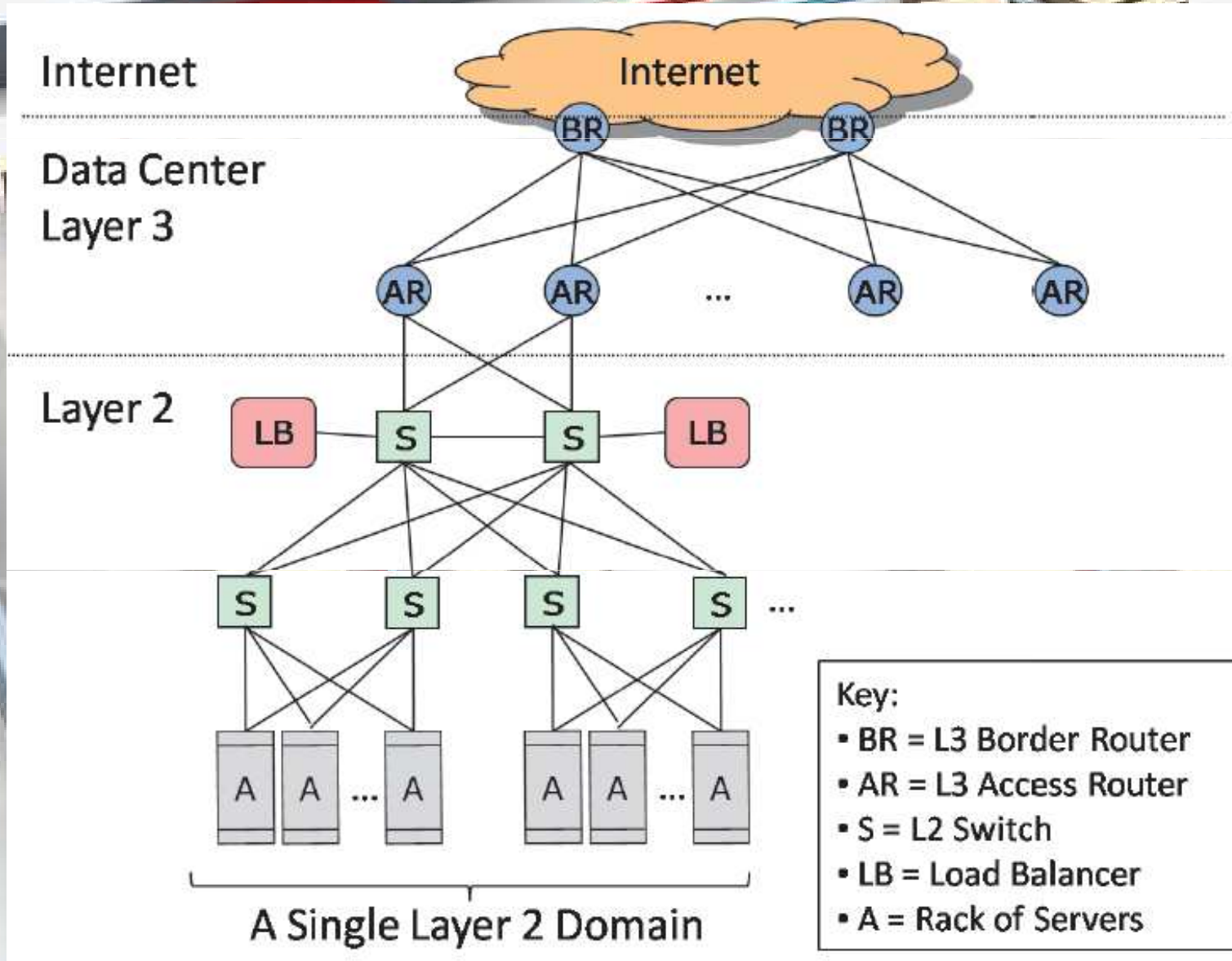
Core benefits:

- Easy deployment
 - High mobility
 - Just plug in power, network, & chilled water
- Increased cooling efficiency
- Manufacturing & H/W Admin. Savings
- Push modularity throughout the DC



The equipment yard at the Google data center in Belgium features no chillers. (Photo from Google) b.

Current DC network architectures



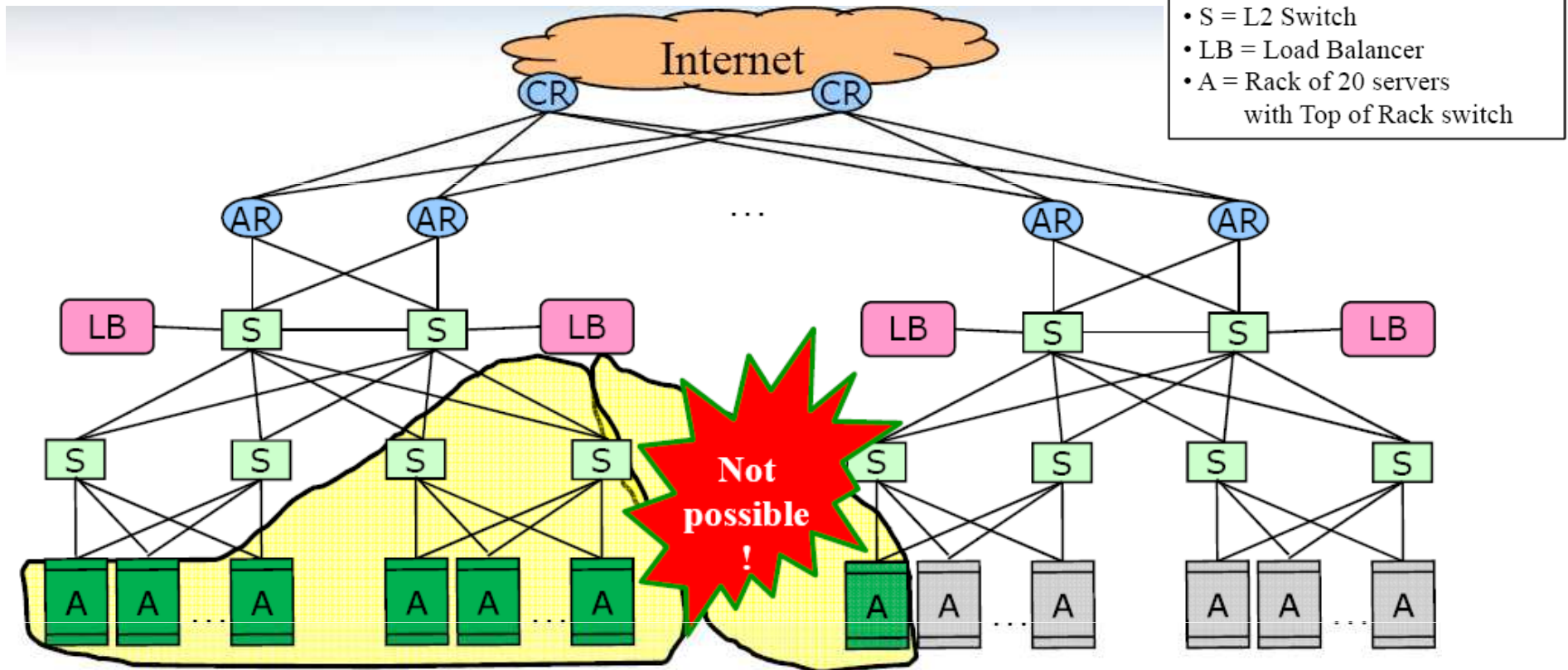
Some issues with conventional DC designs

Networking constraints of traditional L2/L3 hierarchical organization:

- Fragmentation of resources
- Limited server-to-server capacity
- Ethernet scalability
- Low performance under cloud application traffic patterns



Fragmentation of resources

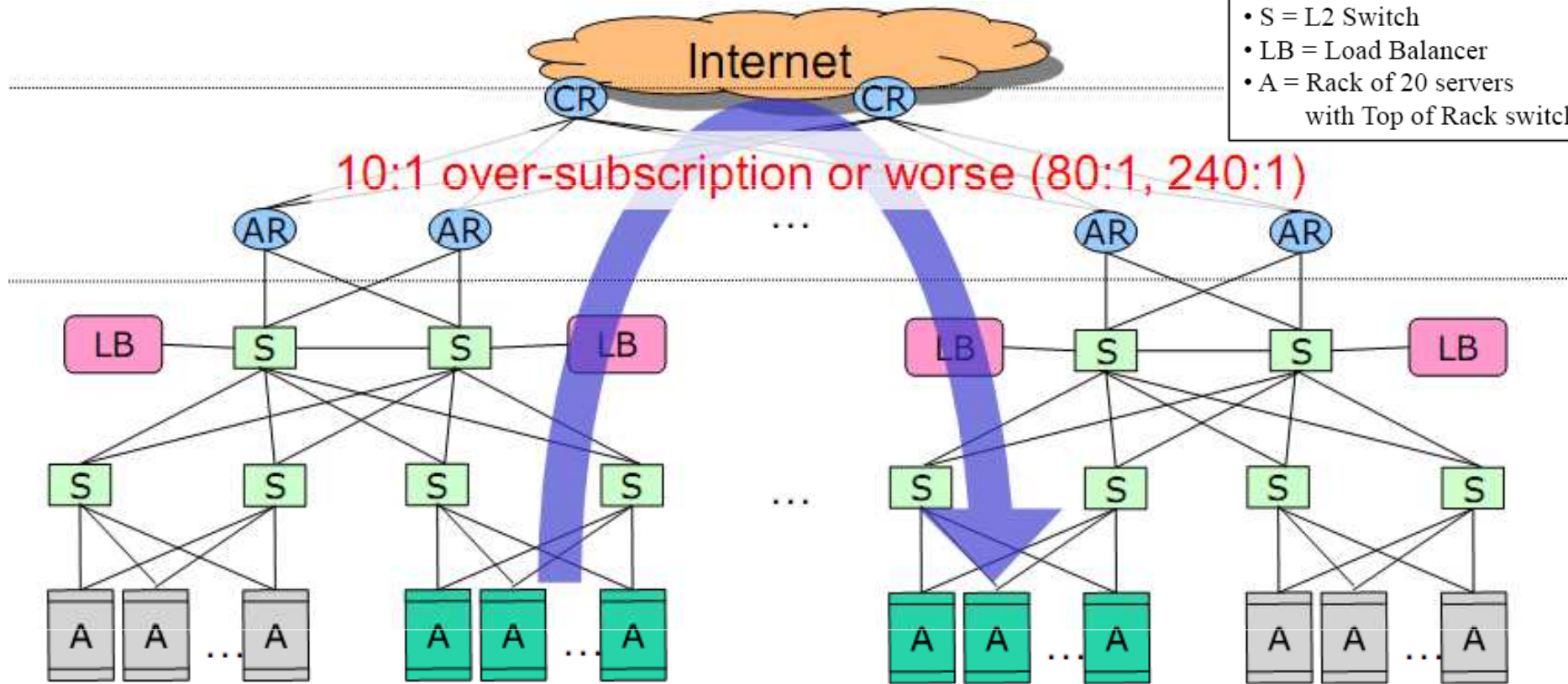


- Fragmentation of resources due to load balancers, IP subnets, ...
 - limits *agility* to dynamically assign services anywhere in the DC.
- Static Network assignment due to application to VLAN mappings, in-path middleboxes, ...

Limited server-to-server capacity

Key:

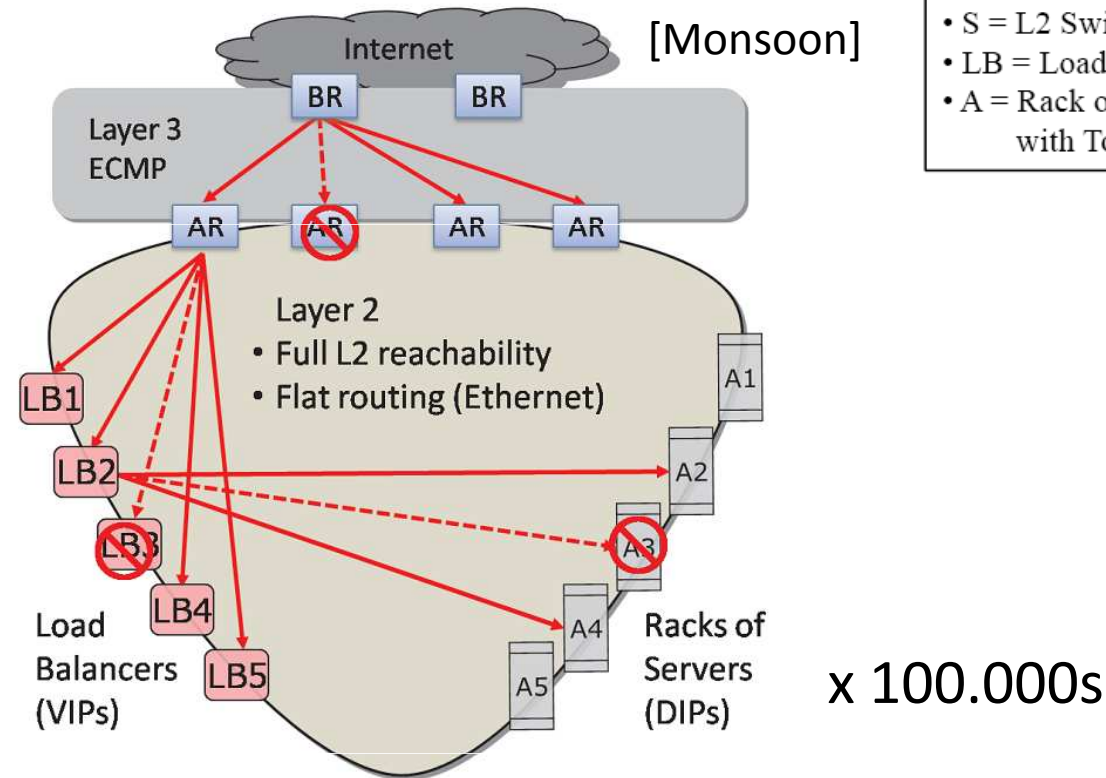
- CR = L3 Core Router
- AR = L3 Access Router
- S = L2 Switch
- LB = Load Balancer
- A = Rack of 20 servers with Top of Rack switch



Costly *scale up* strategy to support more nodes and better transfer rates

- Expensive equipment at the upper layer of the hierarchy.
- High over-subscription rates i.e. poor server bisection BW

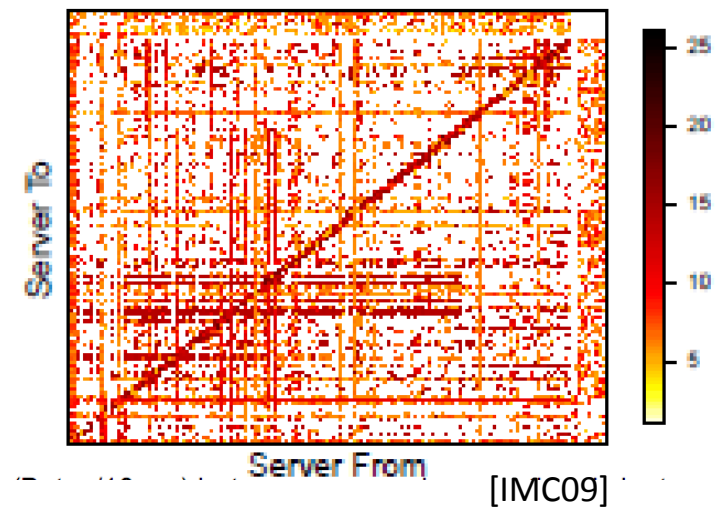
Layer 2 (Ethernet) scalability



Current layer 2 architectures cannot scale

- limited switch *state* for forwarding tables (flat routing)
- *performance* (bisection BW) limitations (i.e. standard spanning tree protocol limits fault tolerance and multipath forwarding)
- ARP broadcast overhead

DC “traffic engineering”



- DC traffic is highly dynamic and bursty
 - 1:5 ratio of external vs. internal traffic
 - Traditional traffic engineering does not work well (TM changes constantly)
- Goal of DC traffic engineering
 - Location-independent uniform BW and latency between any two servers
 - For any TM! DC patterns (1:1, 1:M, N:N)
- Approach
 - Avoid spanning tree to make all available paths could be used for traffic
 - Load balancing: E.g., TM oblivious routing, VLB [Monsoon, VLB]
- Additional requirement
 - Force application traffic through middleboxes (firewalls, DPI, intrusion det., load balancers, WAN opti., SSL offloaders)

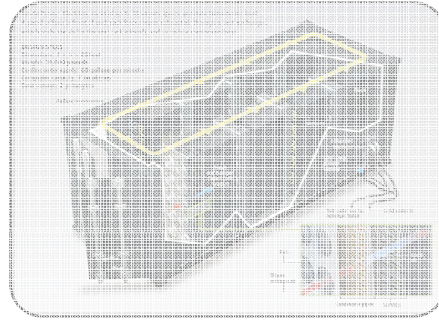
New Generation Data Center Networking

Goals	Requirements	Features
Resource Pooling (servers and network eq.) & Agility	R1: Any VM to any physical machine. <ul style="list-style-type: none"> - Let services “breathe”: Dynamically expand and contract their footprint as needed - L2 semantics 	<ul style="list-style-type: none"> · ID/loc split · Scalable L2
	R2: High network capacity <ul style="list-style-type: none"> - Uniform BW and latency for various traffic patterns between any server pair - 1:1, 1:M, N:N efficient communications along any available physical paths 	<ul style="list-style-type: none"> · Multipath support · New TE (load-balancing)
Reliability	R3: Design for failure. <ul style="list-style-type: none"> - Failures (servers, switches) will be common at scale. 	<ul style="list-style-type: none"> · Fault-tolerance
Low Opex	R4: Low configuration efforts <ul style="list-style-type: none"> - Ethernet plug-and-play functionality 	<ul style="list-style-type: none"> · Auto-config.
	R5: Energy efficiency <ul style="list-style-type: none"> - Networking design for idle link/server optimization 	<ul style="list-style-type: none"> · Energy/Cost-awareness
Low Capex	Use commodity hardware (scale-out strategy)	
Control	Include middlebox services in the data path as required	<ul style="list-style-type: none"> · Network ctrl.

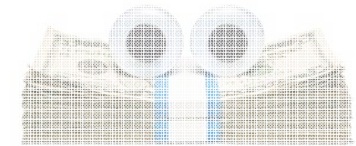
Agenda



Motivation

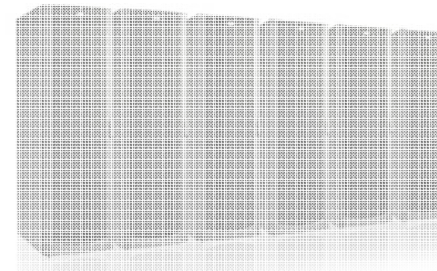


New Data Center Designs



Cost & Control

Requirements



Features



Green Internetworking



Inter-Cloud



Unicamp



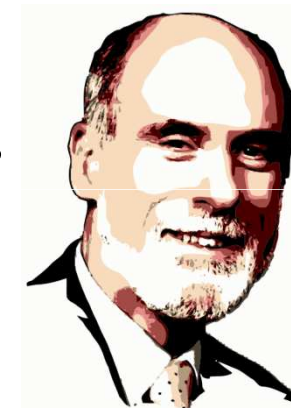
Impacts of the cloud data centers on the Future Internet?

THE INTER-CLOUD



The Inter-Cloud

*“The Cloud represents a **new layer** in the Internet architecture and, like the many layers that have been invented before, it is an open opportunity to add functionality to an increasingly global network” - Vint Cerf, 2009 ^[1]*



“History doesn’t repeat itself, but it does rhyme.” - Mark Twain

Cloud Initiatives that have an analogue in the Internet’s past ^[2]:

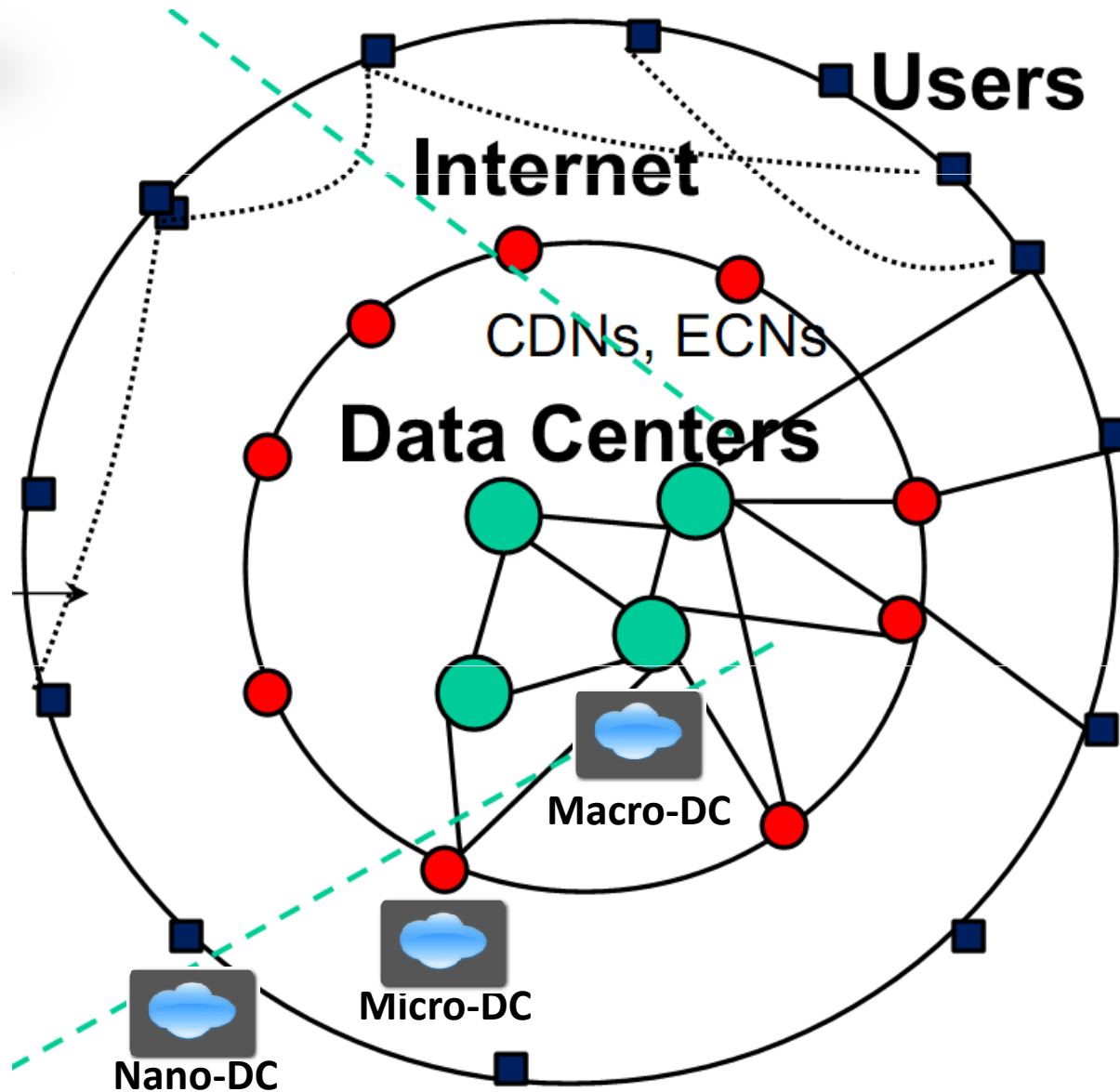
- The rising importance of *academia*.
- Increasing interest in *interoperability* among cloud vendors.
 - Today’s clouds like network islands before IP
- *Carrier* interest in new service opportunities.

^[1] <http://googleresearch.blogspot.com/2009/04/cloud-computing-and-internet.html>

^[2] http://blogs.cisco.com/datacenter/comments/is_the_intercloud_history_repeated/

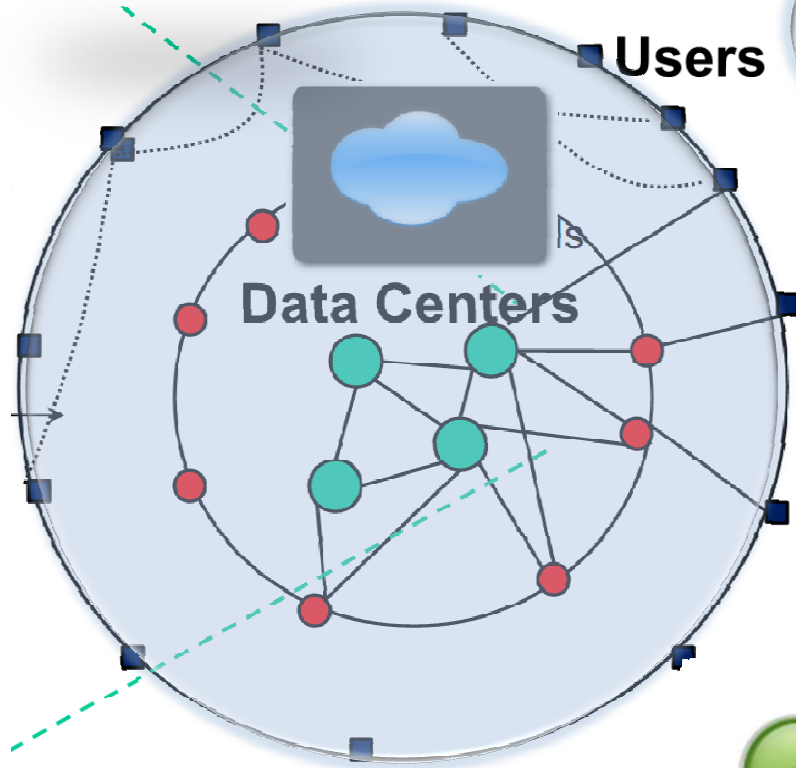


Impacts of the cloud on the FI?

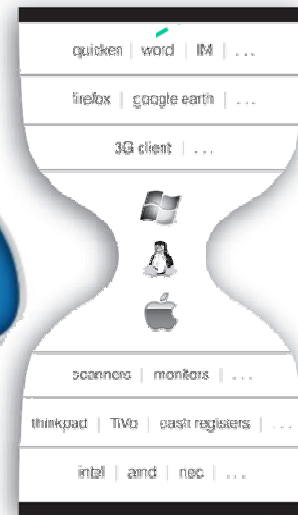
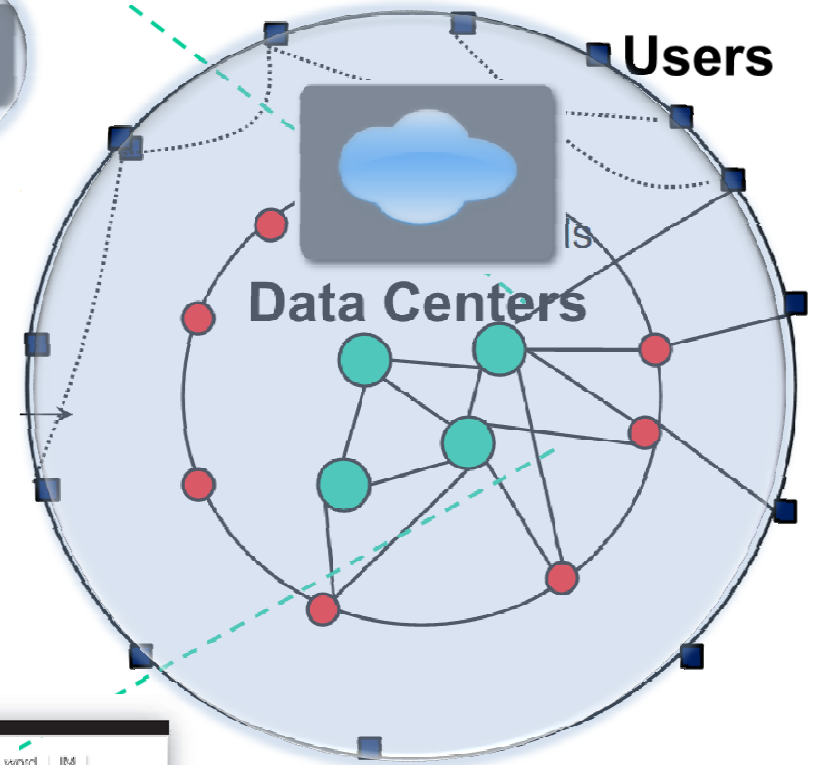




Isolated Clouds over IP

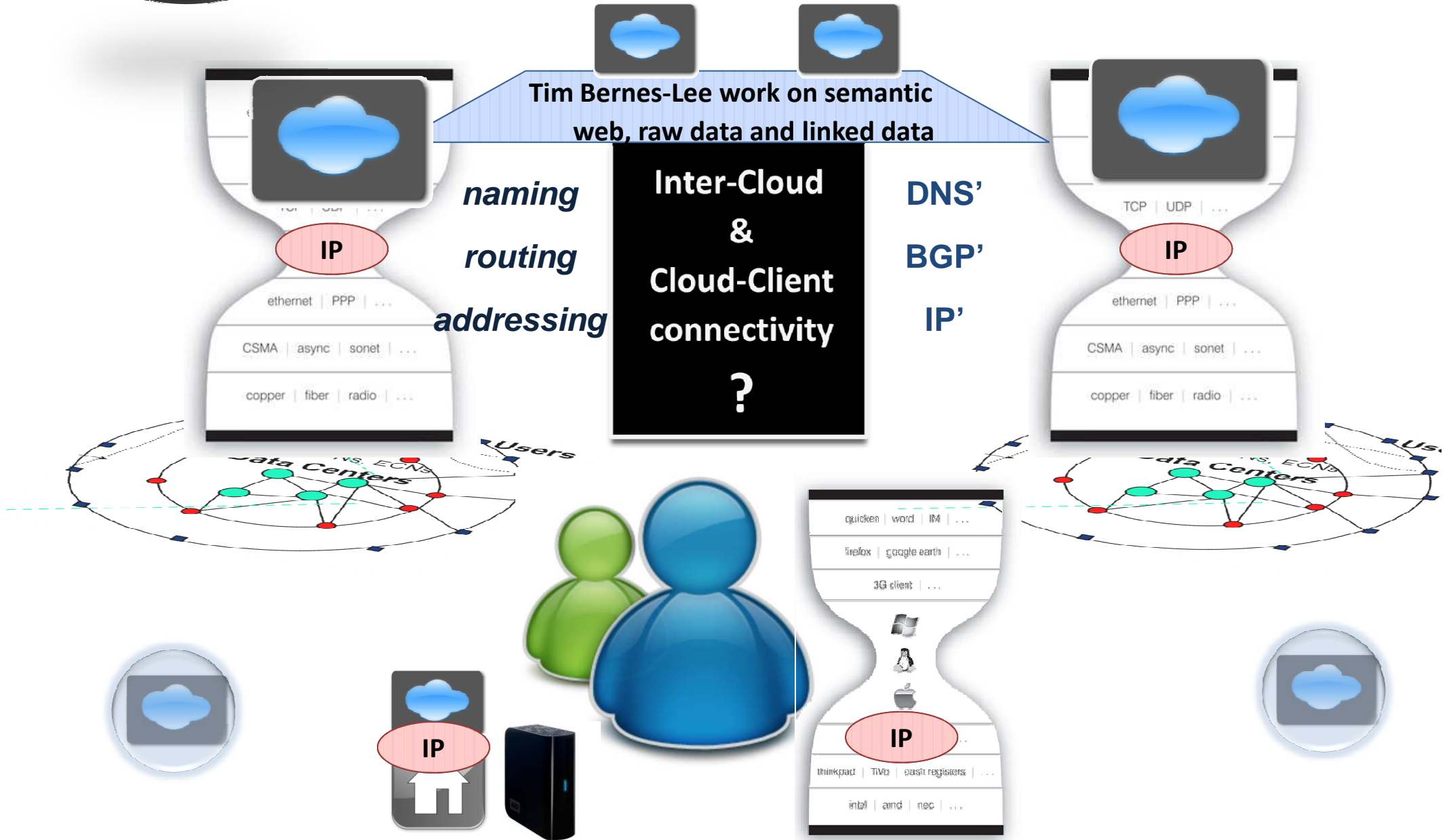


naming DNS
routing BGP
addressing IP

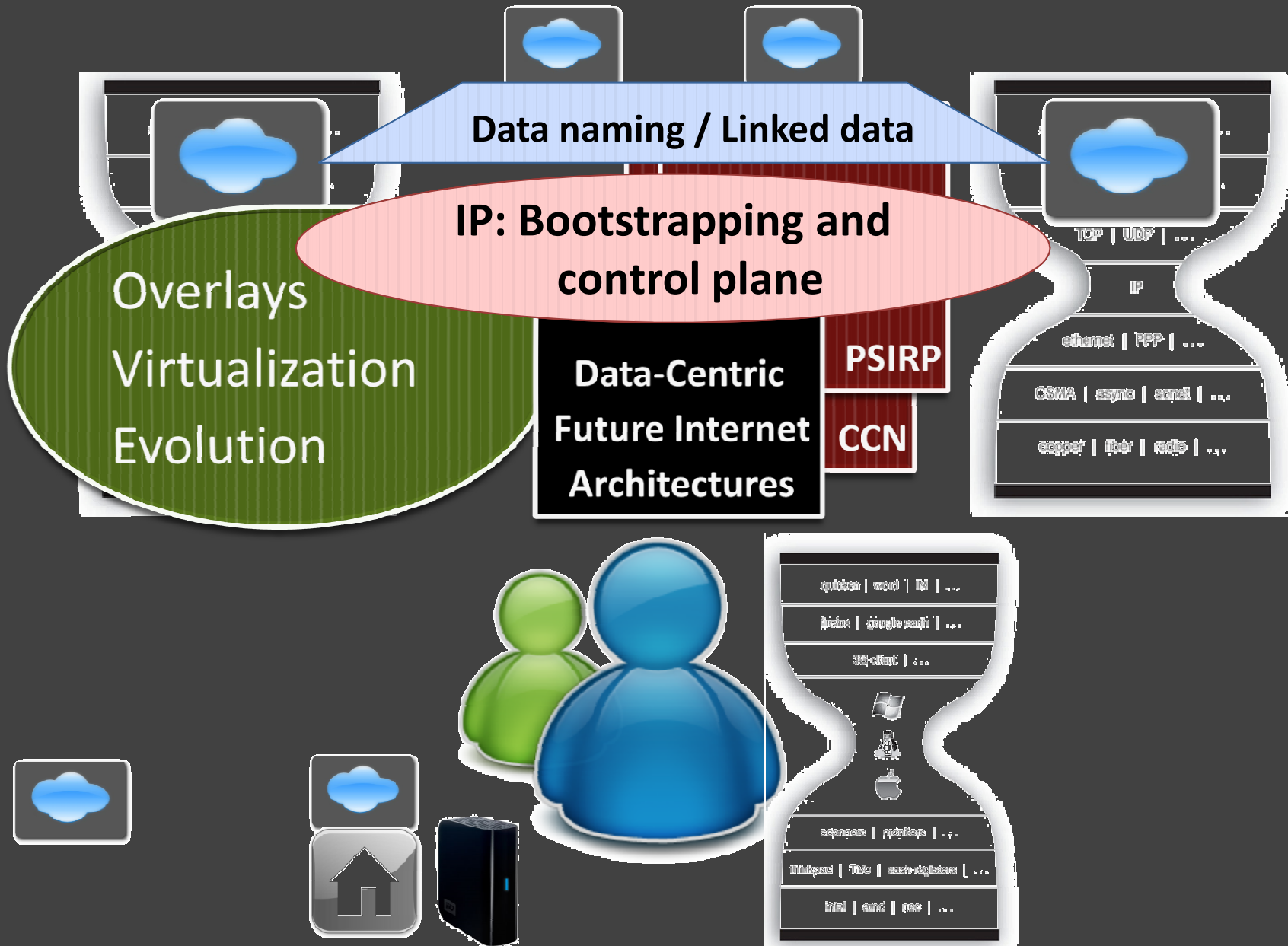




Re-thinking cloud connectivity



Emergence of data-centric architectures



Networking Impacts of the Inter-Cloud

Drivers

- User demand for Virtual Private Clouds
 - QoS, privacy, security, availability, etc.
- Inter-Cloud Connectivity
 - Identity of information, security, agility, cost, etc.

Shorter term

- Incentives for adoption of Sec-DNS, Sec-BGP, IPv6 and so-forth patches
- Demand for end-to-end optical paths
- Emergence of Transit Portals (disruption in traditional peering practices)


Longer term

- Novel, scalable, information-oriented connection services i.e. next-gen. MPLS or IPsec VPNs
- *Put your favourite research here* (e.g., Van Jacobson CCN, EU FP7 PSIRP)

More research questions

- Role of CDN overlays and infrastructure providers (e.g., with e2e virtualization in place)
- From Green Computing to Energy/Cost-aware Internetworking



A perspective view of a server room aisle. The aisle is flanked by rows of server racks on both sides. The ceiling and walls are covered in dense, vibrant green foliage, including various leafy plants and vines, creating a lush, natural environment. The floor is a light-colored, reflective surface. The overall scene suggests a harmonious blend of technology and nature.

Towards a green future Internet

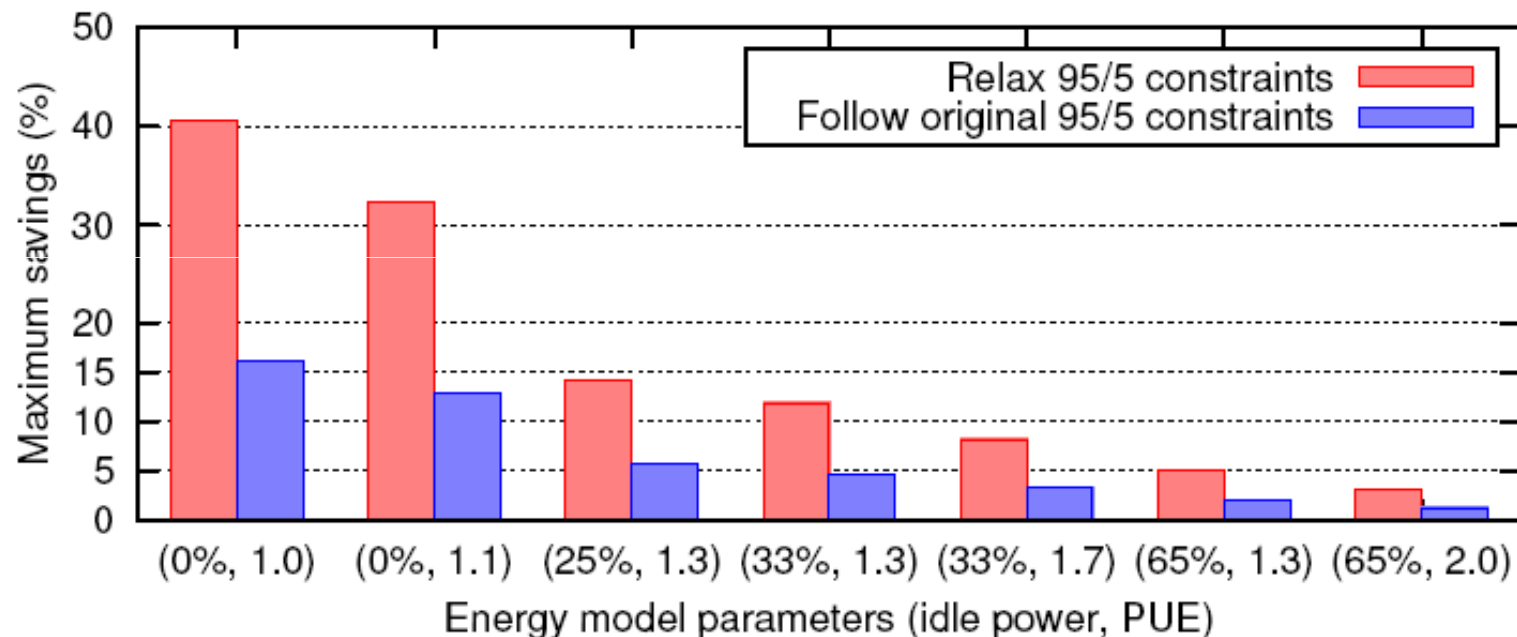
Cost-Aware Internet Routing



40%

savings of a cloud computing installation's power usage by dynamically re-routing service requests to wherever electricity prices are lowest on a particular day, or perhaps even where the data center is cooler.

From "Follow the energy price!" to "Follow the wind, the sun or the moon!"



[Qureshi et al, "Cutting the Electric Bill for Internet-Scale Systems", SIGCOMM'09]

Green Internetworking



- Internet-routing algorithms that track electricity price fluctuations
 - Take advantage of daily and hourly fluctuations
 - Weight up the physical *distance* needed to route information against the potential cost *savings* from reduced energy use.
- Reduce DC electricity costs
 - + tax incentives for (near) zero-carbon-emission DCs
 - DCN designs to optimize idle links and idle servers ?

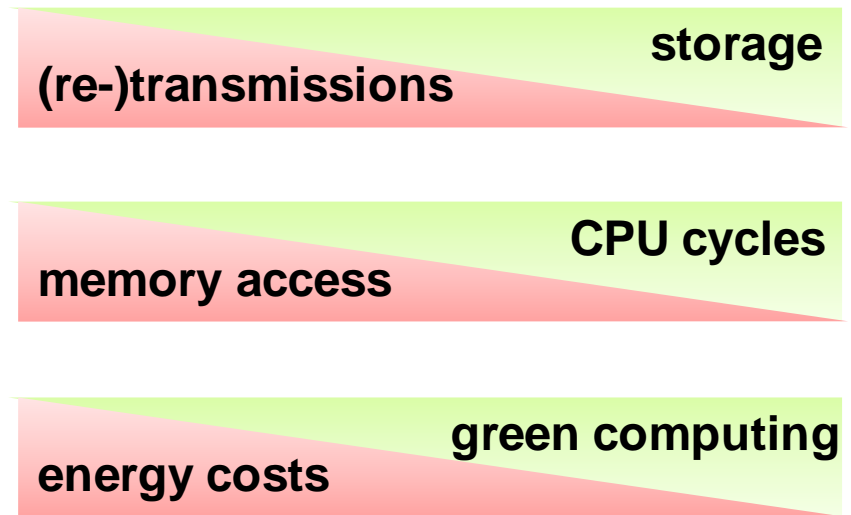
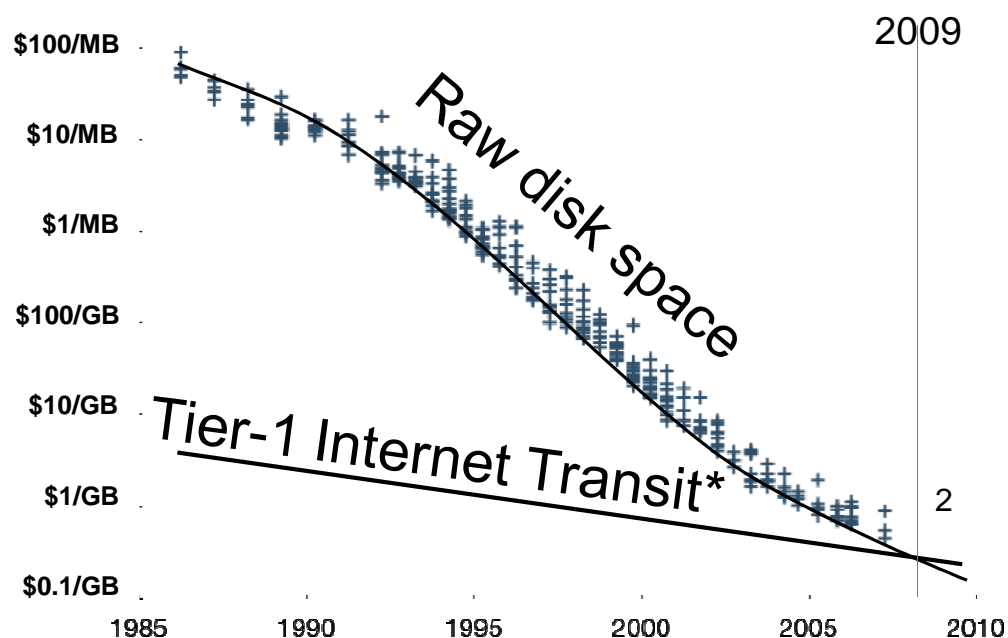
“Next generation cloud computing to distribute data centers so that, when the wind is blowing in Wyoming, computing tasks are shifted to the data center there, and when the wind stops blowing, computing shifts back elsewhere – to where the sun is shining, for example. The same could be done for network routers using standard routing protocols”

Bill St. Arnaud, chief research officer at CANARIE

[<http://telephonyonline.com/global/news/carbon-trade-arnaud-0626/>]

Network economics & Future Internet

- **Data Centers are like Factories¹**
 - Number 1 Goal: Maximize useful work per dollar spent
- **And the future network of networks?**
 - Incentives for re-architeting the Internet? DC-driven incentives???
- **Think like an economist/industrial engineer as well as a computer scientist**
 - Understand where the dollar costs come from
 - Use computer science to reduce/eliminate the costs / complexity



¹ cf. Greenberg SIGMETRICS tutorial
² Preliminary data [Nikander'09]

Activities at Unicamp



Embracing the Data Center Networking research:

- From “commoditization in the DC network is the next frontier”
- To “DC network customization (switch programmability) is the next frontier”

Long-time cooperation with Ericsson Research

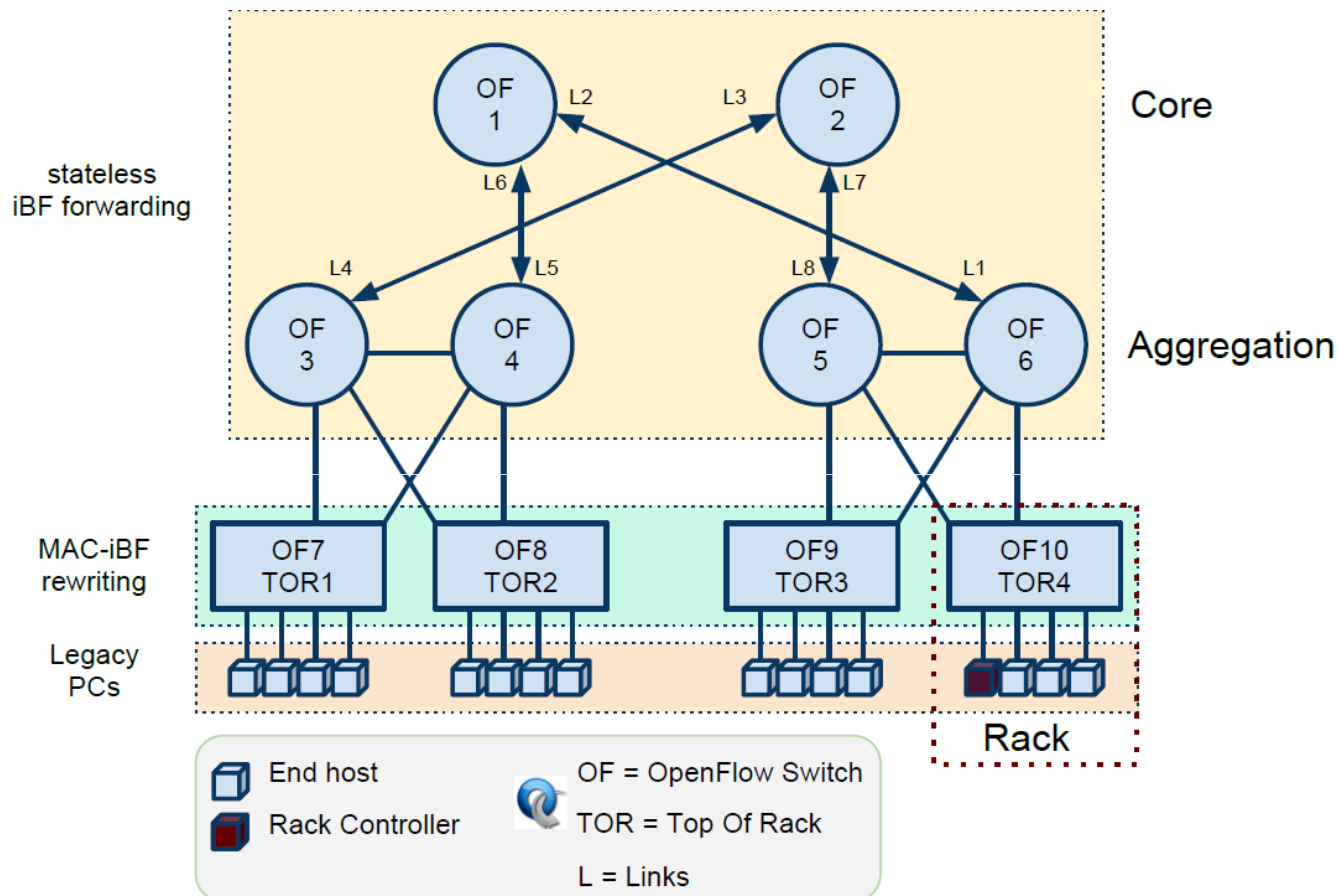
- Control plane of optical networks
- Node ID Architecture
- Routing on flat identifiers



Activities at Unicamp



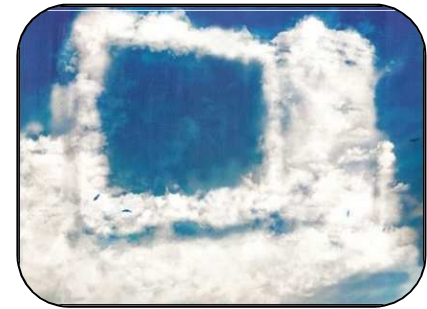
- Load balanced DCN with in-packet Bloom filters (iBF)¹
 - OpenFlow testbed



¹ DC application of P. Jokela et al., [LIPSIN: Line Speed Publish/Subscribe Inter-Networking](#). SIGCOMM'09

Conclusion

- Lots of interesting networking research issues towards novel DC and service provider network designs
 - Driven by cloud-computing demands and cost + control goals
- Potential impacts for the future Internet
 - The Inter-Cloud shaped by how geo-distributed DC footprints communicate among them and with edge clients
 - Energy-awareness

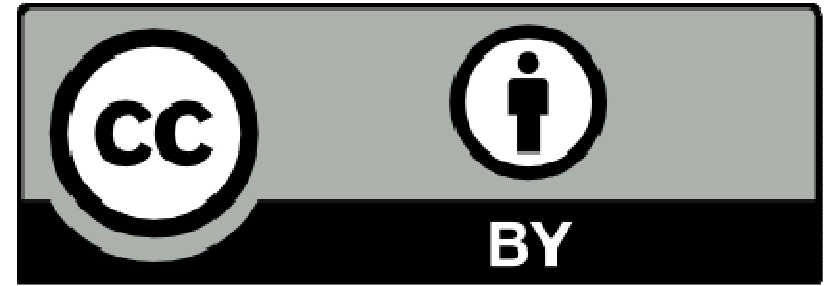


Thank you!

Questions?

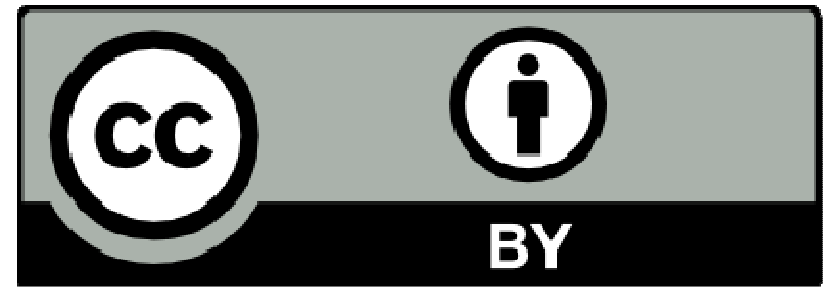


REFERENCES

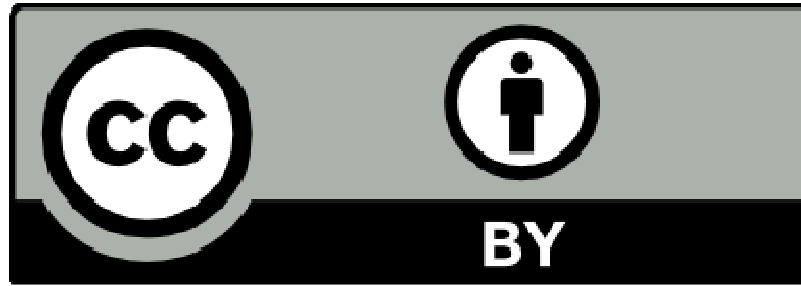


- A. Greenberg and et al., “The cost of a cloud: research problems in data center networks.” *SIGCOMM CCR.*, 2009.
- A. Greenberg and et al., “Monsoon: Towards a Next Generation Data Center Architecture: Scalability and Commoditization”
- A. Greenberg and et al., “VL2: A Scalable and Flexible Data Center Network” , SIGCOMM 09
- R.Niranjan et al., “PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric” , SIGCOMM 09
- “BCube: A High Performance, Server-centric Network Architecture for Modular Data Centers”, SIGCOMM 09
- Wu et al. “MDCube: A High Performance Network Structure for Modular Data Center Interconnection”, CoNext09.
- Benson et al., “Understanding Data Center Traffic Characteristics”, WREN 09
- Costa et al. “Why should we integrate services, servers, and networking in a Data Center?”, WREN 09
- Valancius et al. “Transit Portal: Bringing Connectivity to the Cloud”

REFERENCES



- Vaquero et al., “Break in the Clouds: Towards a Cloud Definition”
- EU Commission, “Code of Conduct on Data Centres Energy”
- Guo et al., “DCell: A Scalable and Fault-Tolerant Network Structure for Data Centers”
- Joseph et al. “A Policy-aware Switching Layer for Data Centers”
- Greg Schulz, “The Green and Virtual Data Center”
- Al-Fares et al. “A Scalable, Commodity Data Center Network Architecture”
- Nano Data Center, EU FP7 NADA, www.nanodatacenters.eu
 - http://www.computerworld.com.au/article/253324/set_top_boxes_revolutionise_internet_architecture
- Laoutaris et al., “ECHOS: Edge Capacity Hosting Overlays of Nano Data Centers”
- Prachi Patel-Predd et al., “Cutting the Power in Data Centers”,
- Qureshi et al., “Cutting the Electric Bill for Internet-Scale Systems” SIGCOMM 09
- <http://telephonyonline.com/global/news/carbon-trade-arnaud-0626/index.html>
- http://blogs.cisco.com/datacenter/comments/is_the_intercloud_history_repeated/
- P. Jokela et al., “LIPSIN: Line Speed Publish/Subscribe Inter-Networking” SIGCOMM'09



Credits

- Ericsson Research
- Prof. Mauricio Magalhaes, F. Verdi et al.
- Sudipta Sengupta, Slides on “issues with conventional DC designs”, from “Oblivious Routing and Applications”, Tutorial at IEEE ICC 2009.
- Guo et al, Slide on “Container-based modular DC”
- A. Greenberg and D.A. Maltz, „What Goes into a Data Center”, SIGMETRICS 2009 Tutorial, Image on slide on “Impacts of the cloud on the FI?”.

Images

- Switch slide 10, <http://lcg.web.cern.ch/LCG/lhcgridfest/partners.htm>
- Interconnection hw, slide 4, http://www.microsoft.com/presspass/events/msrtechfest/images/LowPowerProcessors_print.jpg

Images

BACK-UP

Re-thinking the cloud service infrastructure design

COST

- Commoditization (hosts + network)
- Scale-out strategy
- Virtualization
- Energy

&

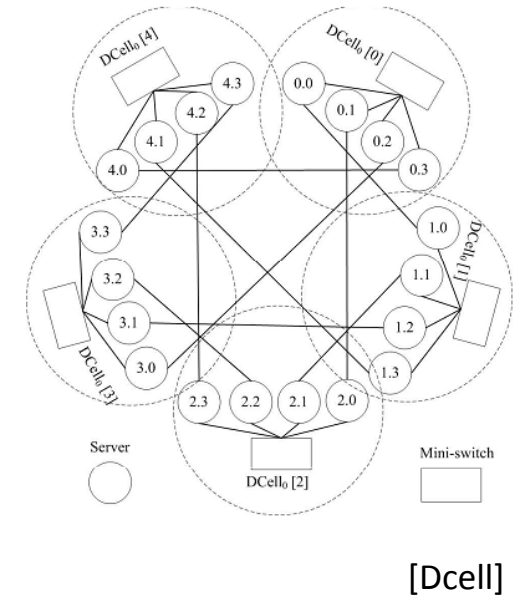
CONTROL

- Customization (host & network)
- Scalability
- Agility



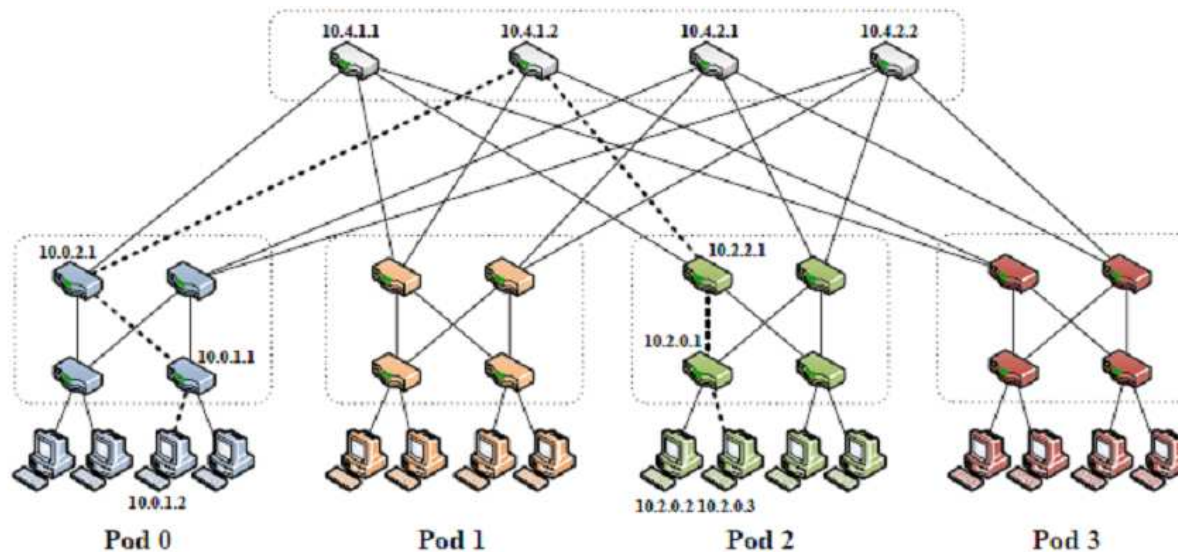
Server-centric designs

- Philosophy:
 - “Commoditization in the network is the next frontier”
 - “End-host customization”
- Leading examples:
 - Microsoft Research designs: [Monsoon, VL2, (MD)Bcube, FiConn, Dcell]
- Routing intelligence *solely* into servers to handle load-balance and fault-tolerance
 - Servers with multiple NICs act as routers (aka P2P)
 - Switches do not connect to switches (aka crossbars)
 - Leverage commodity instead of high-end switches to scale out
- Server-centric interconnection network in the spirit of mesh, torus, ring, hypercube and de Bruijn graphs
 - Different to HPC are the scale and the Ethernet/IP/TCP considerations



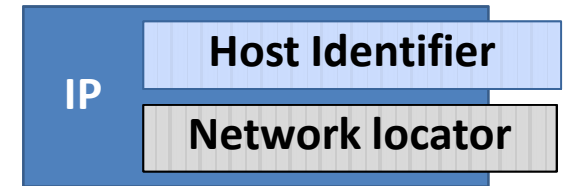
Network-centric designs

- Servers connect to a *switching fabric* such as a Clos network, Butterfly and a fat-tree topology.
- Modification of the *network control plane* of the network, leaving the switch hardware and end hosts untouched.
- Network-wide *controllers*
 - E.g., a centralized fabric manager resolves IP-to-PMAC mappings and responds to ARP requests intercepted by edge switches [Portland].
- Network customization through *switch programmability*



[Fat-tree, Portland]

ID/loc separation

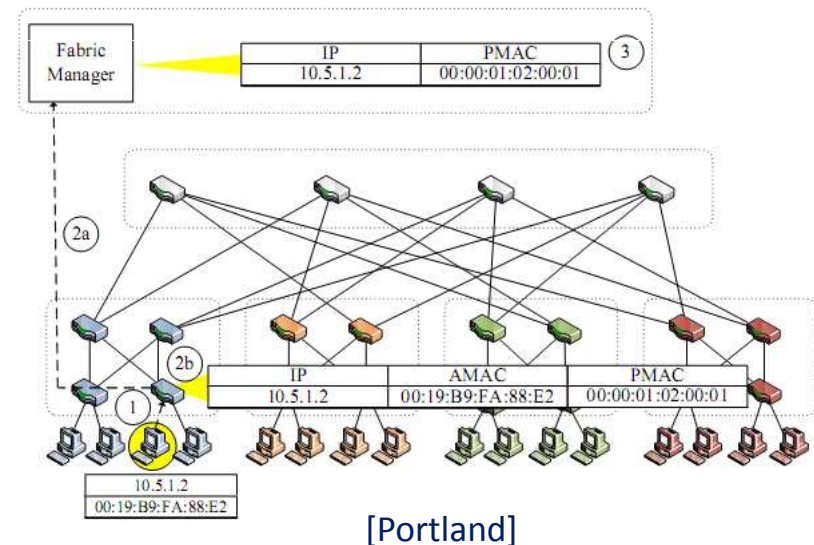


- ID/loc split is a common approach in DCN designs
 - Location-independent Addressing: Services use location-independent addresses that decouple the server's location from its address.
- Goal: Agility - Any server, any service
 - Let services “breathe”
 - dynamically expand and contract their footprint as needed
 - Any server can become part of any server pool while simplifying configuration management, enable anycast and live VM migration.
- Different flavors:
 - App. Address / Netw. Address [VL2] : IP-in-IP with anycast-based ECMP
 - Virtual IP / Direct IP [Monsoon] : MAC-in-MAC forwarding
 - IP / location-based Pseudo MAC [Portland] : Edge switches rewrite MAC
- Beyond id/loc split
 - Shim header approaches to encode network paths [(MD)BCube]
 - Source routing with in-packet Bloom filters [LIPSIN, Unicamp]

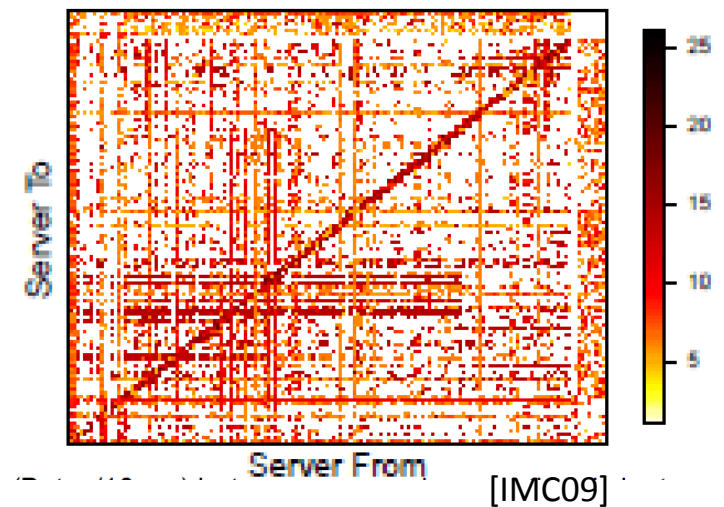
Network Controllers



- Logically Centralized
 - But implemented as a distributed service (fault tolerant, consistent, etc.) in commodity servers.
 - Centralized directory and control plane acceptable [4D]
- Provide Routing Services maintaining network state
 - Resolve location-independent Application Address into (set of) locators
 - E.g., Resolve ARP requests for service IPs into a (list of) MAC addresses of servers running the application identified by the service IP
 - Application-specific load balancers
 - Health services
 - Multicast management
- Examples:
 - Fabric Manager in [Portland]
 - Directory Service [VL2, Monsoon]



DC “traffic engineering”



- DC traffic is highly dynamic and bursty
 - 1:5 ratio of external vs. internal traffic
 - Traditional traffic engineering does not work well (TM changes constantly)
- Goal of DC traffic engineering
 - Location-independent uniform BW and latency between any two servers
 - For any TM! DC patterns (1:1, 1:M, N:N)
- Approach
 - Avoid spanning tree to make all available paths could be used for traffic
 - Load balancing: E.g., TM oblivious routing, VLB [Monsoon, VLB]
- Additional requirement
 - Force application traffic through middleboxes (firewalls, DPI, intrusion det., load balancers, WAN opti., SSL offloaders)

Inefficient enforcement of middlebox policies

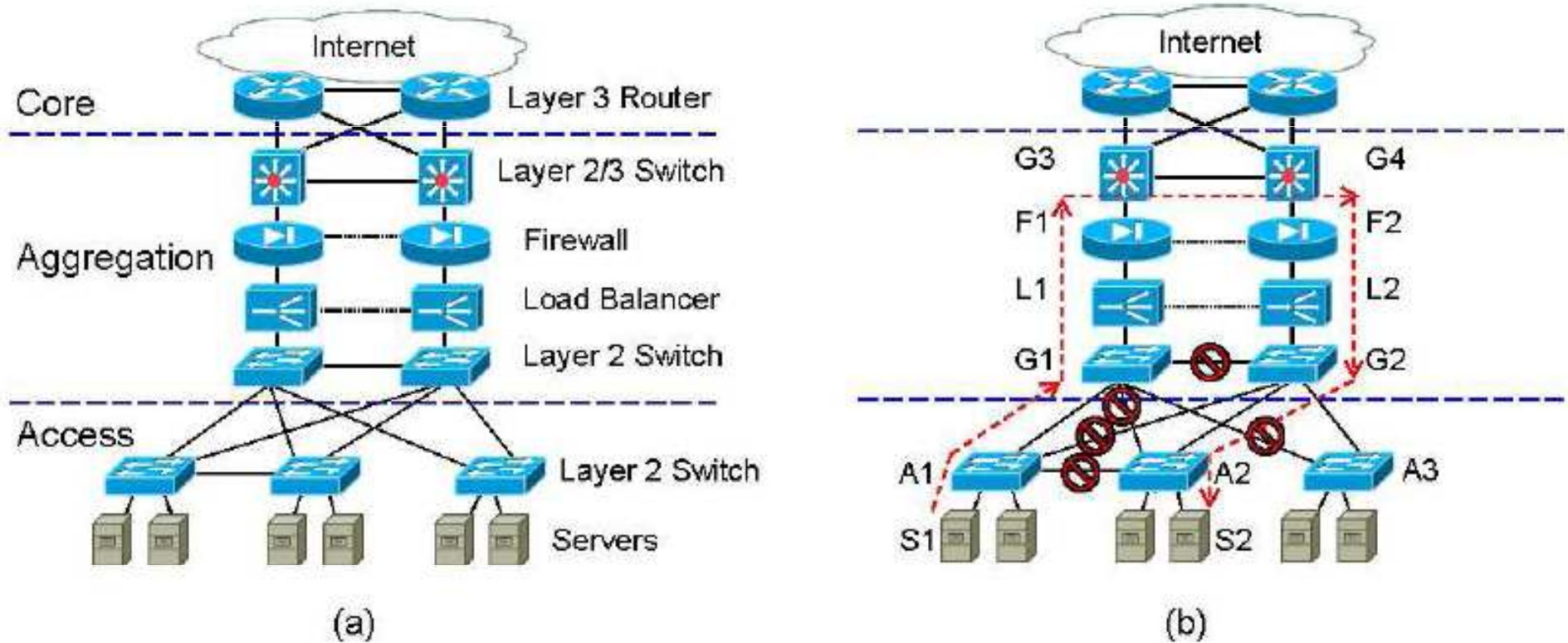
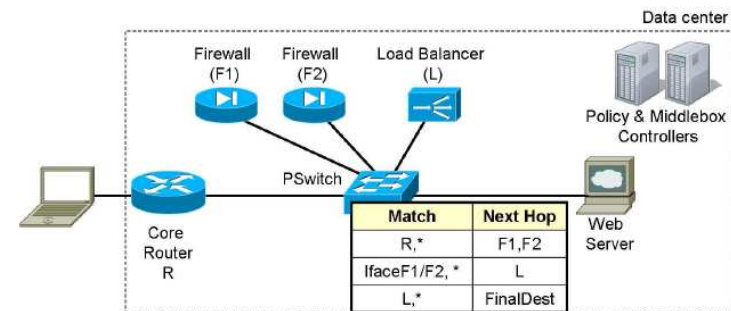


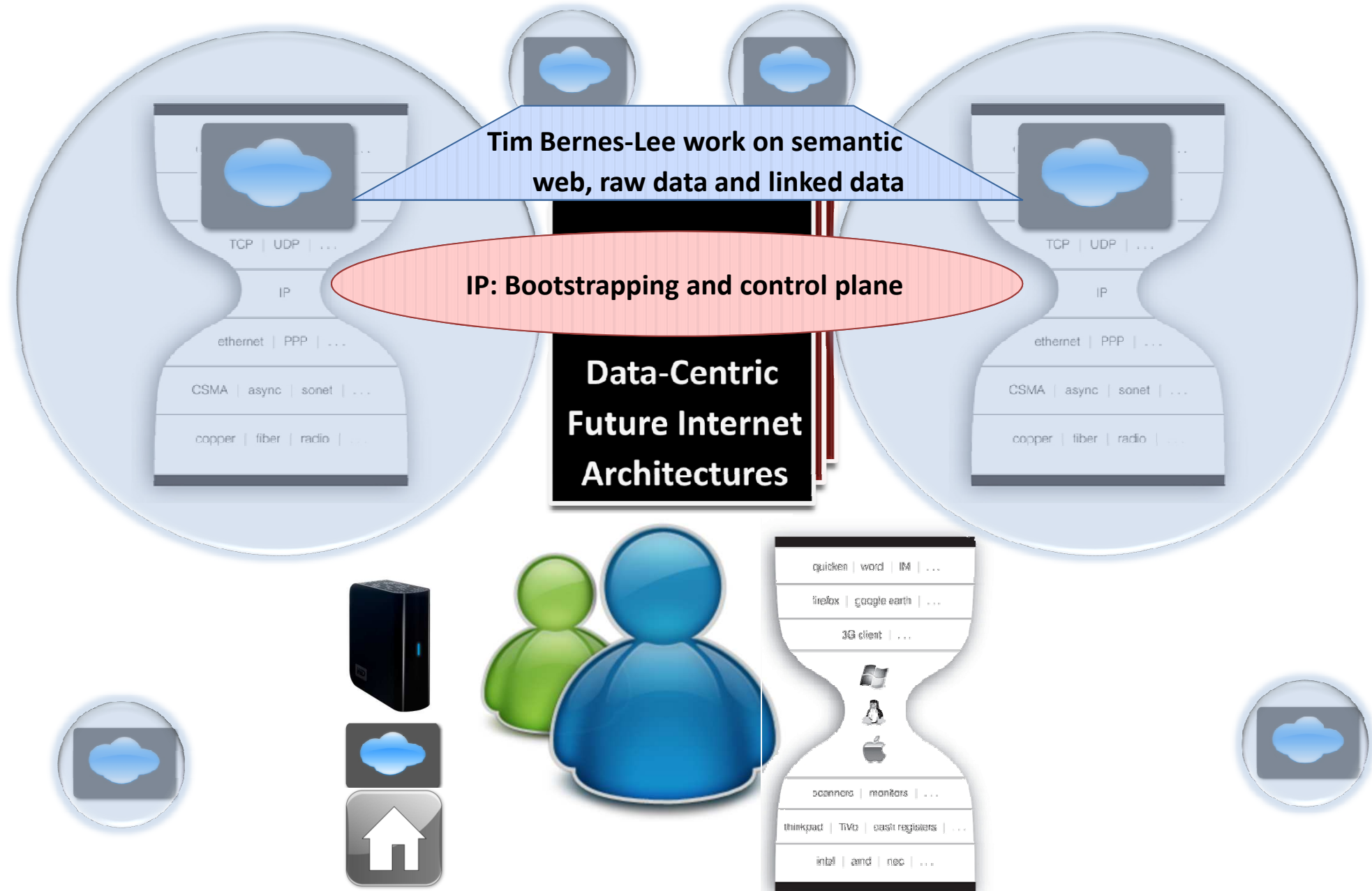
Figure 1: (a) Prevalent 3-layer data center network topology. (b) Layer-2 path between servers $S1$ and $S2$ including a firewall.

Path enforcement options:

- Remove physical connectivity:
- Manipulate link costs:
- Separate VLANs:



The Inter-Cloud shaping the Future Internet?



Vint Cerf's open questions on inter-cloud

- How should one reference another cloud system?
- What functions can one ask another cloud system to perform?
- How can one move data from one cloud to another?
- Can one request that two or more cloud systems carry out a series of transactions?
- If a laptop is interacting with multiple clouds, does the laptop become a sort of “cloudlet”?
- Could the laptop become an unintended channel of information exchange between two clouds?
- If we implement an inter-cloud system of computing, what abuses may arise?
- How will information be protected within a cloud and when transferred between clouds.
- How will we refer to the identity of authorized users of cloud systems?
- What strong authentication methods will be adequate to implement data access controls

<http://googleresearch.blogspot.com/2009/04/cloud-computing-and-internet.html>

