Resilient Optical Network Virtualization

C. Develder⋆
B. Jaumard⋆⋆
M. Bui

⋆ Ghent University – iMinds
Dept. of Information Technology – IBCN

Concordia University
Computer Science and Software Engineering Department
⋆⋆ Concordia University Research Chair – Optimization of Communication Networks

C. Develder

- PhD, Ghent University, 2003
  Design and analysis of optical packet switching networks
- Professor at Ghent University since Oct. 2007
- Research Interests: dimensioning, modeling and optimizing optical (grid/cloud) networks; smart grids; information retrieval and information extraction
- Visiting researcher at UC Davis, CA, USA, Jul-Oct. 2007 (optical grids)
- Visiting researcher at Columbia Univ., NY, USA, 2013-14 (IR/IE)
- Industry Experience: network planning/design tools OPNET Technologies (now part of Riverbed), 2004-05
- More info: http://users.atlantis.ugent.be/cdvelder

B. Jaumard

- Concordia University Research Chair, Tier 1, Optimization of Communication Networks
- Professor, Computer Science and Software Engineering (CSE) Department, Concordia University
- Researcher, GERAD (Group for Research in Decision Analysis) & CIRRELT (Inter-University Research Center on Enterprise Networks, Logistics and Transportation) & Calcul Quebec
- Research Interests: mathematical modeling and algorithm design for large-scale optimization problems arising in communication networks, transportation networks and artificial intelligence.
- Industry Experience: SDN & Virtualization, CIENA ; Anycast & QoS in Clouds, Siemens ; Train & Locomotive Optimization Problems (Energy Consumption), CPR - Canadian Pacific Railway ; Crew Scheduling, ViaRail.
- More info: http://users.encs.concordia.ca/ bjaumard/

M. Bui

- PhD student in the Computer Science and Software Engineering Department, Concordia University, 2010 - 2014
  Survivable logical topologies for optical networks.
- M.Sc, Institute de la Francophonie pour l'Informatique (IFI) in Hanoi, Vietnam, 2006
- B.Sc, Hanoi University of Technology, Vietnam, 2002
- Research Interests: Mathematical programming, large scale optimization techniques for design and planning resilience optical networks and cloud.

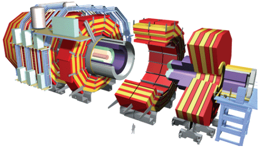
Outline

- Optical and DCs
- Virtualization
- Server Location
- Resiliency
- Resiliency & QoS
- Energy Issues

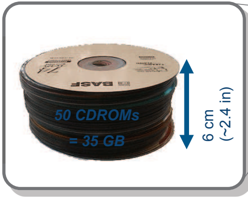
Converged Optical and Data Centers

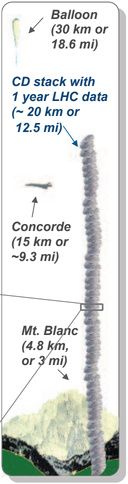
Background: “Optical grids”

- eScience:
    - By 2015 it is estimated that **particle physicists** will require exabytes (10^{18}) of storage and **petaflops** (10^{15}) per second of computation
    - CERN’s LHC Computing Grid (LGC), when fully operational generates **15 petabytes** annually (that’s ~2Gbit/s)



LHC = Large Hadron Collidor: particle accelerator





“Optical grids” for consumer services?

E.g., **video editing**:
2Mpx/frame for HDTV, suppose effect requires 10 flops/px/frame, then evaluating 10 options for 10s clip is **50 Gflops** (today’s high performance PC: <5 Gflops/s)

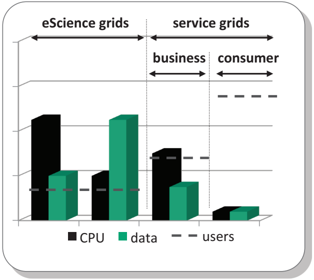
Online gaming:
e.g. Final Fantasy XI: 1.500.000 gamers

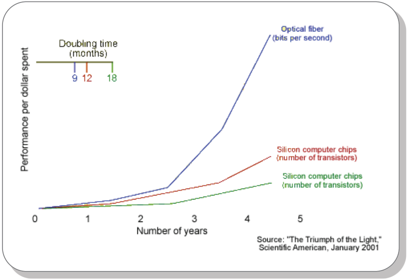
*Virtual reality: rendering of 3*10^8 polygons/s → 10^4 GFlops*

Multimedia editing

Why Optics

- Grid opportunities ranging from academia over corporate business to home users
- Optical data speeds ≥ internal PC bus speeds ⇝ remove the electronics bottleneck
- Today: Light speed (slowed only by routing and switching) ⇝ a bottleneck in fastest networks, e.g., finance networks



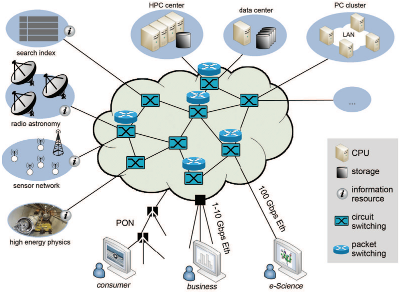


Source: "The Triumph of the Light," Scientific American, January 2001

Today: Toward Optical Grid and Cloud Computing

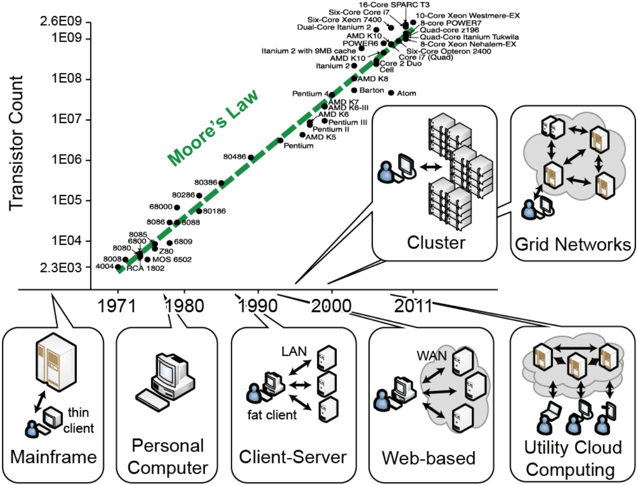
Optical networks crucial for increasingly demanding cloud services, e.g.,

- Computing:
  - High energy physics
  - Amazon EC2, Microsoft Azure
- Online storage:
  - Dropbox, Google Drive, etc.
- Collaboration tools:
  - MSOffice 365, Google Doc
- Video streaming:
  - Netflix, YouTube



C. Develder, et al., *"Optical networks for grid and cloud computing applications"*, Proc. IEEE, Vol. 100, No. 5, May 2012, pp. 1149-1167.

A historical perspective ...

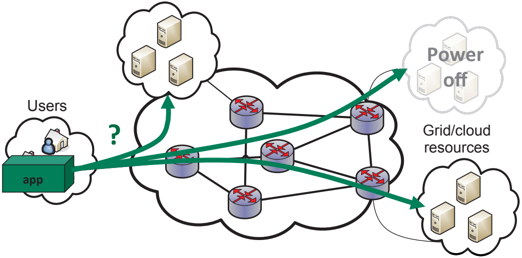


Cluster vs. Grid vs. Cloud Computing

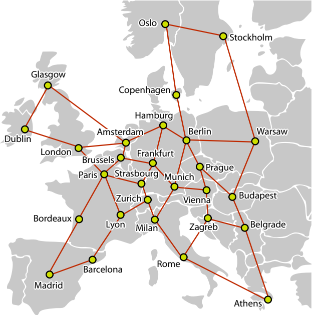
- Cluster Computing: group of parallel and distributed systems, which consists of a collection of inter-connected stand-alone computers working together as a single integrated computing.
- Grid Computing: set of different heterogeneous resources which include three units: computational (and data processing), storage, and networking, which are geographically spread over different domains.
- Cloud Computing: Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Anycast

- Users do (in general) **NOT** care where applications are served
  - E.g., virtual machines in IaaS can be instantiated anywhere
  - E.g., bag of tasks grid jobs can be run at any server, or distributed on any set of servers.



Case study set-up



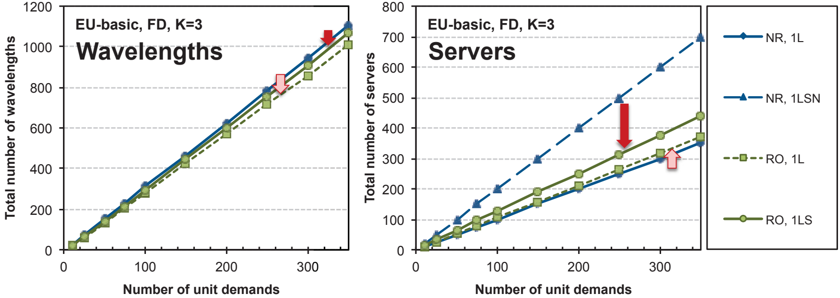
- Topology
  - European network
  - 28 nodes and 41 bidirectional links
- Demand
  - Randomly generated requests (10-350)
  - 10 instances for each number of requests
- **Four scenarios:**

	No relocation		Exploiting relocation
Single <u>link</u> failures:	***1L, NoReloc***	⇔	***1L, Reloc***
Single failures of either <u>link or server</u>:	***1LSN, NoReloc***	⇔	***1LS, Reloc***

C. Develder, J. Buysse, B. Dhoedt and B. Jaumard, *"Joint dimensioning of server and network infrastructure for resilient optical grids/clouds"*, IEEE/ACM Trans. Netw., Vol. PP, Oct. 2013, pp. 1-16. doi:10.1109/TNET.2013.2283924

The impact of relocation

- Single Link failures (1L):
  - Reduction of backup wavelengths
  - Slight increase in server capacity
- Single link/server failure (1LS)
  - Reduction of backup wavelengths
  - Fewer servers than 1:N server protection (N=1)



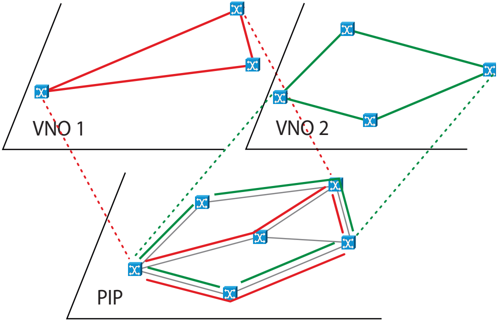
Manycast

- Manycast is a group communication scheme that enables communication with an arbitrary (user specified) number of group members.
- Anycast and Multicast communication are special cases of Manycast in which the target number of group members is one and infinity, respectively.

Virtualization

Network virtualization

Physical network is logically partitioned in isolated virtual networks



- **Virtual Network Operators (VNO)** operate logically separated networks
- **Physical Infrastructure Providers (PIP)** have full control over infrastructure (fibers, OXCs)

J.A. García-Espín, et al., *"Logical Infrastructure Composition Layer: the GEYSERS holistic approach for infrastructure virtualisation"*, in Proc. TERENA Networking Conference (TNC 2012), Reykjavík, Iceland, 21-24 May 2012.

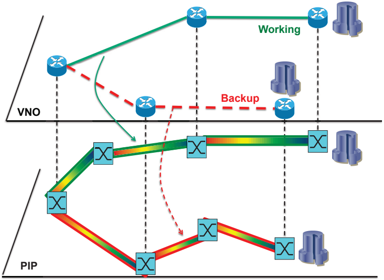
Why virtualization

- Existence of multiple stakeholders with conflicting goals and policies ⇝ alterations to the existing Internet architecture are now limited to simple incremental updates
- Network virtualization: A networking environment that allows coexistence of multiple virtual networks on the same physical substrate.
- Virtual network (VN) in a network virtualization environment (NVE) is a collection of virtual nodes and virtual links, i.e., a subset of the underlying physical network resources.

N.M.M.K. Chowdhury and R. Boutaba, A survey of network virtualization, *Computer Networks* 54 (2010) 862876.

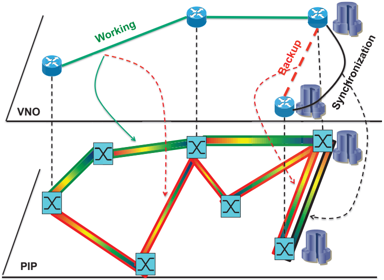
VNO-resilience

- Protection handled by the VNO
- Requests rerouted in the virtual network for both physical network failures and DC failures



PIP-resilience

- A virtual link is mapped resiliently to two failure-disjoint paths in the physical network



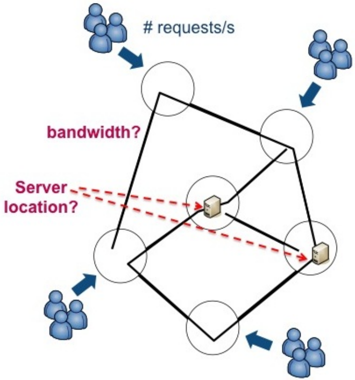
M. Bui, B. Jaumard, and C. Develder, Anycast end-to-end resilience for cloud services over virtual optical networks (invited), ICTON 2013.

I. Barla, D. Schupke, M. Hoffmann, and G. Carle, Optimal design of virtual networks for resilient cloud services, DRCN 2013.

Location of the DC/servers

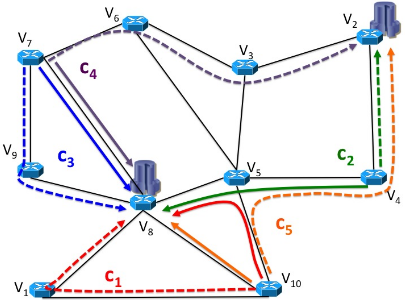
Problem Statement

- Given:
    - Network topology
    - Anycast requests
- Find
    - Locations of servers
    - Request provisioning
    - Backup paths



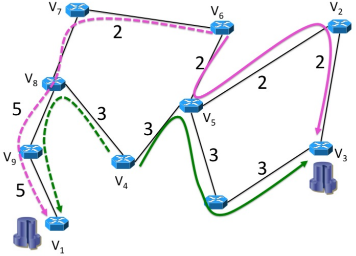
Model I: Configuration examples
Single link failure scenario

- Each configuration
    - One source node v
    - One primary path rooted at v
    - One backup path rooted at v (link/node disjoint)
- Selection of several configurations per source node (multiple primary/backup paths)

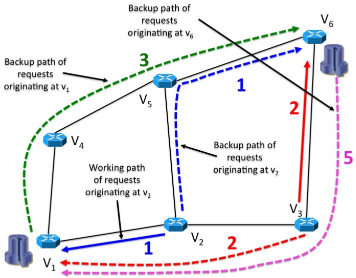


Model I: Configurations for single node or single server failure scenarios

Single node protection



Single server protection


Backup path of requests originating at v6

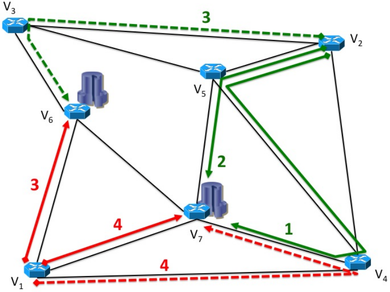
Backup path of requests originating at v1

Backup path of requests originating at v2

Working path of requests originating at v2

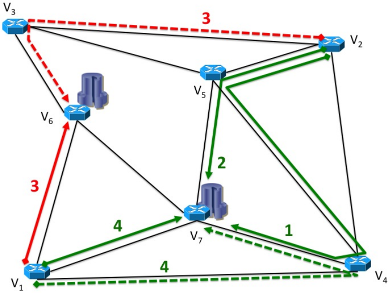
Model II: Configuration examples

- Each configuration
  - One source node v
  - A set of primary paths rooted at v
  - A single backup path rooted at v
- Selection of one configuration per source node
- For a given source node
  - Multiple primary paths
  - A single backup path

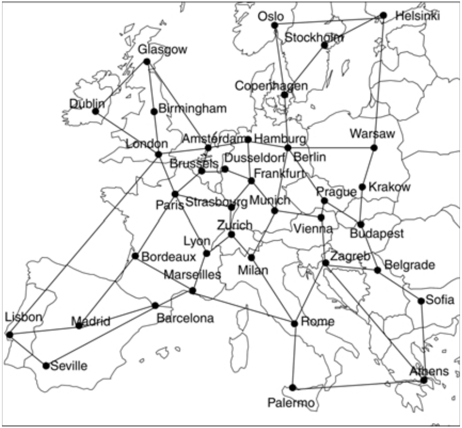


Model III: Configuration examples

- Each configuration
  - One server node v_s
  - A set of primary paths terminated at v_s
  - A set of backup paths terminated at v_s
- Selection of one configuration per server node
- For a given source node
  - Multiple primary paths
  - A single backup path

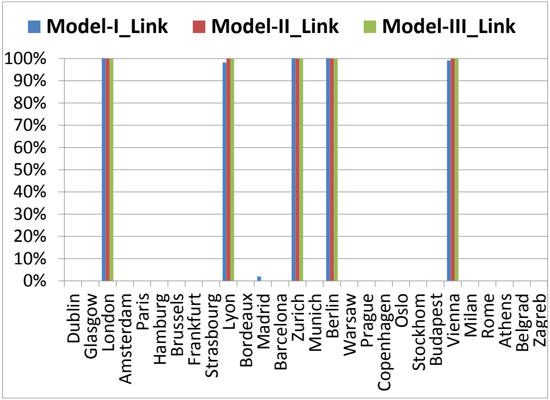


Numerical Results - European Network



Selection of server nodes under link protection scheme:

mostly the same servers

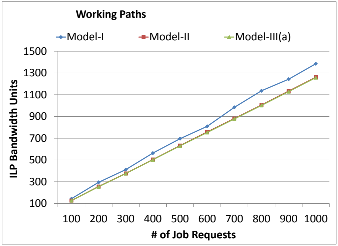


Comparison of the bandwidth requirements of the three models

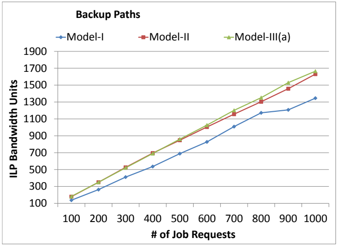
Mostly the same for Models II & III

Model I: Larger working, smaller backup

Working bandwidth requirements



Backup bandwidth requirements



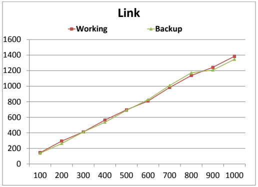
ILP solution for working and backup bandwidth units

Model I (5 Servers)

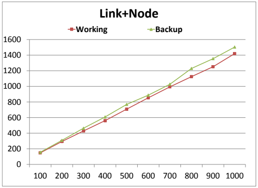
Similar for link & link + node protection

More backup bandwidth is needed when adding server protection

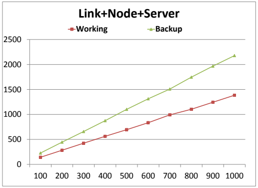
Link



Link+Node



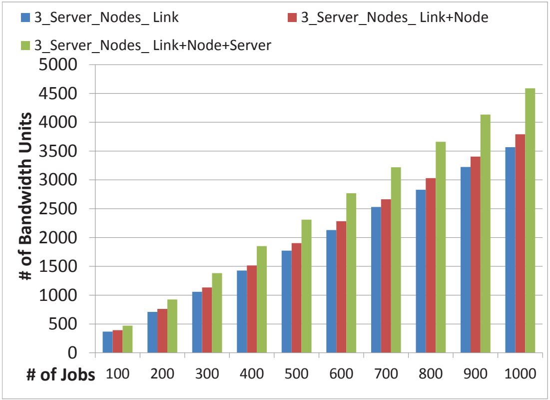
Link+Node+Server



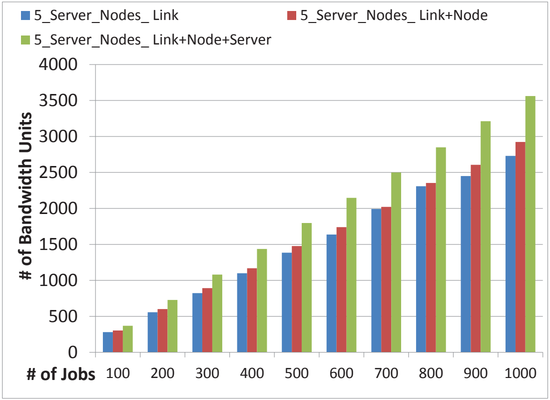
Different protection schemes - Model I - 3 server nodes

Similar for Link and Link + Node

30 % relative difference for Link+Node+Server



Different protection schemes - Model I - 5 server nodes



Model I - Working paths

Single link protection scheme (5 servers)

After basic pattern, largest percentages ⇝ 2 different paths toward the same destination

# of destinations	**1**	1	1	1	2	2	3	2	3	4	Σ
# of paths	**1**	0	2	3	2	3	3	4	4	4	
Size	%										
100	**58.3**	17.9	11.5	1.2	6.3	2.8	0.0	2.0	0.0	0.0	100
200	**51.2**	17.9	12.3	2.0	10.3	6.0	0.0	0.4	0.0	0.0	100
300	**53.2**	17.9	15.0	1.8	9.6	2.1	0.0	0.4	0.0	0.0	100
400	**42.5**	17.9	16.3	5.2	8.7	6.3	0.4	2.0	0.4	0.4	100
500	**42.1**	17.9	19.4	4.4	9.1	4.0	0.8	0.8	0.8	0.8	100
600	**45.7**	17.9	16.8	4.3	7.9	5.0	0.0	1.8	0.4	0.4	100
700	**50.4**	17.9	15.6	0.9	9.4	3.6	0.4	1.8	0.0	0.0	100
800	**62.5**	17.9	9.5	0.0	5.4	2.4	0.0	1.2	0.6	0.6	100
900	**49.3**	17.9	15.7	1.4	9.6	3.6	0.4	1.4	0.4	0.4	100
1000	**31.1**	17.9	8.6	0.4	23.6	6.8	4.6	0.7	3.2	3.2	100
AVERAGE	**48.6**	**17.9**	**14.1**	**2.1**	**10.0**	**4.3**	**0.7**	**1.2**	**0.6**	**0.6**	100

Model I - Working paths under Single Link or Node Failure (5 servers)

# of destinations	**1**	1	1	1	2	2	3	2	3	4	Σ
# of paths	**1**	0	2	3	2	3	3	4	4	4	
Size	%										
100	**54.3**	17.9	7.9	2.1	10.4	6.1	0.4	1.1	0.0	0.0	100
200	**47.9**	17.9	11.4	2.9	11.1	6.8	0.4	1.8	0.0	0.0	100
300	**45.4**	17.9	15.4	1.8	9.3	6.1	0.7	3.6	0.0	0.0	100
400	**45.4**	17.9	12.1	1.8	12.5	7.1	0.7	1.8	0.4	0.4	100
500	**42.9**	17.9	16.1	3.9	11.8	6.1	0.0	0.7	0.4	0.4	100
600	**42.5**	17.9	12.1	2.9	15.4	7.5	0.4	0.7	0.4	0.4	100
700	**40.4**	17.9	14.3	5.0	13.6	5.0	1.1	2.9	0.0	0.0	100
800	**41.3**	17.9	16.7	2.4	12.7	7.1	0.8	0.4	0.4	0.4	100
900	**43.2**	17.9	12.9	2.1	13.2	6.8	0.4	2.1	0.7	0.7	100
1000	**37.7**	17.9	8.1	0.0	25.3	5.8	2.6	0.6	1.0	1.0	100
AVERAGE	**44.1**	**17.9**	**12.7**	**2.5**	**13.5**	**6.4**	**0.7**	**1.6**	**0.3**	**0.3**	100

Model I - Working paths
Link/node/server node protection scheme (5 servers)

# of destinations	1	1	1	1	2	2	3	2	3	4	$\sum$
# of paths	1	0	2	3	2	3	3	4	4	4	
Size	%										
100	**64.6**	17.9	10.7	0.4	5.7	0.7	0.0	0.0	0.0	0.0	100
200	**61.8**	17.9	13.6	0.7	5.7	0.4	0.0	0.0	0.0	0.0	100
300	**55.7**	17.9	12.9	1.1	10.4	2.1	0.0	0.0	0.0	0.0	100
400	**57.5**	17.9	12.1	0.7	7.5	3.2	0.4	0.0	0.4	0.4	100
500	**55.0**	17.9	13.6	0.7	10.7	1.4	0.4	0.4	0.0	0.0	100
600	**58.2**	17.9	13.9	0.7	8.6	0.0	0.4	0.4	0.0	0.0	100
700	**56.8**	17.9	16.8	0.0	7.1	1.4	0.0	0.0	0.0	0.0	100
800	**59.1**	17.9	13.5	0.0	7.1	1.6	0.4	0.4	0.0	0.0	100
900	**59.3**	17.9	13.9	0.4	7.1	1.4	0.0	0.0	0.0	0.0	100
1000	**49.0**	17.9	5.5	0.3	24.0	0.6	1.9	0.0	0.3	0.3	100
AVERAGE	**57.7**	**17.9**	**12.7**	**0.5**	**9.4**	**1.3**	**0.3**	**0.1**	**0.1**	**0.1**	100

Backup: one vs. multiple paths

- Largest percentages: 2 different paths with two different destinations, if we exclude the basic pattern
- Single link and single link/node do not differ much with respect to the number of bandwidth units (3%) as well as with respect to the ratio of multiple paths (6%).

Conclusions

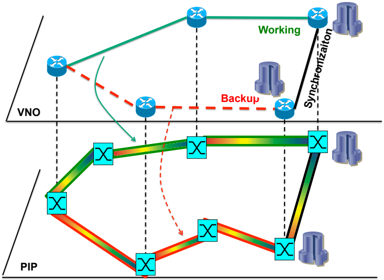
- Three different decomposition models, with Model III being the straightforward extension of the model for p-center problem
- Model I, the most simple one $\rightsquigarrow$ the most efficient
- In addition, Model I offers the largest flexibility
- Best location of the servers
    - is independent of the number of backup paths
    - offers a compromise between distribution of the server locations and availability of short and highly sharable backup paths

VNO vs. PIP resilience & Synchronization

Bui, M., B. Jaumard and C. Develder, Scalable Algorithms for QoS Aware Virtual Network Mapping for Cloud Services, ICC 2014

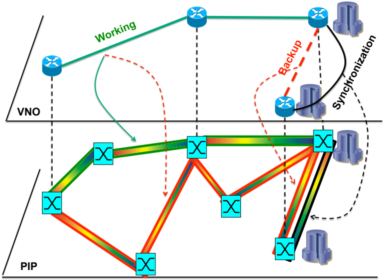
VNO-resilience

- Protection is handled by the virtual network operator
- Requests are rerouted in the virtual network both in case of physical network failure and DC failure



PIP-resilience

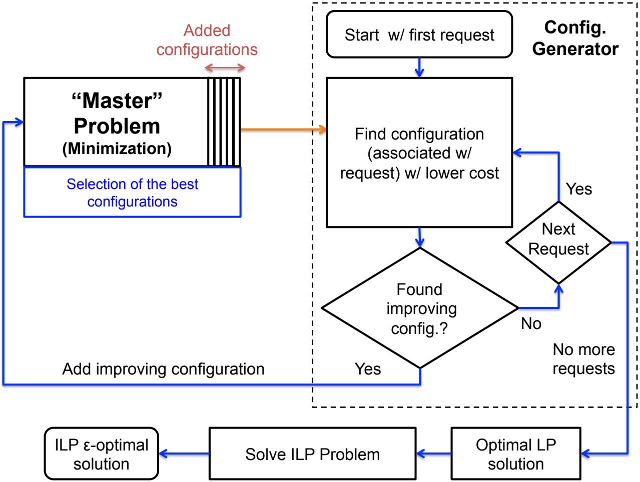
- A virtual link is mapped resiliently to two failure-disjoint paths in the physical network.



M. Bui, B. Jaumard, and C. Develder, Anycast end-to-end resilience for cloud services over virtual optical networks (invited), ICTON 2013.

I. Barla, D. Schupke, M. Hoffmann, and G. Carle, Optimal design of virtual networks for resilient cloud services, DRCN 2013.

Flowchart - Solution Mechanism



VNO Optimization Model

$$\min \quad \sum_{\ell \in L} (\beta_\ell^{\mathrm{W}} + \beta_\ell^{\mathrm{B}}) \cdot \|\ell\|,$$

subject to:

$$\sum_{c \in C} \alpha_k^c \, z_c \geq 1 \qquad k \in K \qquad \text{Demand} \tag{1}$$

$$\sum_{c \in C} \Delta_{k_c} \left(p_{\ell,c}^{\mathrm{W}} + \delta_k \, p_{\ell,c}^{\mathrm{S}} \right) z_c = \beta_\ell^{\mathrm{W}} \qquad \ell \in L \qquad \text{W. bandwidth} \tag{2}$$

$$\sum_{c \in C} \Delta_{k_c} \, p_{\ell',c}^{\mathrm{W}} \, p_{\ell,c}^{\mathrm{B}} \, z_c \leq \beta_\ell^{\mathrm{B}} \qquad \ell' \in L, \ell \in L \setminus \{\ell'\} \qquad \text{B. bandwidth/links} \tag{3}$$

$$\sum_{c \in C} \Delta_{k_c} \, a_{v,c}^{\mathrm{W}} \, p_{\ell,c}^{\mathrm{B}} \, z_c \leq \beta_\ell^{\mathrm{B}} \qquad v \in V_{\mathrm{D}}, \ell \in L \qquad \text{B. bandwidth/nodes} \tag{4}$$

$$z_c \in \{0,1\} \qquad c \in C \tag{5}$$

$$\beta_\ell^{\mathrm{W}}, \; \beta_\ell^{\mathrm{B}} \in \mathbb{R} \qquad \ell \in L. \tag{6}$$

PIP Optimization Model

Replace constraints (4) with (7). Remark that S will need to support the full request bandwidth when a node failure occurs at the primary data center (but it can be shared among different failure cases):

$$\sum_{c \in C} \Delta_{k_c}\, a^{\mathrm{W}}_{v,c}\, p^{\mathrm{S}}_{\ell,c}\, z_c \leq \beta^{\mathrm{B}}_{\ell} \qquad v \in V_{\mathrm{D}}, \ell \in L. \tag{7}$$

Note that the synchronization bandwidth on the S path will be reserved on top of that (see (2) in the master problem).

Configuration Generation: The path and data center variables have to obey:

$$\overline{\text{COST}} = 0 - \sum_{\ell \in L} u_\ell^{(2)}\, \Delta_{k_c} \left(p^{\text{W}}_{\ell,c} + \delta_k\, p^{\text{S}}_{\ell,c} \right) - \sum_{k \in K} u_k^{(1)}\, \alpha_k$$

$$- \sum_{\ell \in L} \sum_{\ell' \in L \setminus \{\ell\}} u^{(3)}_{\ell\ell'}\, \Delta_k p^{\text{W}}_\ell\, p^{\text{B}}_{\ell'} - \sum_{v \in V_{\text{D}}} \sum_{\ell \in L} u^{(4)}_{v\ell} \Delta_k a^{\text{W}}_v p^{\text{B}}_\ell \tag{8}$$

$$\sum_{\ell \in \omega(v)} p^{\text{W}}_\ell = \begin{cases} 1 & \text{if } v = v_s \\ 2\, d^{\text{W}}_v - a^{\text{W}}_v & \text{otherwise} \end{cases} ; \quad \sum_{\ell \in \omega(v)} p^{\text{B}}_\ell = \begin{cases} 1 & \text{if } v = v_s \\ 2\, d^{\text{B}}_v - a^{\text{B}}_v & \text{otherwise} \end{cases} \qquad v \in V \tag{9}$$

$$\sum_{\ell \in \omega(v)} p^{\text{S}}_\ell = 2\, d^{\text{S}}_v - a^{\text{W}}_v - a^{\text{B}}_v \qquad v \in V \tag{10}$$

$$p^{\text{W}}_\ell + p^{\text{B}}_\ell \leq 1 \qquad \ell \in L \quad ; \quad a^{\text{W}}_v + a^{\text{B}}_v \leq 1 \qquad v \in V_{\text{D}} \tag{11}$$

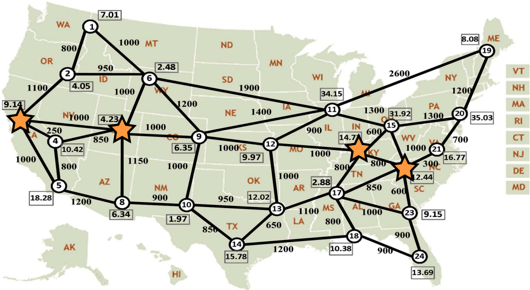
$$\sum_{v \in V_{\text{D}}} a^{\text{W}}_v = 1\,; \quad \sum_{v \in V_{\text{D}}} a^{\text{B}}_v = 1\,; \quad \sum_{v \notin V_{\text{D}}} a^{\text{W}}_v + a^{\text{B}}_v = 0 \tag{12}$$

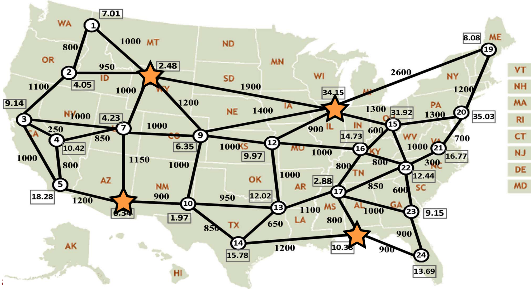
$$a^{\text{W}}_v, a^{\text{B}}_v \in \{0,1\} \qquad v \in V \quad ; \quad p^{\text{W}}_\ell, p^{\text{B}}_\ell, p^{\text{S}}_\ell \in \{0,1\} \qquad \ell \in L. \tag{13}$$

Numerical Results

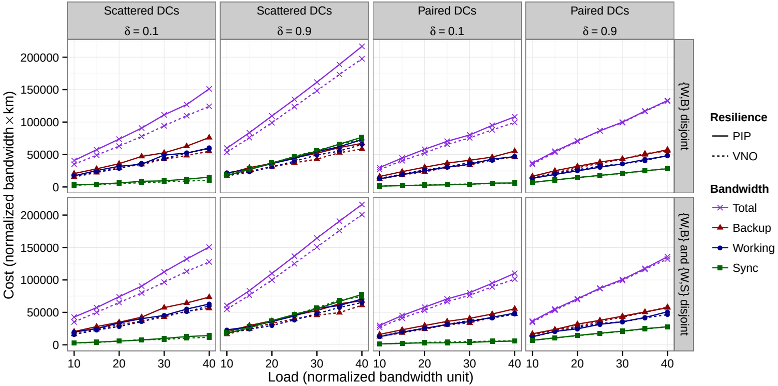
- 24-node US nationwide backbone network with 4 data centers
- Bandwidth requirement for each service request is generated randomly with uniform distribution between 0 and 1 normalized bandwidth units
- Uniform traffic, i.e., the source node of a request is chosen randomly, and vary the total requested bandwidth (i.e., the total load) from 10 to 40 units (the number of generated requests varied from 22 to 83).
- Each request is individually provisioned: requests originating from the same source node are not forced to follow the same paths towards the same data centers.
- Two sets of DC locations
    - Scattered: DCs spread fairly uniformly: {WY(6), AZ(8), IL(11), AL(18)}.
    - Paired: DCs in paired locations: {CA(3), UT(7), KY(16), NC(22)}.
- Low vs high synchronization bandwidth: $\delta = 0.1$ and $\delta = 0.9$.

US topology





US topology: costs vs. traffic load



Conclusions

- Scalable models to find routings and DC allocations for cloud requests, with minimal cost, for the VNO vs PIP resilience frameworks and options for the synchronization path (one or two disjoint ones)
- Intuitively expected advantage of VNO resilience actually can be quite limited, when DCs occur in paired configurations (which may be desirable to obtain similar latencies towards both primary and backup DC)
- If the synchronization bandwidth becomes a substantial fraction of the actual traffic bandwidth, relative cost advantage becomes very limited.

QoS-aware virtual network mapping for cloud services

Bui, M., B. Jaumard, I.B. Barla, and C. Develder, Resilience options for provisioning anycast cloud services with Virtual Optical Networks, ONDM 2014

Statement of the problem

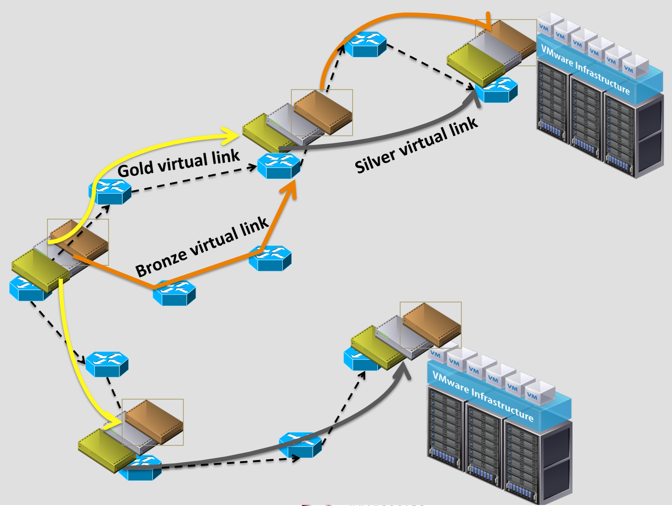
Given

- Network topology:
  - $G^{\text{PHY}} = (V^{\text{PHY}}, L^{\text{PHY}})$, $G^{\text{VIR}} = (V^{\text{VIR}}, L^{\text{VIR}})$.
  - $V^{\text{DC}} \subseteq V^{\text{VIR}}$ – the set of data center locations.
- Cloud requests $d \in D$, each one characterized by:
  - Source node $\text{SRC}_d \in V^{\text{VIR}}$
  - Requested bandwidth Δ_d^{BW}
  - Requested number of virtual machines Δ_d^{VM}
  - Minimal QoS class of the VMs, $q_d \in Q$
  - Maximal end-to-end delay δ_d

Find: For each request d, a working (W) and backup (B) data center and routes in the virtual network G^{VIR}, such that d can always be served under any failure scenario with minimum total network cost. subject to:

- Physical network capacity constraints
- Demand QoS requirements

An Illustrative Example



Objective function

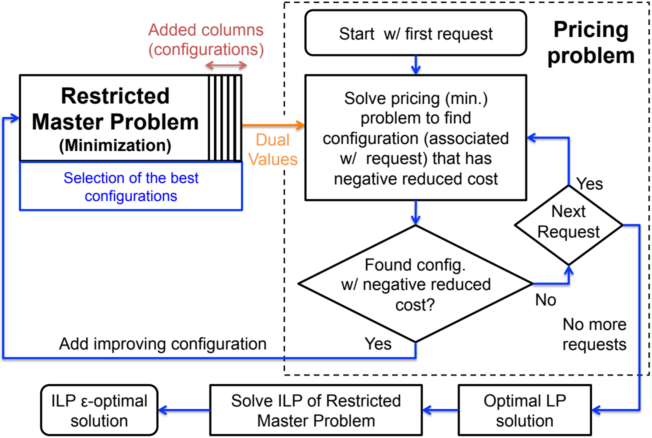
$$\min \quad \sum_{\gamma\in\Gamma} \text{COST}_{\gamma}\; z_{\gamma} + \sum_{\ell'\in L^{\text{VIR}}} \text{C\_SETUP}^{\text{LINK}}_{\ell'}\; x^{\text{LINK}}_{\ell'} + \sum_{v'\in V^{\text{VIR}}} \sum_{q\in Q} \text{C\_SETUP}^{\text{NODE},q}\; x^{\text{NODE},q}_{v'}$$

$$\underbrace{\text{COST}_{\gamma}}_{\text{Cost of configuration } \gamma} = \Delta_d^{\text{BW}} \underbrace{\Bigg[\sum_{\ell'\in V^{\text{VIR}}\times V^{\text{VIR}}} \sum_{q\in Q} \text{C}^{\text{LINK},q} \left(y^{\text{W},q,\gamma}_{\ell'} + y^{\text{B},q,\gamma}_{\ell'} \right) +}_{\text{Unit cost of virtual links}} \underbrace{\sum_{v\in V^{\text{VIR}}} \sum_{q\in Q} \text{C}^{\text{NODE},q}\; y^{\text{NODE},q,\gamma}_{v} \Bigg]}_{\text{Unit cost of virtual nodes}} + \Delta_d^{\text{VM}} \underbrace{\Bigg[\sum_{v\in V^{\text{DC}}} \sum_{q\in Q} \text{C}^{\text{VM},q}\; y^{\text{VM},q,\gamma}_{v} \Bigg]}_{\text{Unit VM cost}}$$

Constraints

$\sum_{\gamma \in \Gamma_d} z_\gamma \geq 1$	$d \in D$	**Demand**
$M \, x_{\ell'}^{\text{LINK}} \geq \sum_{\gamma \in \Gamma} p_{\ell'}^{\gamma} \, z_\gamma$	$\ell' \in L^{\text{VIR}}$	**# of distinct virtual link maps**
$M \, x_v^{\text{NODE},q} \geq \sum_{\gamma \in \Gamma} y_v^{\text{NODE},q,\gamma} \, z_\gamma$	$v \in V^{\text{VIR}}, q \in Q$	**g, s, b nodes**
$\text{CAP}_\ell^{\text{LINK}} \geq \sum_{\gamma \in \Gamma} \Delta_\gamma^{\text{BW}} \, p_\ell^{\gamma} z_\gamma$	$\ell \in L^{\text{PHY}}$	**Link Capacity**
$\text{CAP}_v^{\text{NODE}} \geq \sum_{\gamma \in \Gamma} \sum_{q \in Q} \Delta_\gamma^{\text{BW}} y_v^{\text{NODE},q,\gamma} z_\gamma$	$v \in V^{\text{VIR}}$	**Node Capacity**
$\text{CAP}_v^{\text{VM}} \geq \sum_{\gamma \in \Gamma} \sum_{q \in Q} \Delta_\gamma^{\text{VM}} \, \text{CAP}^{\text{VM},q} \, y_v^{\text{VM},q,\gamma} \, z_\gamma$	$v \in V^{\text{DC}}$	**VM Capacity**

Decomposition flow chart



Added columns (configurations)

Restricted Master Problem (Minimization)

Selection of the best configurations

Dual Values

Pricing problem

Start w/ first request

Solve pricing (min.) problem to find configuration (associated w/ request) that has negative reduced cost

Found config. w/ negative reduced cost?

Yes

No

Next Request

Yes

No more requests

Add improving configuration

Optimal LP solution

Solve ILP of Restricted Master Problem

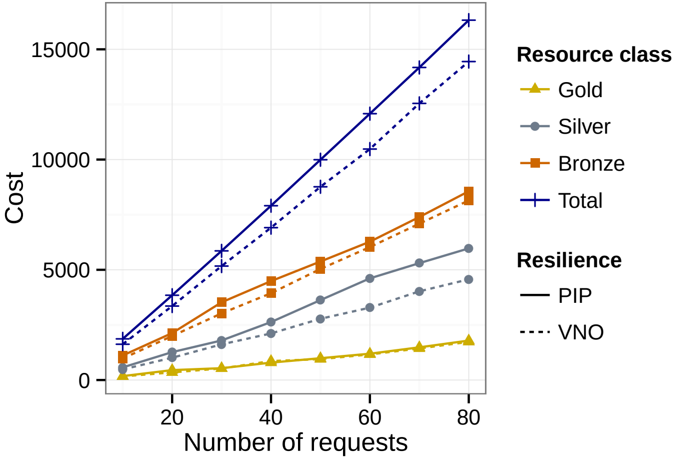
ILP ε-optimal solution

Network instance

Cost parameters

Parameters	Cost
Virtual node setup cost (gold, silver, bronze)	10, 6, 4
Virtual node bandwidth unit cost (gold, silver, bronze)	5, 3, 2
Virtual link setup cost (gold, silver, bronze)	10, 6, 4 + 10 × physical hopcount
Virtual link bandwidth unit cost (gold, silver, bronze)	5, 3, 2
Virtual machine unit cost (gold, silver, bronze)	5, 3, 2
Delay of a physical node	1
Delay of a physical link	1
Delay of a logical node (gold, silver, bronze)	2, 3, 5
Capacity of a virtual machine (gold, silver, bronze)	5, 3, 2

Results - 10% gold, 30 % silver, 60 % bronze



Comments

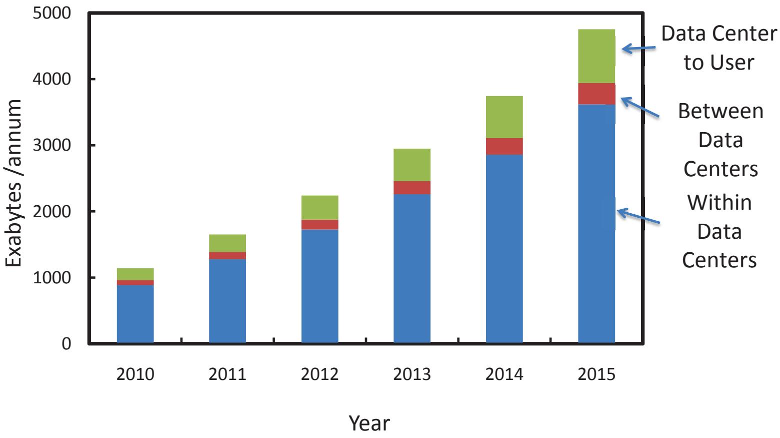
- Overall cost of the PIP scheme is about 10% higher than the cost of the VNO scheme: VNO model has a greater flexibility for selecting DC than the PIP model
- The major difference stems from the Silver: Bronze links are high delay and hence less likely to be feasible to reuse. Gold links are to keep the delay under control and hence there are few opportunities to split them without violating the delay requirement.

Energy Issues

Forecast Overview

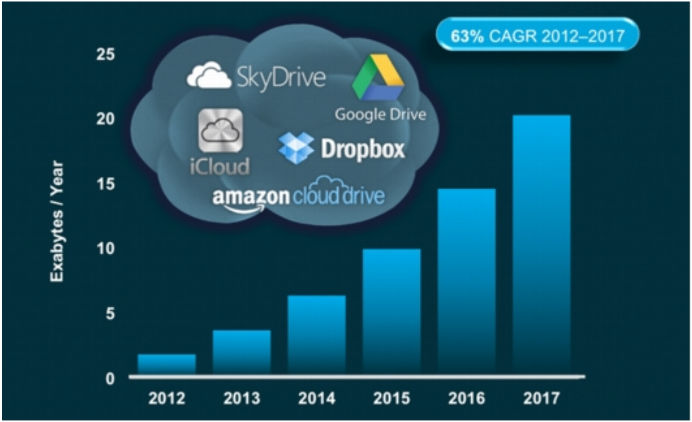
- **Global data center traffic**
    - Global data center IP traffic will nearly triple over the next 5 years. Overall, data center IP traffic will grow at a compound annual growth rate (CAGR) of 25 percent from 2012 to 2017.
- **Global cloud traffic**
    - Global cloud IP traffic will increase nearly 4.5-fold over the next 5 years. Overall, cloud IP traffic will grow at a CAGR of 35 percent from 2012 to 2017
    - Global cloud IP traffic will account for more than two-thirds of total data center traffic by 2017

Projections of data centre traffic



Source : Cisco Cloud Index 2011

Personal Content Locker Traffic Growth



Why does energy matter?

- If nothing is done to address the growing efficiency gap:
  - ICT will consume ever larger proportion of global energy
  - Energy consumption could become a barrier to network growth
- Economic and engineering imperatives:
  - Energy is a growing component of OPEX
  - Increased energy consumption ⇝ increased footprint

Estimating ICT power consumption

- Inventory based estimates
  - Look at what is out there
    - Use sales and deployment data from vendors and surveys
    - Accessing accurate data is problematic
- Network design and dimensioning based estimates
  - Design a network that will satisfy current and projected demands
    - Use typical network design rules
    - Difficult to include network inefficiencies, overlays & legacies
- Transaction based estimates
  - Look at services required and design a network to provide them
    - Similar to network design approach

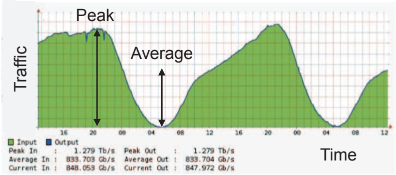
Summary of estimates

Author	Year	% national electricity use	Country	PC's, office equip. & servers	Wireless access	Notes
Huber	1999	13%	USA	Yes	No	Severe over estimate
Koomey	1999	2%	USA	Yes	No	Users & equipment estimate
Kawatomo	2001	3%	USA	Yes	No	Users & equipment estimate
Turk	2001	0.5 - 1.7%	Germany	Yes	No	Users & equipment estimate
Barthel	2001	0.9 – 1.5%	Germany	Yes	Yes	Users & equipment estimate
Roth	2002	< 2.3%	USA	Yes	Yes	Users & equipment estimate
Cremer	2003	7.1%	Germany	Yes	Yes	Users & equipment estimate
Baliga	2007	0.5%	OECD	No	No	Network design & dimensioning
Vereecken	2010	Not given	Not given	No	No	Network design & dimensioning
Lange	2010	Not given	Not given	No	Yes	Network design & dimensioning
Kilper	2011	Not given	USA	No	Yes	Transaction
Pickavet	2007 2012		Global	Yes	Yes	Users and equipment estimate

Design and dimensioning approach (Baliga *et al.*, 2007)

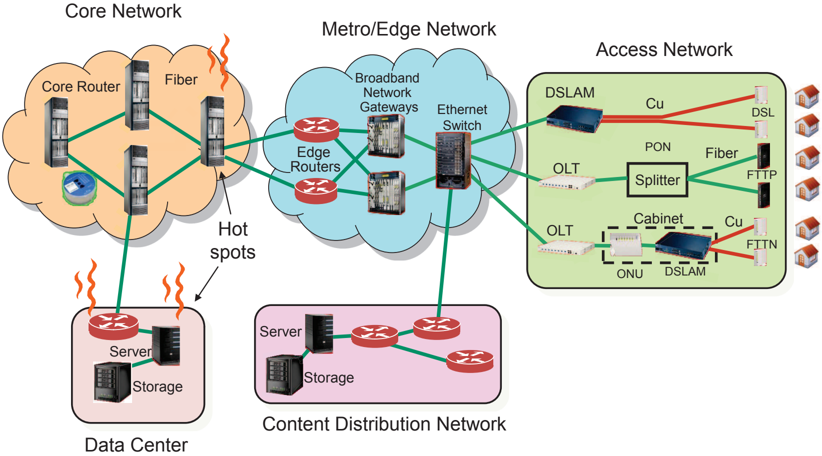
1. Split network into
    - Access
    - Metro/Edge
    - Core
    - Data centres, content storage
2. Model with representative architecture and equipment
3. Dimension network to accommodate expected traffic
4. Calculate power consumption per customer for network

Key Parameter Models

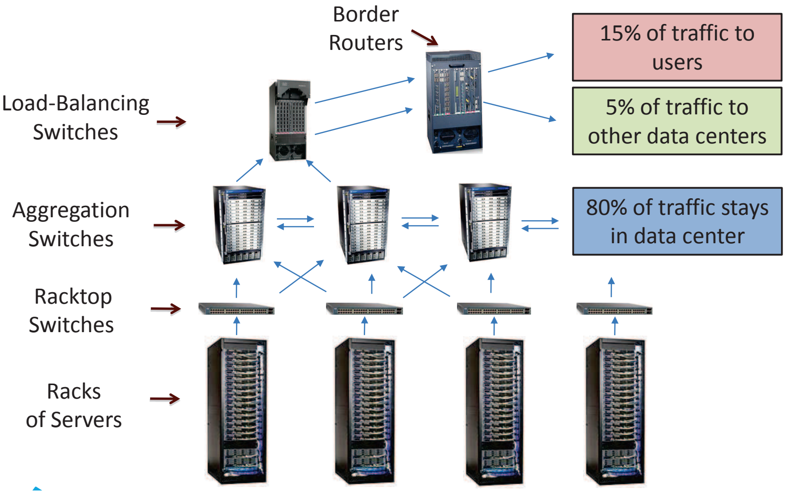


- Peak vs average access speed
  - Contention & aggregation
- Network dimensioning
  - Traffic growth
    - Deployed capacity > demanded capacity
  - Equipment redundancy
    - Multi-homing , back-up storage
  - Service protection
    - 1 + 1, 1:1, 1:N protection
- Router hops between source and destination
- Data centre Power Usage Effectiveness (PUE)

Network segmentation



Data centers and content servers



Border Routers

Load-Balancing Switches

Aggregation Switches

Racktop Switches

Racks of Servers

15% of traffic to users

5% of traffic to other data centers

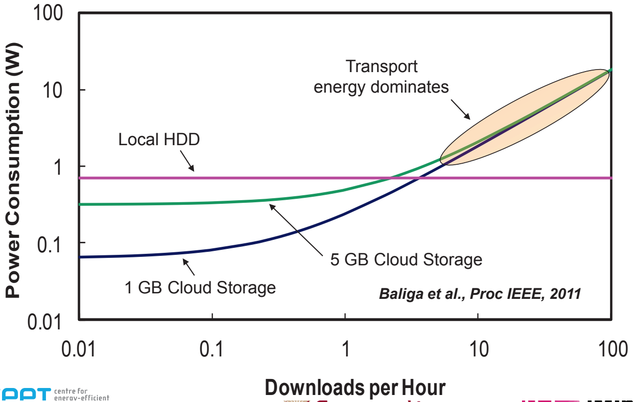
80% of traffic stays in data center

The Cloud

- Cloud services widely promoted as greener than on-site facilities:
  - Cloud Computing The IT Solution for the 21st Century
    - Carbon Disclosure Project Study 2011
  - Salesforce.com & the Environment
    - WSP Environment & Energy 2011
- Strong case for enterprise private cloud
- What about the public cloud?
  - Apple iCloud
  - Google drive
  - Microsoft sky drive

Example: Public storage as a service (SaaS)

Storage of application & data in the cloud compared with storing on a local disk. 50 MBytes per download. Modern laptop-style HDD 20% read/write and 80% idle.

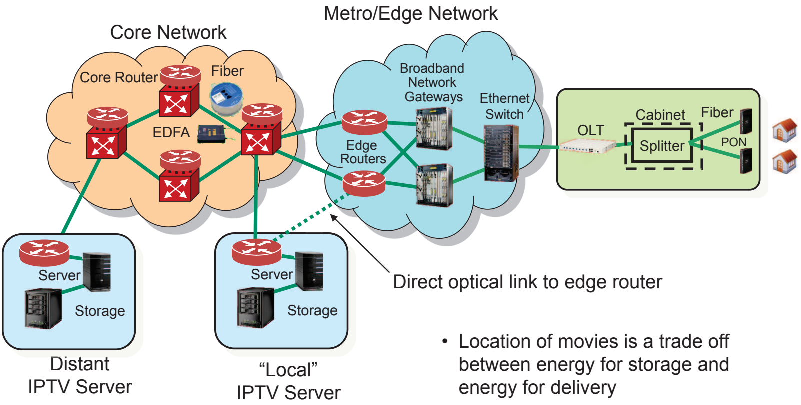


Baliga et al., Proc IEEE, 2011

Rethinking the Green Cloud

- Need to improve access energy efficiency
  - Small wireless cells
  - PON
- Keep some processing power in users device
- Reduce the number of router hops
  - Avoid public Internet
  - Use optical layer by-pass of routers
- Improve protocol efficiency
  - Less overhead bytes
  - Smart scheduling

IPTV over the public Internet



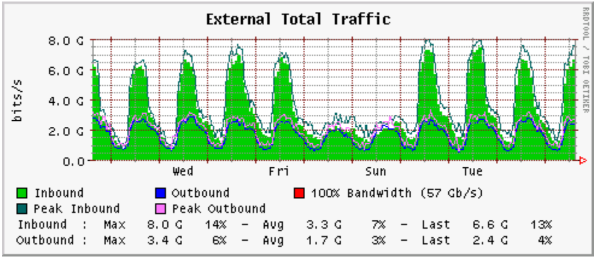
- Location of movies is a trade off between energy for storage and energy for delivery
- Each movie is replicated on R servers throughout the network

Energy-efficient routing

B. Puype et.al., *"Multilayer traffic engineering for energy efficiency"*, Photonic Netw. Commun., vol. 21, no. 2, Apr. 2011, pp. 127-140

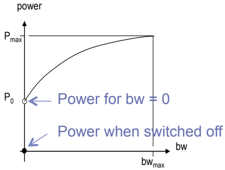
Diurnal traffic variations

- Night time volume is only a fraction of day time volume
  (20-30% typically; 20% for WiFi hotspot uplinks; 10% for smartphone network)
- Variation amplitude is increasing in recent years



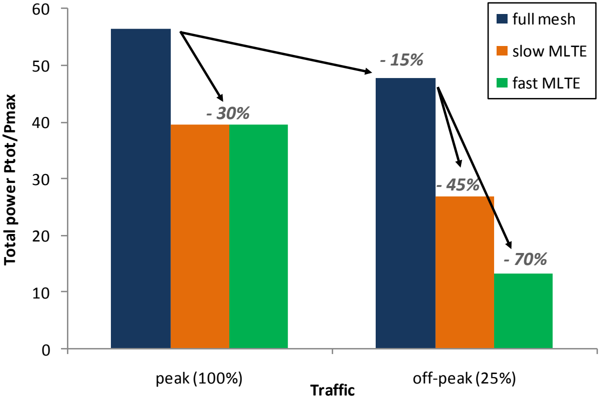
K.Cho et al., Observing Slow Crustal Movement in Residential User Traffic, Proc. ACM CoNext '08
M.Afanasyev et al., Analysis of a Mixed-Use Urban WiFi Network, Proc. IMC '08, 85-98

Power optimization techniques

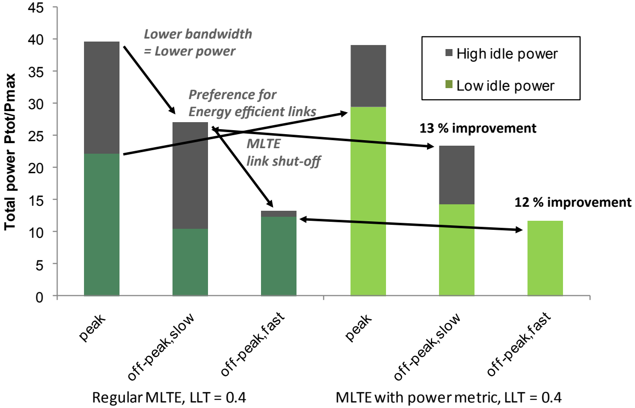


- Energy-efficient multilayer **traffic engineering** (MLTE)
  - Online routing of requests
  - Link metric reflects power characteristic
- Slow MLTE (days) = Logical topology configuration
  - Only set up required IP/MPLS links for traffic of that day
- Fast MLTE (hours) = Reconfiguration & shut-down if off-peak
  - Reroute traffic of shut-down links over remaining links

Effect of MLTE on power



Diurnal traffic variations



Conclusion for energy-efficient routing

- Opportunities for power savings:
  - Line-card power usage depends on line rate
  - Power scales with bandwidth use
  - Idle power maybe substantial part of max. power

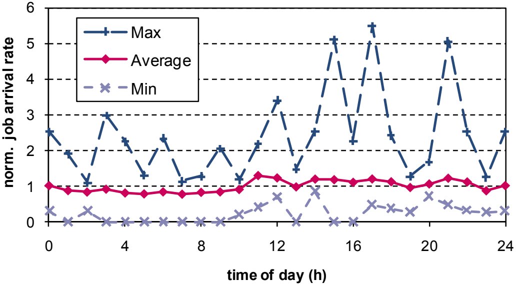
- MLTE = multilayer traffic engineering,
  - Optimizing for link utilization beneficial for power (obviously), but also explicit power optimization is possible *(see paper)*
  - **Even 'slow' MLTE offers substantial savings**
  - **'Fast' MLTE achieves further reduction** (but requires fast control plane)

Energy-efficient scheduling

C. Develder, M. Pickavet, B. Dhoedt and P. Demeester, *"A power-saving strategy for Grids"*, in Proc. 2nd Int. Conf. on Networks for Grid Applications (GridNets 2008), Bejing, China, 8-10 Oct. 2008.

Opportunities for “green grids”

- Fluctuations in server use in LCG Grid



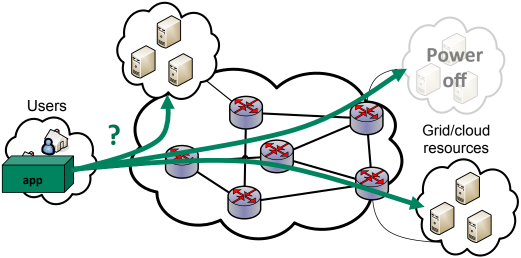
23 LCG Grid sites

$\lambda_{i,norm} = \lambda_i/\lambda_{avg}$
λ_{avg} = avg. over 24h

$\lambda_{max} = \max \lambda_{i,norm}$
$\lambda_{avg} = \text{avg } \lambda_{i,norm}$
$\lambda_{min} = \min \lambda_{i,norm}$

Anycast

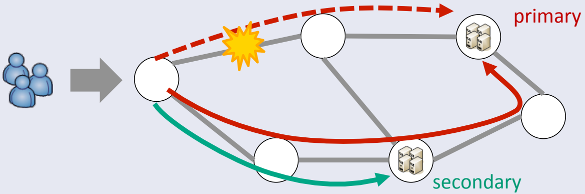
- Users do (in general) NOT care where applications are being served
  - E.g., virtual machines in IaaS can be instantiated anywhere
  - E.g., bag-of-tasks grid jobs can be run at any server



C. Develder, J. Buysse, M. De Leenheer, B. Jaumard and B. Dhoedt, *"Resilient network dimensioning for optical grid/clouds using relocation (Invited Paper)"*, in Proc. Workshop on New Trends in Optical Networks Survivability, at IEEE Int. Conf. on Commun. (ICC 2012), Ottawa, Ontario, Canada, 11 Jun. 2012.

Note: Anycast exploited for resilience = relocation

- Dimension optical grid/cloud so that it is resilient against failures?
- Exploit anycast principle: allow rerouting jobs to other destinations



C. Develder, J. Buysse, M. De Leenheer, B. Jaumard and B. Dhoedt, "Resilient network dimensioning for optical grid/clouds using relocation (Invited Paper)", in Proc. Workshop on New Trends in Optical Networks Survivability, at IEEE Int. Conf. on Commun. (ICC 2012), Ottawa, Ontario, Canada, 11 Jun. 2012.

A. Shaikh, J. Buysse, B. Jaumard and C. Develder, *"Anycast routing for survivable optical grids: scalable solution methods and the impact of relocation"*, IEEE/OSA J. Opt. Commun. Netw., Vol. 3, No. 9, Sep. 2011.

When to turn servers off?

Simple power-scheme:

- **Turn off if a time D after a job finishes**, the CPU it was running on is completely idle (i.e., all cores are idle)

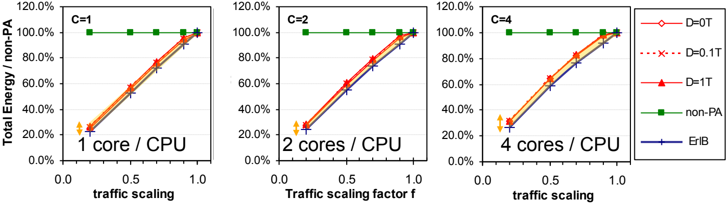
Influence of D?

- Poisson process with arrival rate λ:
  P[no arrivals in D] = $e^{-\lambda D}$
- Increasing D: chance we can turn it off will decrease exponentially

Baseline for power-saving strategy will be ErlangB: number of servers from ErlangB formula for max. loss of 5%

Total server energy

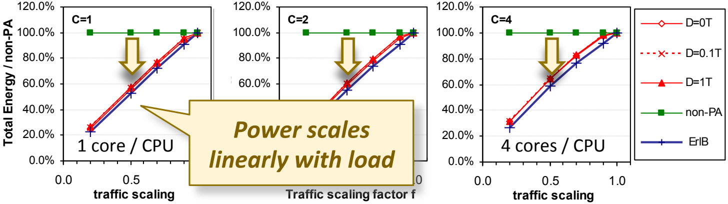
Discrepancy wrt ErlB due to powering on/off (possibly different) CPUs



- We achieve power close to theoretical lower bound of ErlangB
- Negligible influence of delay D on *total* power reduction

ErlangB baseline: Power consumption if same fixed number N_s of servers was always on, calculated for traffic $s \cdot \lambda$ and 5% loss (s = traffic scaling factor)

Total server energy

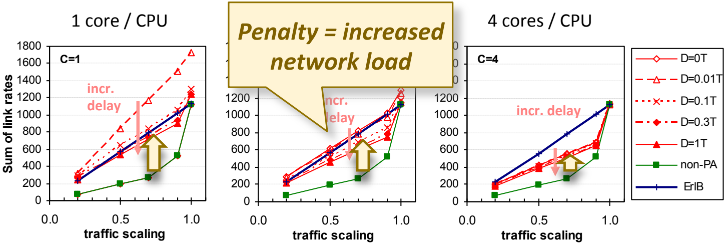


Power scales linearly with load

- We achieve power close to theoretical lower bound of ErlangB
- Negligible influence of delay D on *total* power reduction

ErlangB baseline: Power consumption if same fixed number N_s of servers was always on, calculated for traffic s·λ and 5% loss (s = traffic scaling factor)

Link bandwidth use



Penalty = increased network load

- Increasing delay = lower excess bandwidth
- Increasing # cores = lower excess bandwidth

Conclusion for energy-aware scheduling

- Opportunities for power savings:
  - Substantial variations in load
  - Idle power = large fraction of 100% loaded CPU
  - Limited power usage penalty for turning on server

- Simple power scheme:
  - Power usage close to that of system dimensioned for lower load
  - Limited **penalty: increased network load**

Conclusions: Energy Efficient cloud computing

- Balance between process, storage and transport
- Processing and Storage
  - Workload placement
  - Thermal-aware
  - Cooling-aware
  - Thermal-and-cooling-aware highest thermostat setting
  - Thermal activity monitoring of data centers by WSNs
- Transport
  - Energy-efficient anycasting/manycasting of cloud service demands

Concluding Remarks

Future cloud infrastructure needs

- high-performance, high-capacity and resilient optical network infrastructure for inter-DC connectivity,
- a tighter integration between the optical network and services/resources provided by DCs,
- coordinated virtualization of the optical network and DCs to create application-specific and independent virtual infrastructures (VIs) with guaranteed QoS coexisting over the shared underlying physical substrate
- energy efficient design and management tools

Some Additional Concerns

- Elastic cloud computing
- Resource sharing and scheduling
- Security