UNICAMP

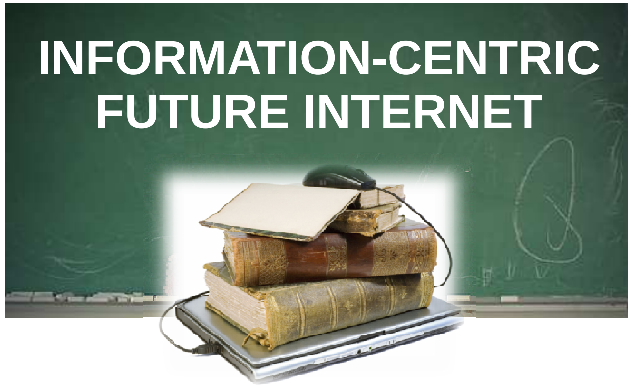
Information-oriented Internetworking

Towards a data-centric forwarding plane

Christian Esteve Rothenberg, 02/07/2009

Department of Computer Engineering and Industrial Automation
School of Electrical and Computer Engineering
State University of Campinas

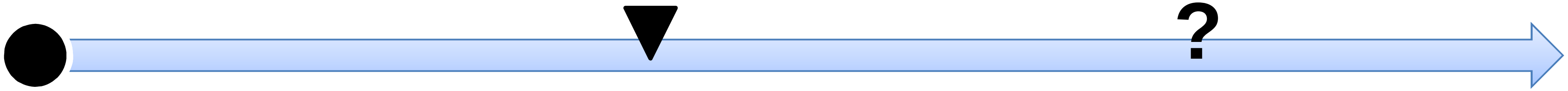
Agenda



INFORMATION-CENTRIC FUTURE INTERNET

Today

Future



information-centrism

PSIRP
PUBLISH-SUBSCRIBE INTERNET ROUTING PARADIGM

Haggle

TRIAD

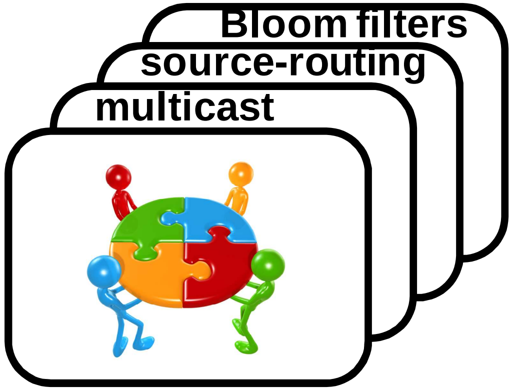
content-centric networking

4Ward

DONA

Thinking “out-of-the-TCP/IP-box”

Exploration



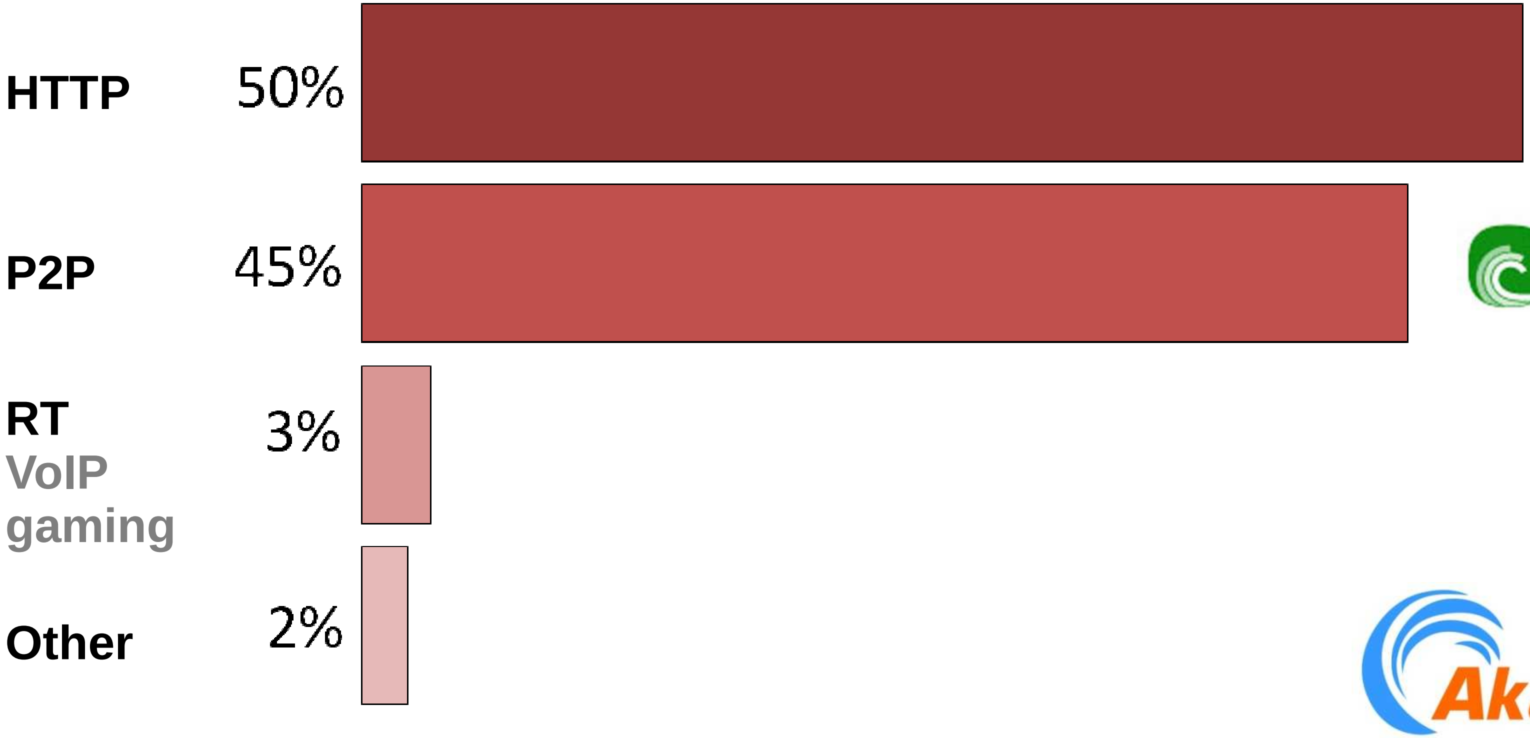
Bloom filters

source-routing

multicast

Components

Internet traffic



YouTube 25%

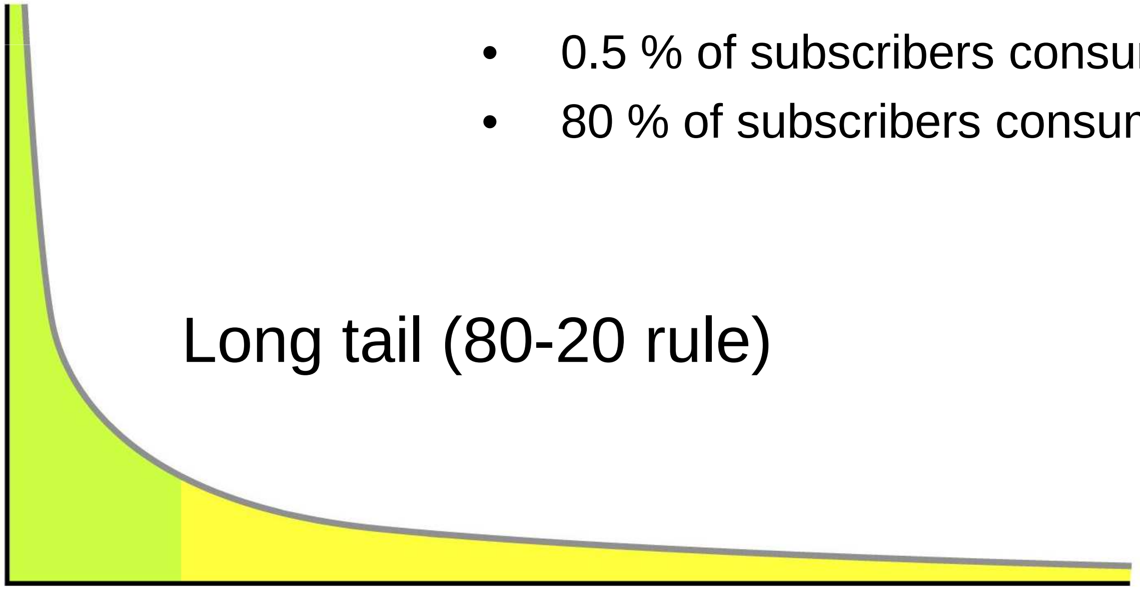
BitTorrent 20%

Akamai 20%

Sources: http://gigaom.com/2008/04/22/shocking-new-facts-about-p2p-and-broadband-usage/
http://www.ipoque.com/resources/internet-studies/internet-study-2007

Internet traffic

- 10 % of subscribers consume 80 % of BW
- 0.5 % of subscribers consume 40 % of BW
- 80 % of subscribers consume < 10 % of BW



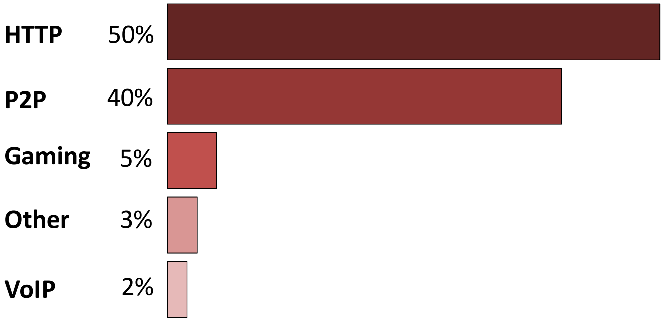
Long tail (80-20 rule)

Source:

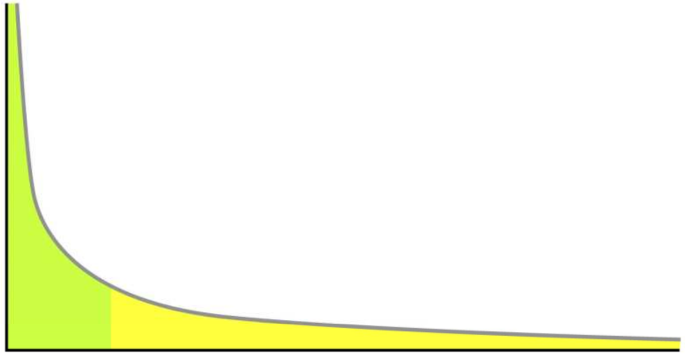
Future
?

Internet traffic

+ High Definition Content
+ Sensor Networks
+ Web 2.0
+ Mobile devices
+ Broadband



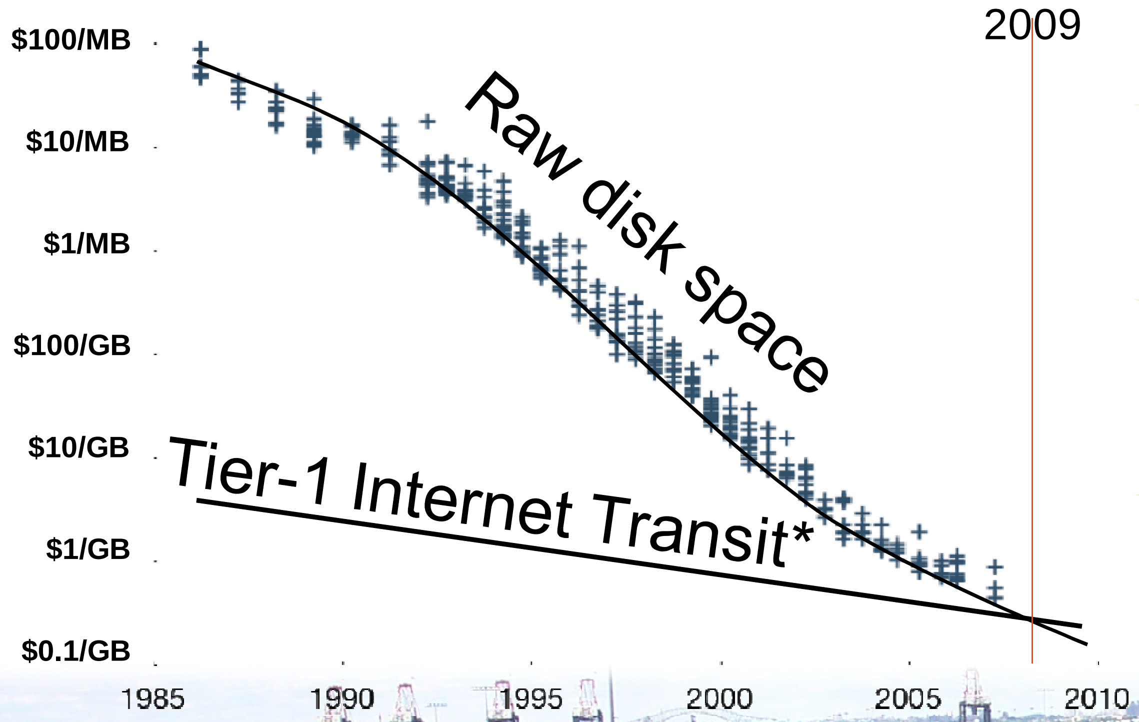
?

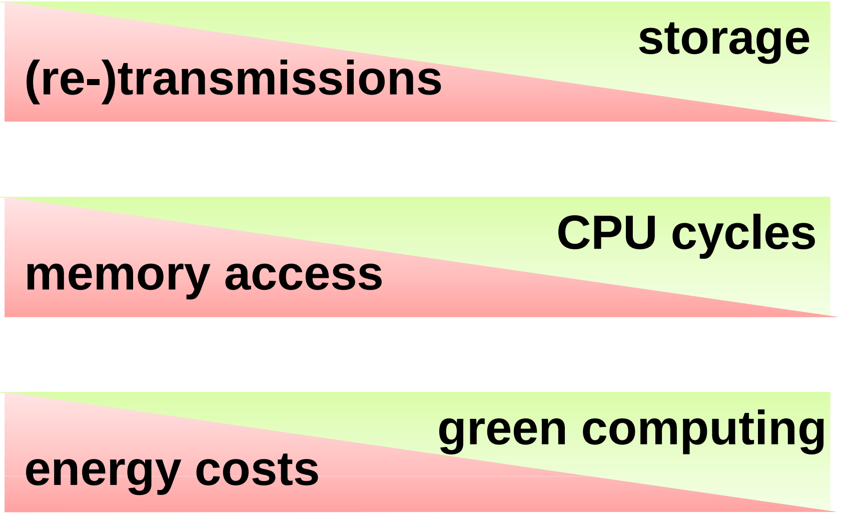


Source:

Network Economics

- Many of the technology assumptions behind the original end-to-end principle may no longer be applicable!





* Preliminary data [Nikander'09]

Network economics arguments for a back-to-basics reconsideration of the end-to-end networking model

TCP/IP

Origin

Solved the problem of resource sharing (FTP, Mail, Telnet, HTTP*) ☺

Today

☹

TCP train wreck applications:

- Massive P2P traffic [Accountability/ re-ECN]
- Multimedia home networking [Wireless losses]
- Cellular networks [E2E control loop]
- High-delay & High-bandwidth links [BW x Delay]
- Data-centers & Cloud computing [Slow start]

- **TCP ignores higher layer needs & lower layer characteristics!**
- **TCP notion of fairness under debate**

DNS

- Identify IP endpoints (computers, routers) ☺
- Handled at human rate

- Identify information objects (URI)! ☹

  - Semantic overload: both info name & location
- Under machine-machine applications

- How to move from server locations to **naming of information** really?
- How robust, scalable, sensitive to attacks and mis-configurations?
- How to HANDLE IP resolution and UPDATE bigger & bigger databases?

Today ▼

Content Delivery Networks

Future ?

Akamai **20%** of total Internet traffic

- Increased Quality of Experience 🙂
  - – Masks current Internet bottlenecks with an *overlay solution*

- **CDN lock-in**
- **Closed innovation**
- **Complex**
  monitoring, DNS tricks

Origin

Observation

Today

Fundamentals of the Internet

VS.

- Collaboration
  - Reflected in forwarding & routing
- Cooperation
  - Reflected in trust among participants
- Endpoint-centric services
  - Mail, FTP, even Web
  - Reflected in E2E principle

Reality in the Internet

- Current economics favor senders
  - Receivers are forced to carry the cost of unwanted traffic
- Phishing, spam, viruses
  - There is no trust any more
- Information-centric services
  - Do endpoints really matter?
  - Information retrieval through, e.g., CDNs, P2P

IP, full end-to-end reachability

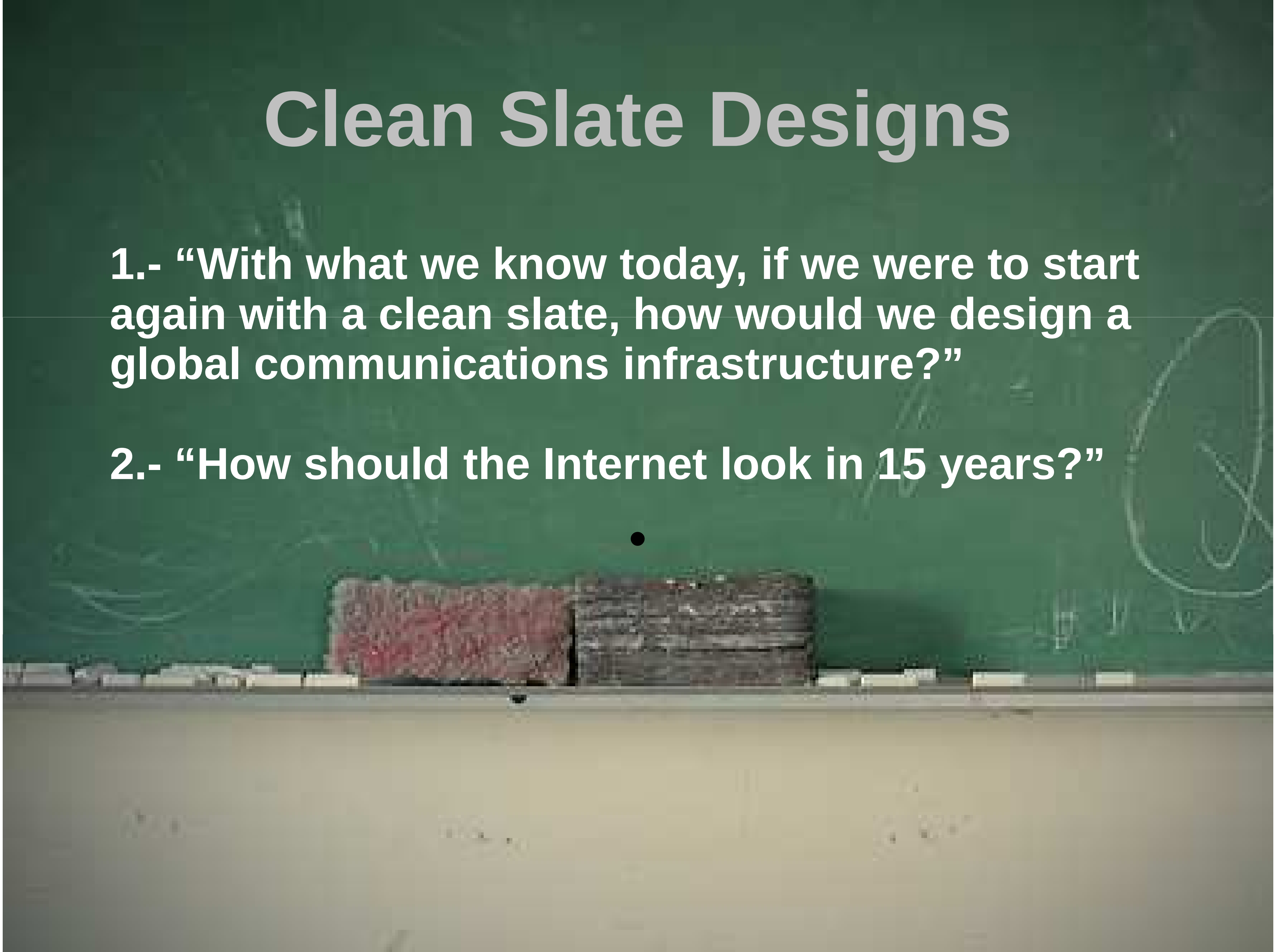
IP with middleboxes & significant decline in trust

Source: EU FP7 PSIRP Project

the future of the Internet & the future Internet ?

Source: Jonathan Zittrain, The Future of the Internet — And How to Stop It, , http://www.jz.org.

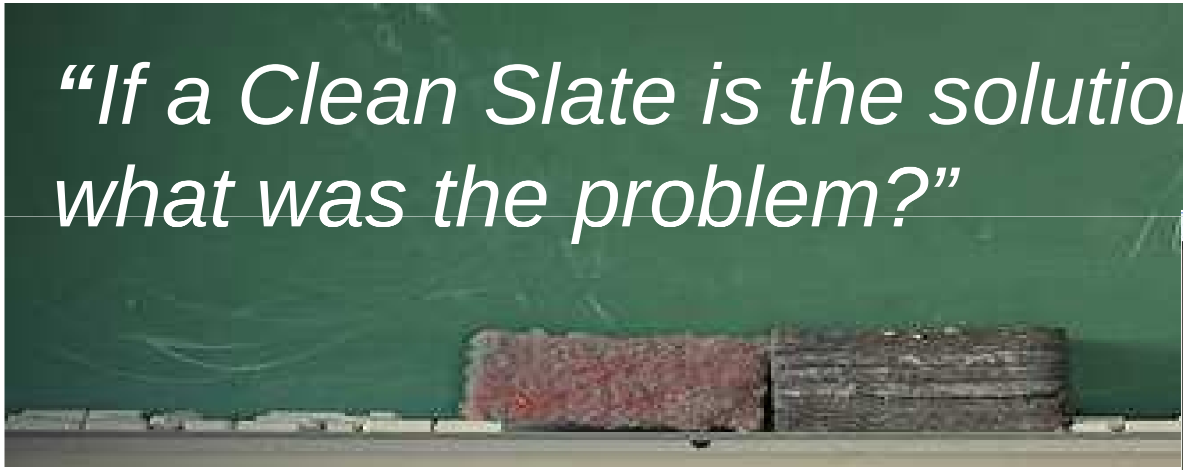
Clean Slate Designs



1.- “With what we know today, if we were to start again with a clean slate, how would we design a global communications infrastructure?”

2.- “How should the Internet look in 15 years?”

Van Jacobson’s waves of networking



“If a Clean Slate is the solution, what was the problem?”

99% Internet traffic:

Named chunks of data (Web, P2P, Video, etc.)

New problem: Dissemination of named pieces of data

Answer: Content-Centric Networking

http://video.google.com/videoplay?docid=-6972678839686672840

Re-Architecturing the Future Internet

information-centrism

PSIRP

DONA

content-centric networking

New ID spaces

TRIAD

ROFL

id-loc

CDN

clean-slate

3 *Interconnecting* ***information***

DPI

P2P

overlays

IPv6

NAT

middleboxes

networking (r)evolution

patching

TCP/IP

2 *Interconnecting* ***hosts***

Telephony

1 *Interconnecting* ***wires***

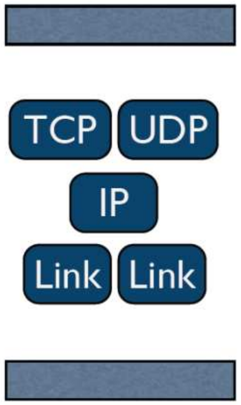
Information-oriented efforts

- Peer-to-Peer Networks (2000)
- The OceanStore Project (2002)
  - Global-Scale Persistent Data
- TRIAD: Content-Based Routing (2002)
  - Routing on FQDN for HTTP req. avoiding DNS resolution
- I3: Internet Indirection Infrastructure (2002)
  - DHT-based rendezvous points in the network
- LNA: Layered Naming Architecture (2004)
  - ID/Loc split at every layer
- DTN: Delay/Disruption Tolerant Networks (2003)
  - CNF: The cache-and-forward network architecture (2008)
  - Haggle: Pocket Switched Networks (2007)
  - IETF activities

Information-oriented efforts

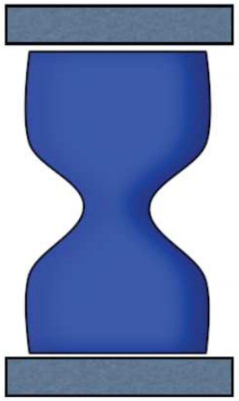
- CCN: Content Centric Networking (2006)
  - Aggregation through structural naming of data pieces
- DONA: Data Oriented Network Architecture (2007)
  - Register / Find P:L
- 4Ward NetInf (2008)
  - Networking of information objects
- Wireless Sensor Networks
  - Data-centric routing approaches
- PSIRP: Publish Subscribe Internet Routing Paradigm (2008)
  - Replace IP with a pure pub/sub based inter-networking stack

PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM



Information-oriented Networking

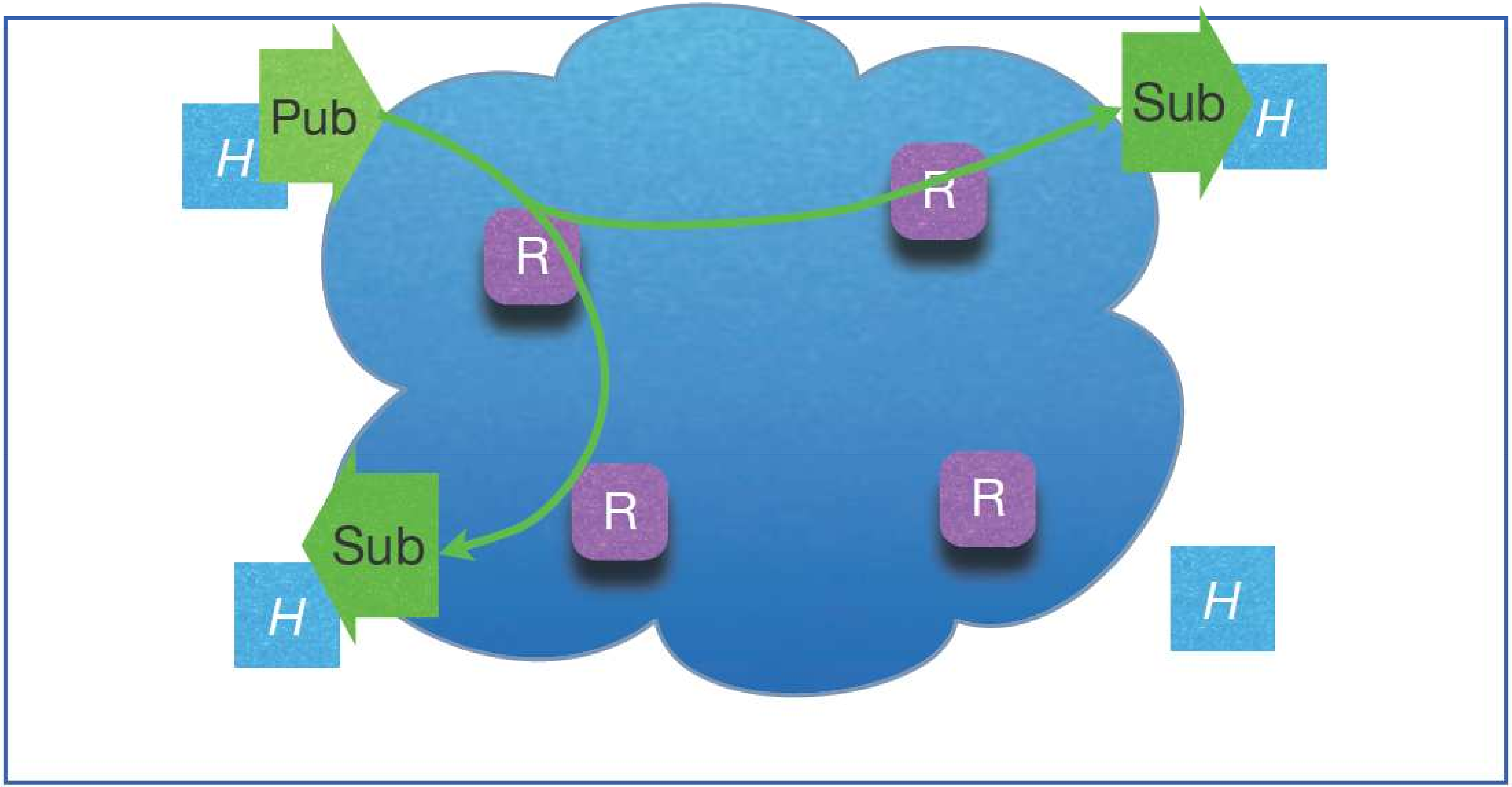
- Rethinking fundamentals -



- Send / Receive → Publish / Subscribe
- Sender-driven → Receiver-driven
- Host names → Data names
- Host reachability → Information scoping
- Channel security → Self-certified metadata
- Unicast → Multicast

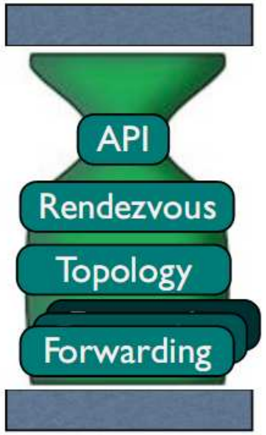
Basic pub/sub networking

PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM

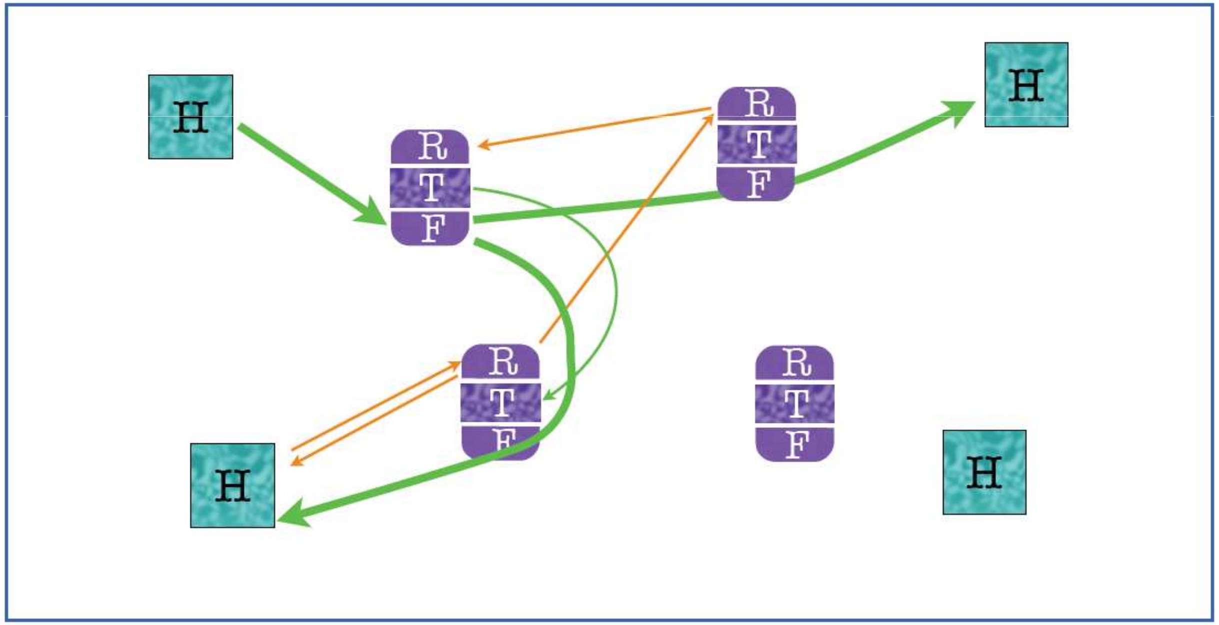


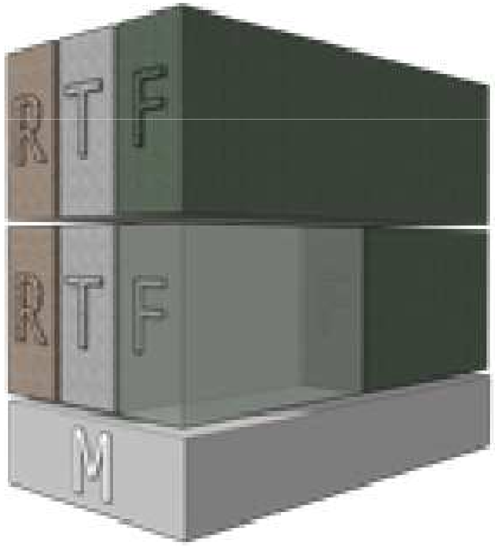
Source: EU FP7 PSIRP Project, http://psirp.org

RTFM Architecture



- **Rendezvous**
  - Matching subscriptions to publications
- **Topology**
  - Creating and maintaining delivery trees used for forwarding publications
- **Forwarding**
  - Data delivery operations. e.g., label switching, fast forwarding
- **and More**
  - Node-to-node link data transfer + e.g., opportunistic caching, collaborative and network coding, lateral error correction etc.

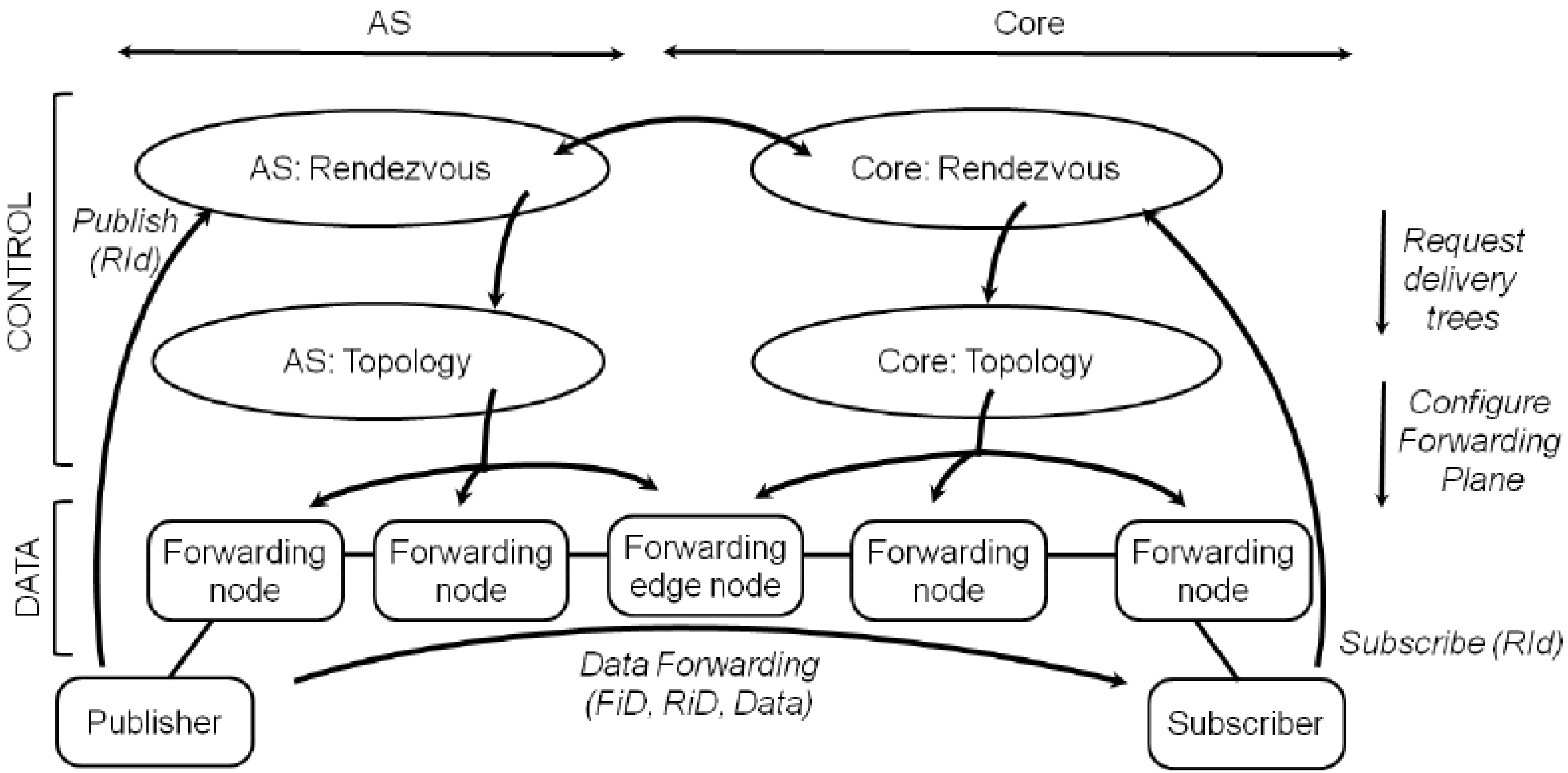




Source: EU FP7 PSIRP Project, http://psirp.org

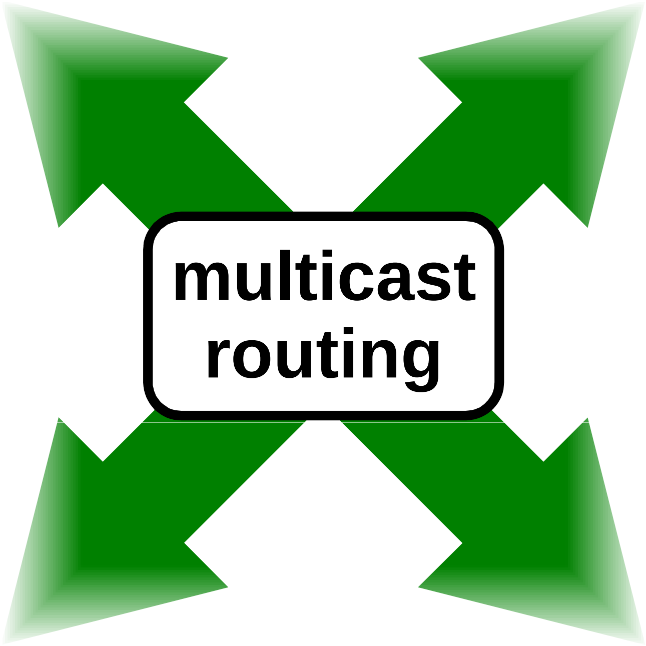
High level architectural overview

- Mapping information to delivery trees -



- **Rendezvous identifier (RiD)**:
  - Self-certifying identifier of data
- **Forwarding identifier (FiD)**:
  - Used for fast forwarding

4-dimensional solution space



multicast routing

Transport efficiency
(Stretch)

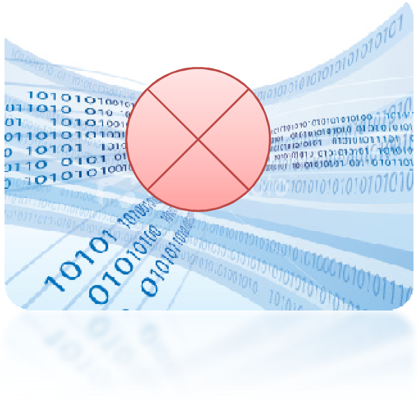
Routing / forwarding
information in packets

Signaling
overhead

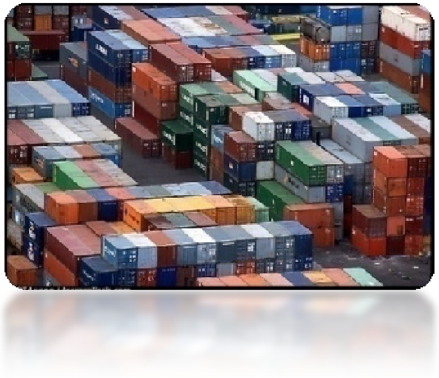
Routing/forwarding
state in network elements

Challenges & Approach

- **Challenges** of an information-centric forwarding plane
  - Take **switching** decisions
    - at *wire speed* (Gbps)
    - on a *large* universe of *flat* identifiers
  - **Scalable native multicast** support
    - no host-based addressing
    - delivery trees of information flows



- **Approach:**
  - Trade **state** for **over-deliveries**
  - Take advantage of a **data-oriented** paradigm
  - Divide & Conquer



Divide and Conquer

Source routing

Hierarchical aggregation

Install network state only when necessary

Stepwise approach for delivery tree management

Transport efficiency ⟷ **Trade-off** ⟷ **Scalability**

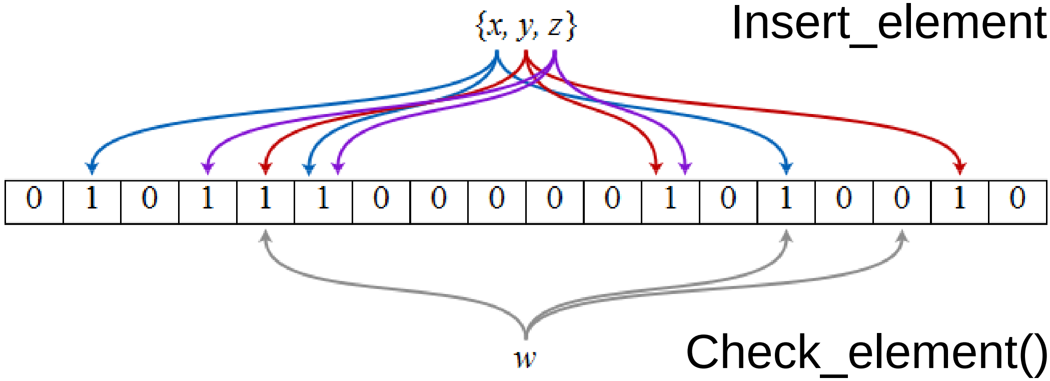
(non-ideal trees, over-deliveries, min. signalling & forwarding tables)

(Image Credit: Scott Maxwell / http://www.lumaxart.com / CC By-SA 2.0)

The role of Bloom and family

- Well-known Bloom filter
  - Efficient *data aggregator*
  - False positive:

    f (memory / # elements)



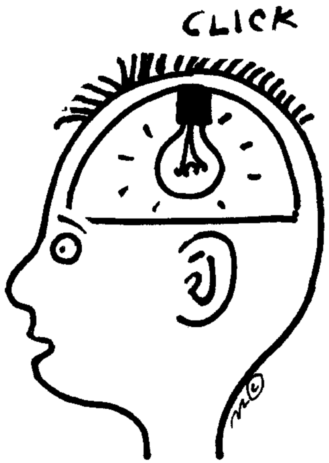
Insert_element()

Check_element()

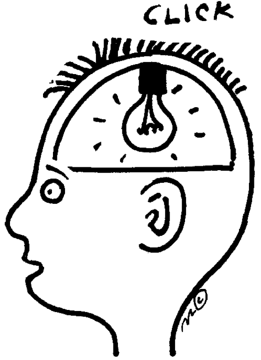
“yes” / no ?

- Wire-speed forwarding requirements
  - Low (bounded) packet processing *time* (time to hash)
  - Limitations in high-speed *memory*

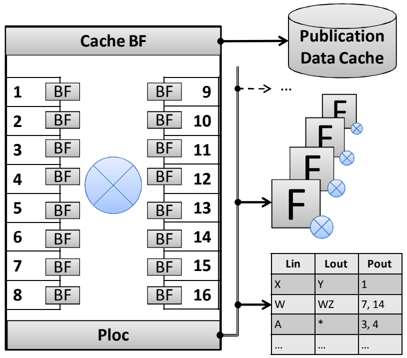
- **A scalable, data-centric forwarding approach:**
  - Bloom-filter-based forwarding
    as *set membership-problems*



Bloom-filter-based forwarding



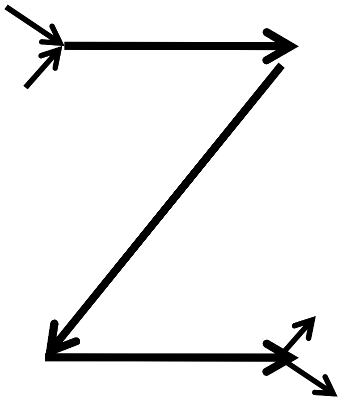
2 extreme & complement *set membership-problems:*



Lin	Lout	Pout
X	Y	1
W	WZ	7, 14
A	*	3, 4
...	...	...

SPSwitch [3]: *Is packet label X in forwarding port P?*

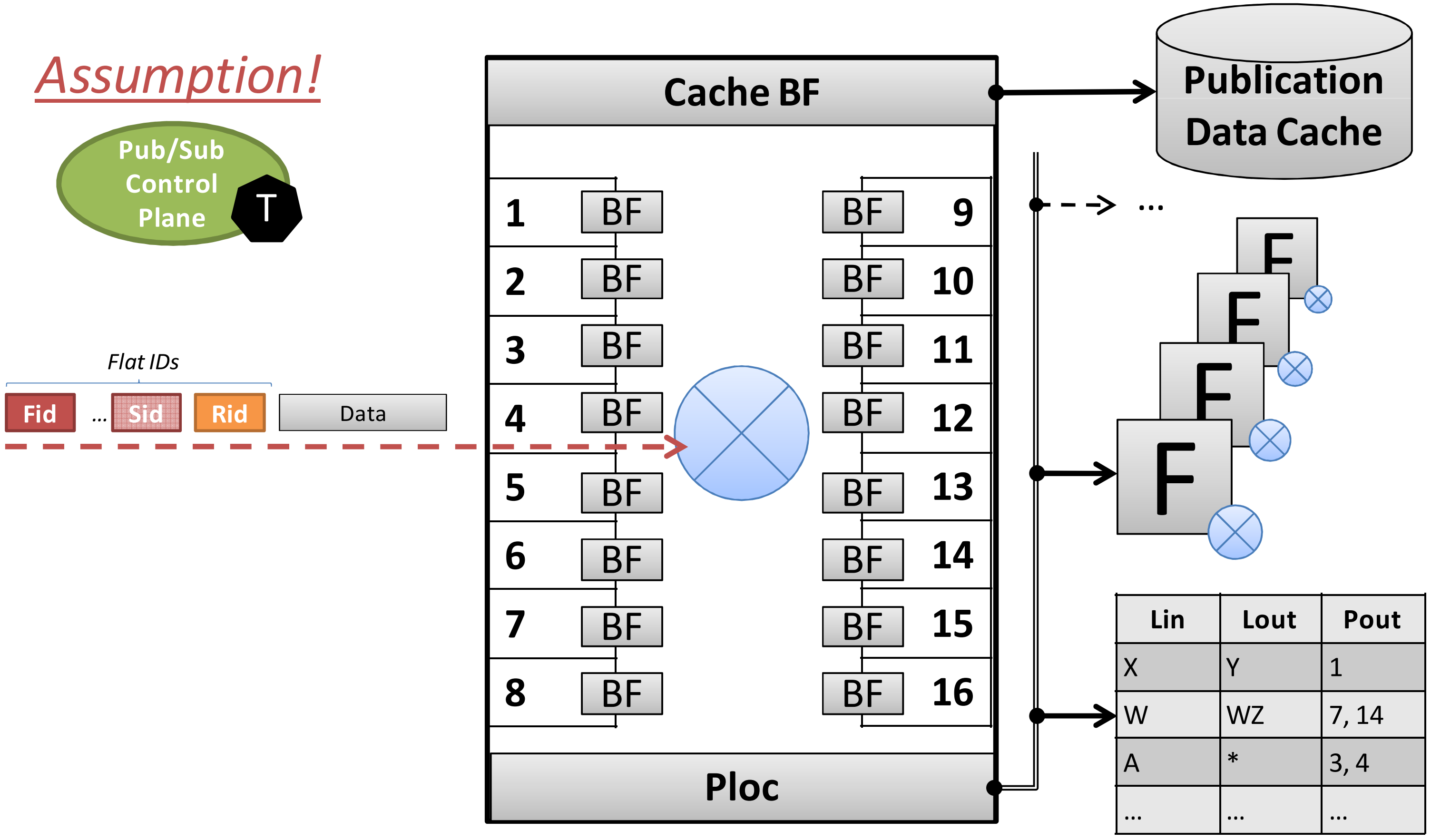
- State in the **network**
- Large Bloom filters maintained in **forwarding tables**



zFilters [5]: *Is outbound link A in packet header Z?*

- State in the **packet header**
- Small in-packet Bloom filter representing a **source route**

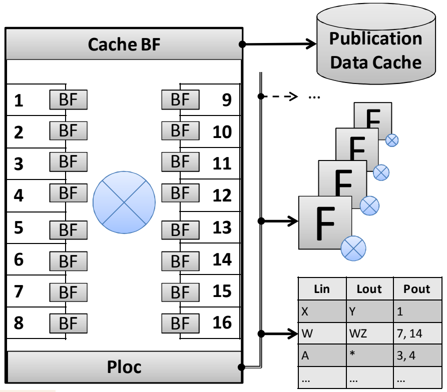
SPSwitch



Lin	Lout	Pout
X	Y	1
W	WZ	7, 14
A	*	3, 4
…	…	…

Is packet label X in forwarding port P?

Experimental results



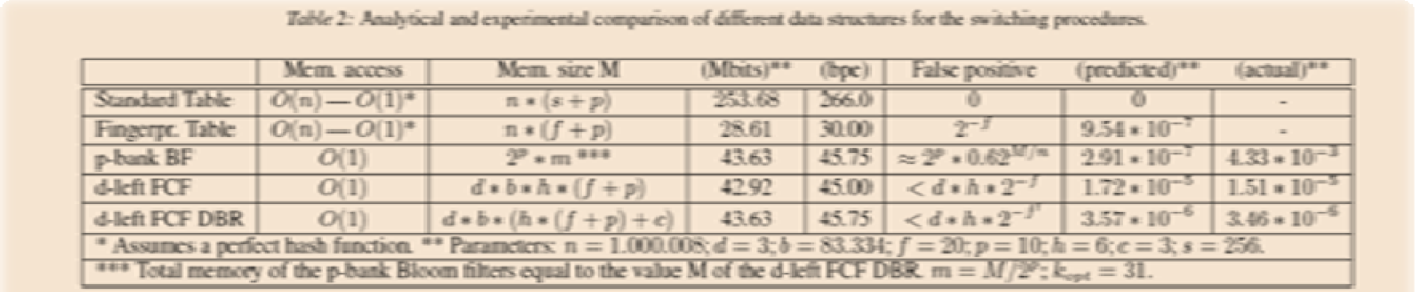


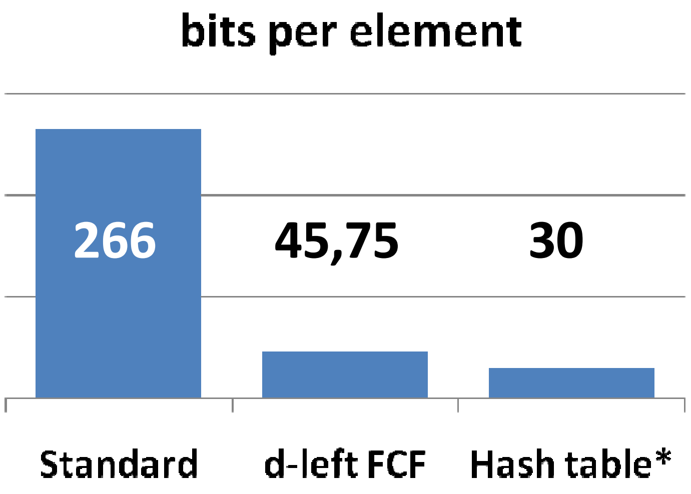
Table 2: Analytical and experimental comparison of different data structures for the switching procedures.

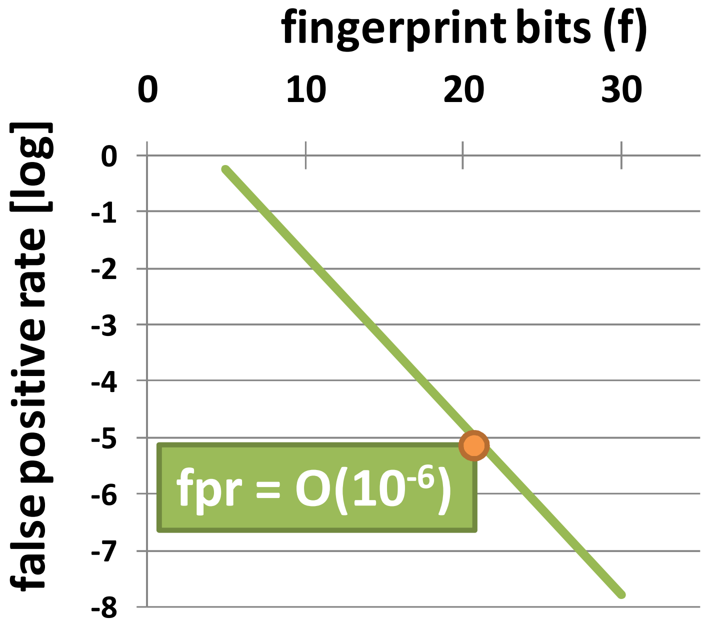
	Mem. access	Mem. size M	(Mbits)**	(bpe)	False positive	(predicted)**	(actual)**
Standard Table	$O(n) - O(1)^*$	$n * (s+p)$	253.68	266.0	0	0	-
Fingerpr. Table	$O(n) - O(1)^*$	$n * (f+p)$	28.61	30.00	2^{-f}	$9.54 * 10^{-7}$	-
p-bank BF	$O(1)$	$2^p * m$ ***	43.63	45.75	$\approx 2^p * 0.62^{M/m}$	$2.91 * 10^{-7}$	$4.33 * 10^{-3}$
d-left FCF	$O(1)$	$d * b * h * (f+p)$	42.92	45.00	$< d * h * 2^{-f}$	$1.72 * 10^{-5}$	$1.51 * 10^{-5}$
d-left FCF DBR	$O(1)$	$d * b * (h * (f+p) + c)$	43.63	45.75	$< d * h * 2^{-f}$	$3.57 * 10^{-6}$	$3.46 * 10^{-6}$

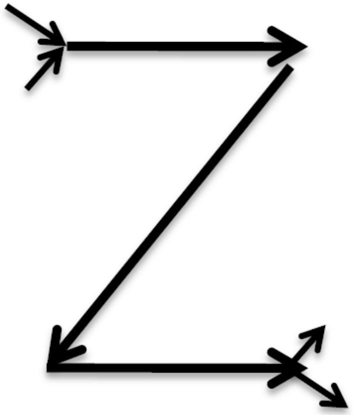
* Assumes a perfect hash function. ** Parameters: $n = 1.000.008$; $d = 3$; $b = 83.334$; $f = 20$; $p = 10$; $h = 6$; $c = 3$; $s = 256$.

*** Total memory of the p-bank Bloom filters equal to the value M of the d-left FCF DBR. $m = M/2^p$; $k_{opt} = 31$.

20-bit fingerprint + 10-bit port







zFilters: in-packet Bloom filter encoding of delivery trees

State in the *packet headers*

- Each network link has an identity and (a series of) *Link IDs:* *LIT: 256 bit vector with just k=5 bit positions set to one*
- Delivery tree by ORing the Link IDs into a fixed-size in-packet Bloom filter (zFilter) representing a *source route*

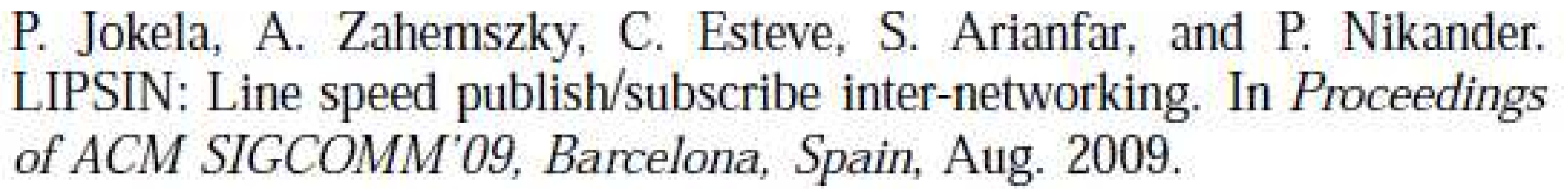
Basic forwarding operation

“Is outbound link A in packet header Z?”

- *Small* forwarding tables (Link ID to neighbors + Virtual Link IDs)
- *Fast* packet forwarding (bitwise AND operations)

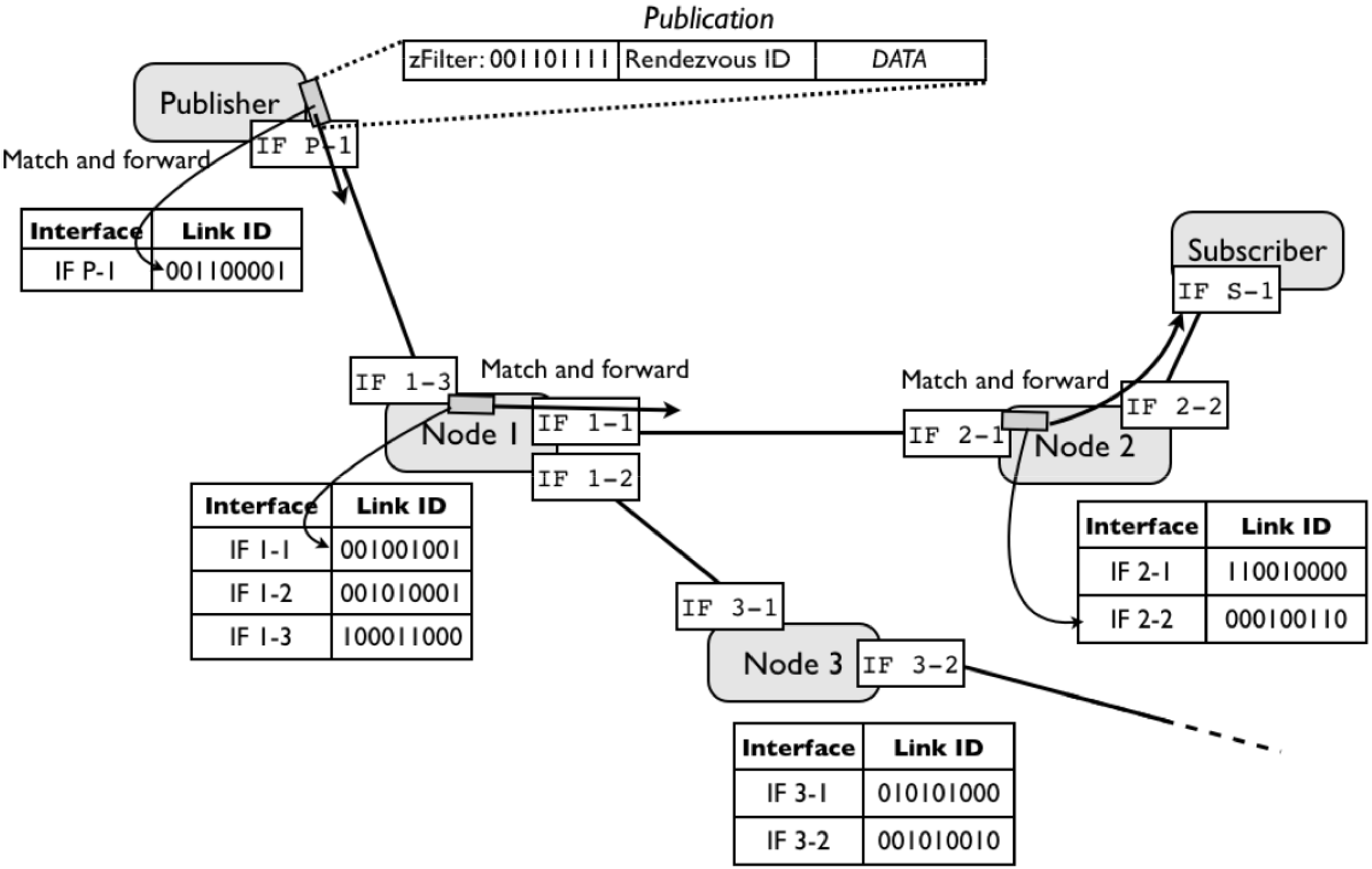
Extensions and details:

[10]



P. Jokela, A. Zahemszky, C. Esteve, S. Arianfar, and P. Nikander. LIPSIN: Line speed publish/subscribe inter-networking. In *Proceedings of ACM SIGCOMM'09, Barcelona, Spain*, Aug. 2009.

Forwarding on Bloomed link identifiers



Publication

zFilter: 001101111	Rendezvous ID	*DATA*

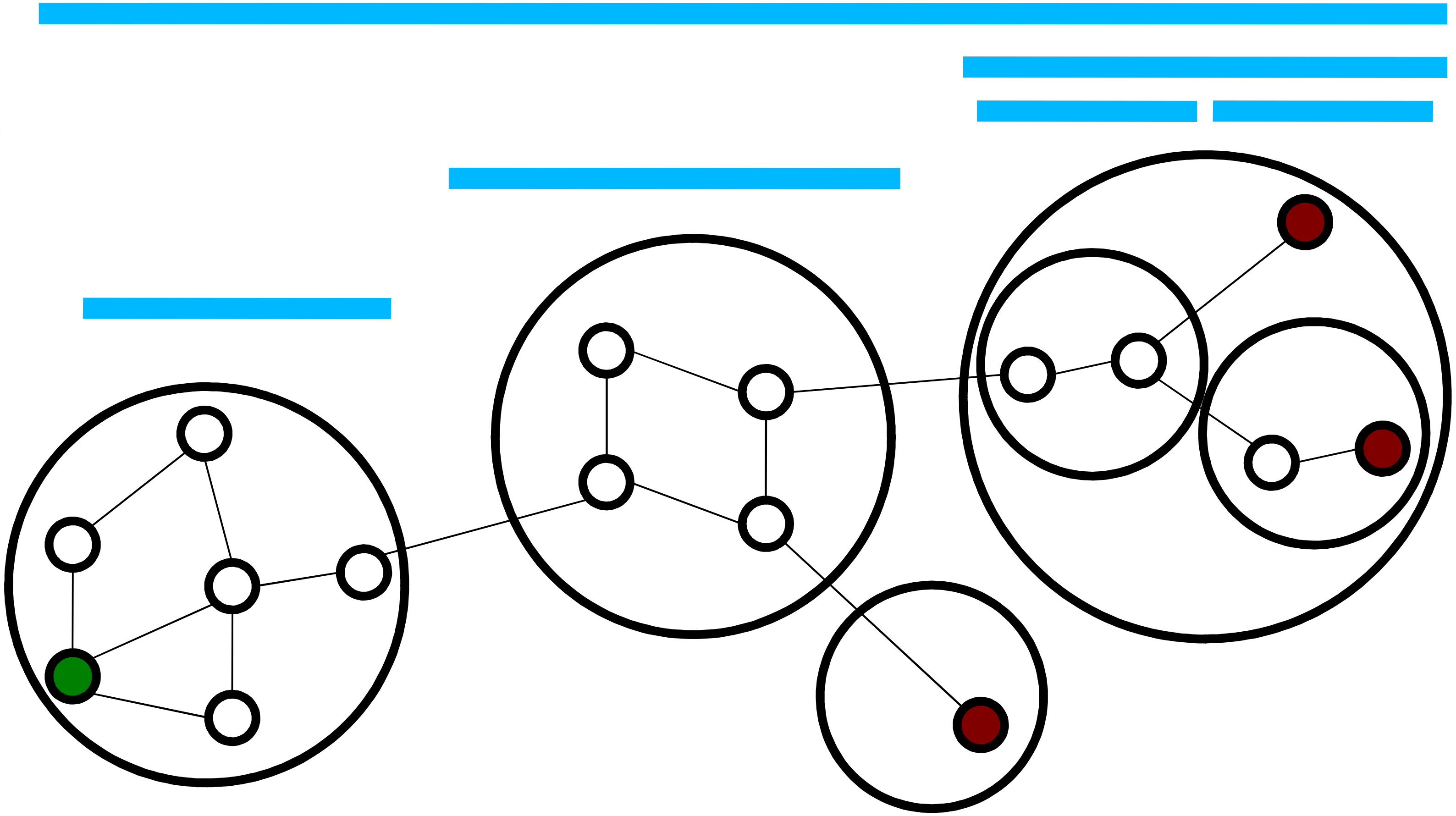
Interface	Link ID
IF P-1	001100001

Interface	Link ID
IF 1-1	001001001
IF 1-2	001010001
IF 1-3	100011000

Interface	Link ID
IF 2-1	110010000
IF 2-2	000100110

Interface	Link ID
IF 3-1	010101000
IF 3-2	001010010

Virtual links

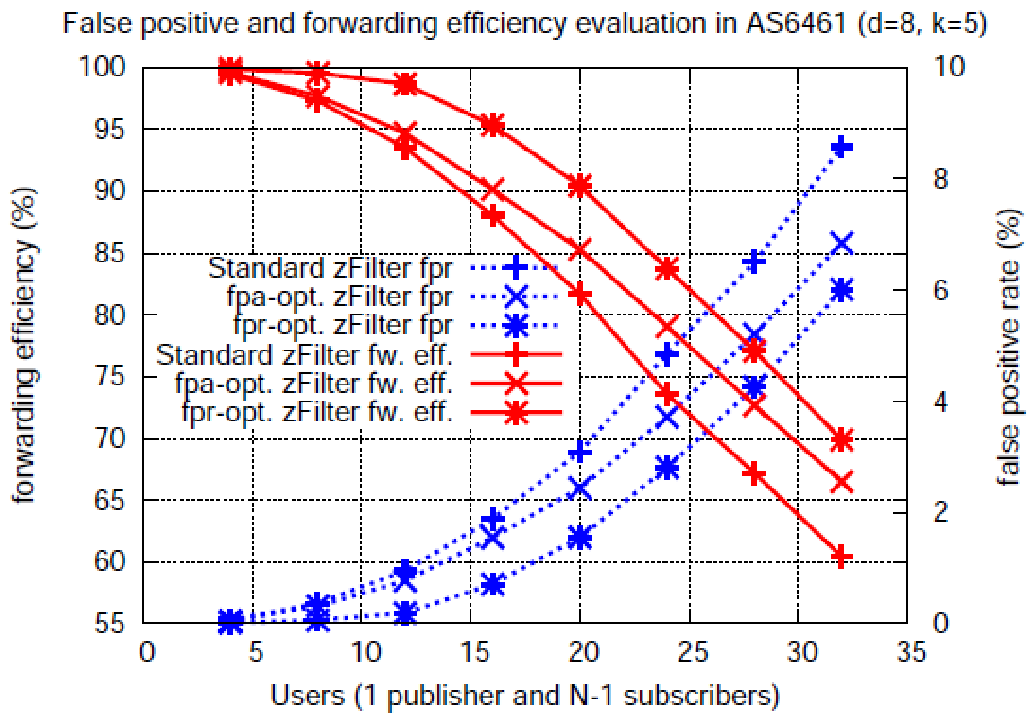


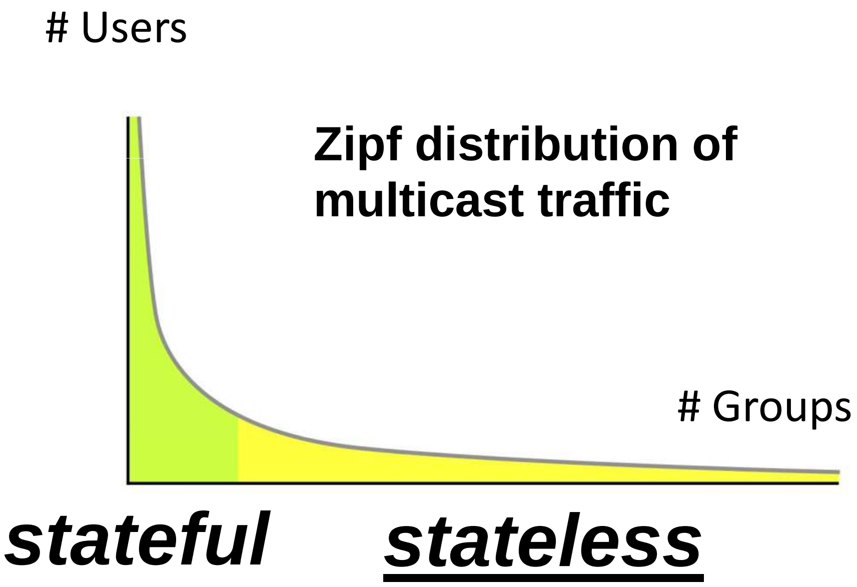
State in network nodes

- One-to-one, one-to-many, many-to-many, many-to-one forw. structures
- Supporting horizontal and/or hierarchical aggregation
- Less overdeliveries

Practical results

- Stateless multicast with 256-bit zFilters (35 links -> 20 subscribers)
- Enough for sparse multicast in typical WAN





Delivery trees in 5 steps

1) Compute an *ideal tree.*

2) Determine the *gaps* between the ideal tree and any existing trees.

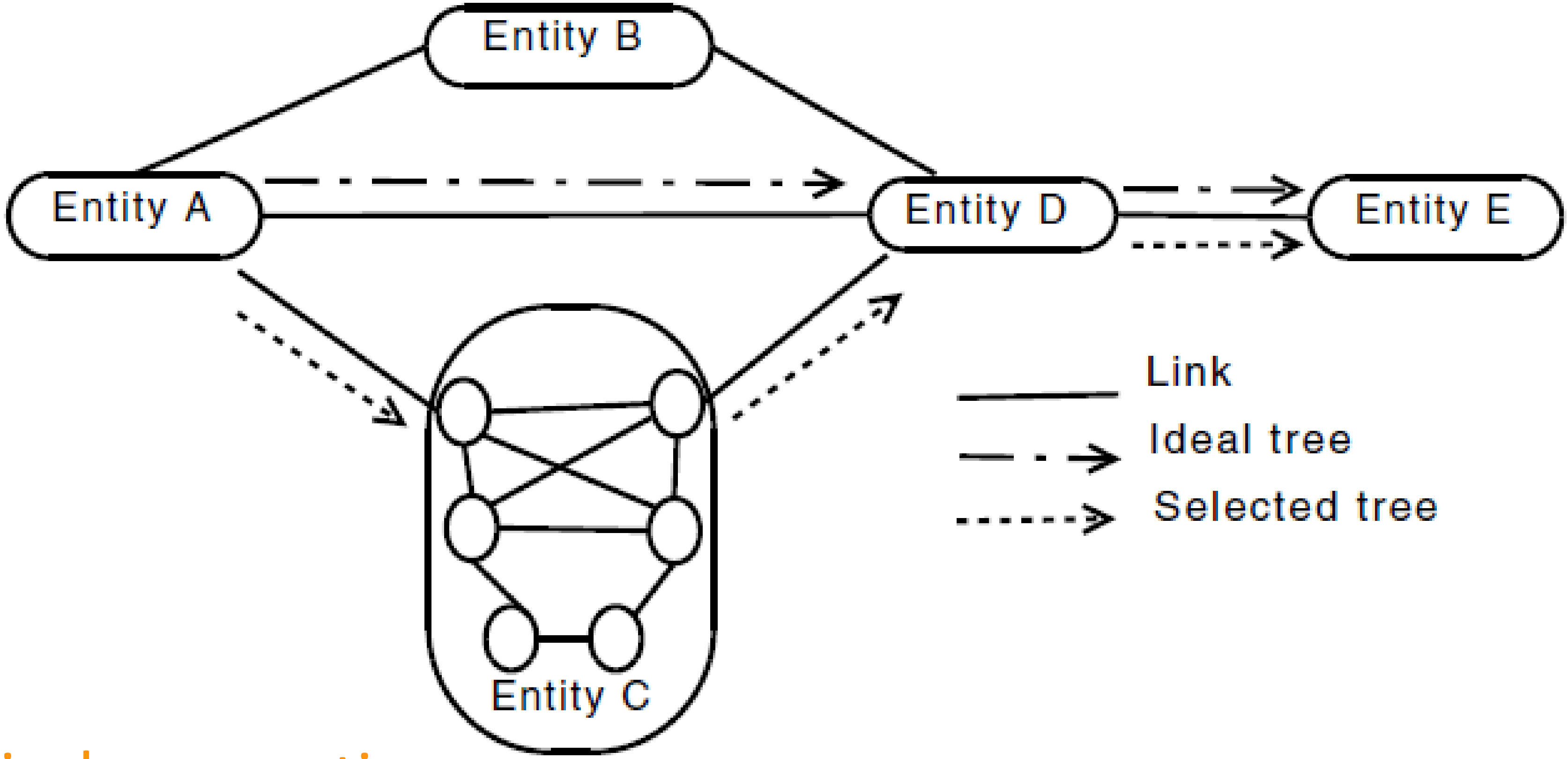
3) Select *tree-creation* strategies or *gap-filling* strategy for each gap.

4) *Compute* the needed *changes* according to the strategies.

5) *Apply* the changes to the network.

(Image Credit: Scott Maxwell / http://www.lumaxart.com / CC By-SA 2.0)

Example



Hierarchical aggregation

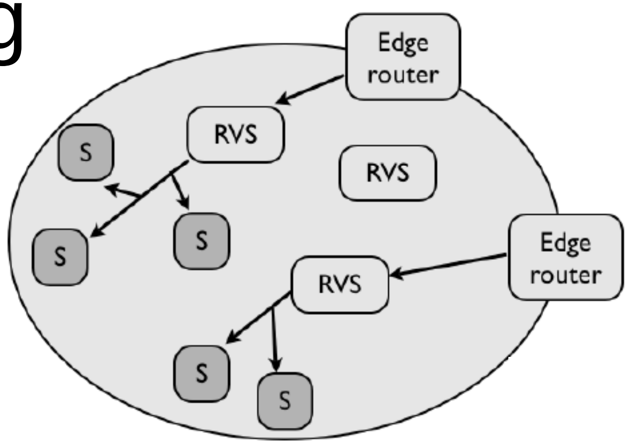
- AS confederations, ASes, intra-domain areas, routers

Selecting a *good enough* tree

- Strict requirement: containing all the subscribers

Challenges and future work

Inter-domain routing and forwarding



Avoid the mapping problem:

- Between intra-AS trees and inter-AS trees no one-to-one mapping exist
- *Do we really need rendezvous identifier-based matching for label swapping?*
- Hints for future directions:
  - Information scopes
  - Non-routable link identifiers for mapping

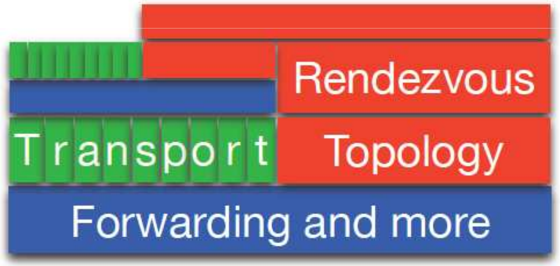
Topology functions:

- performance implications
- delay
- inter-operation between Topology Managers

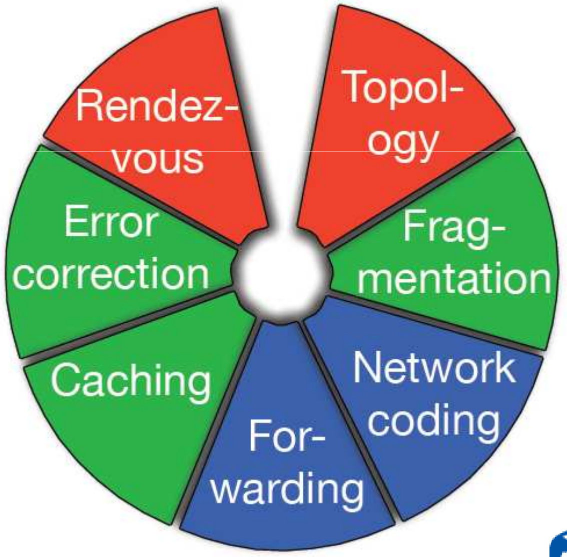
Prototype implementation

PSIRP
PUBLISH-SUBSCRIBE
INTERNET ROUTING
PARADIGM

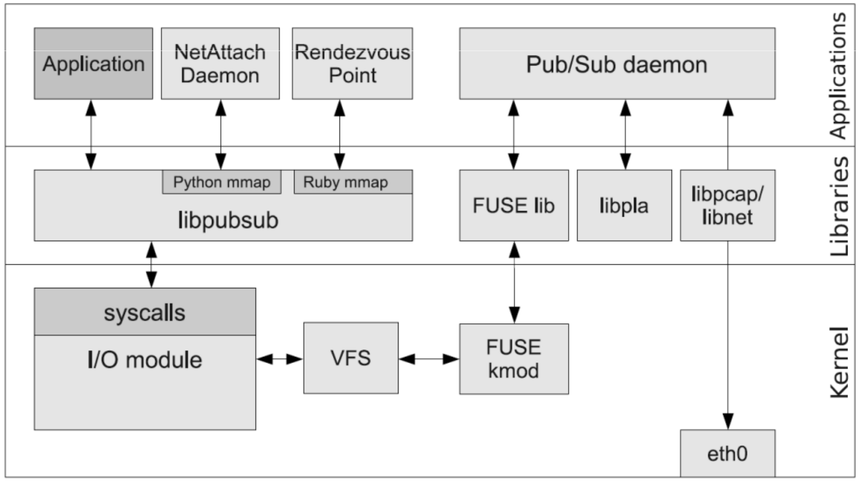
RTFM architecture



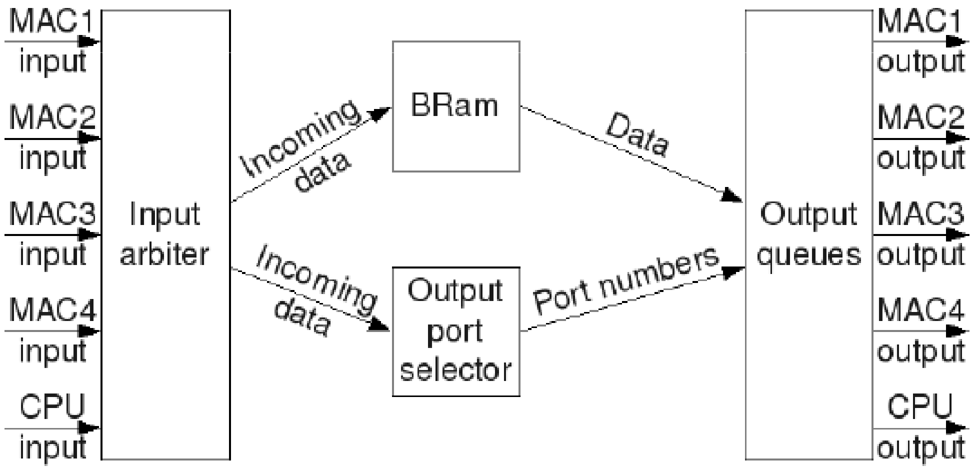
Component Wheel



FreeBSD®

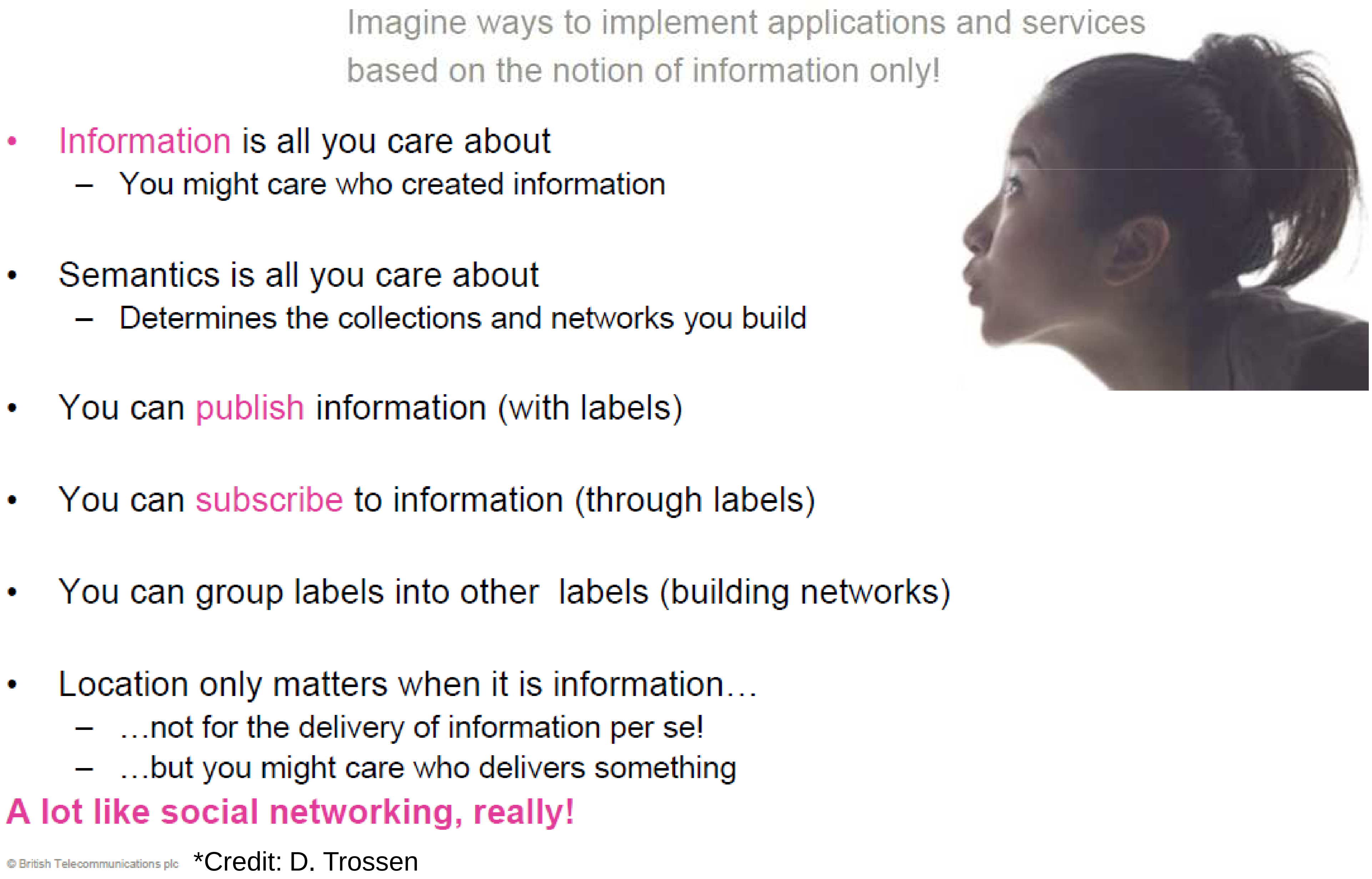


NetFPGA



Source: EU FP7 PSIRP Project, http://psirp.org

Pure pub/sub application development*



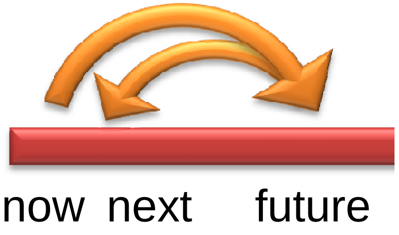
Imagine ways to implement applications and services based on the notion of information only!

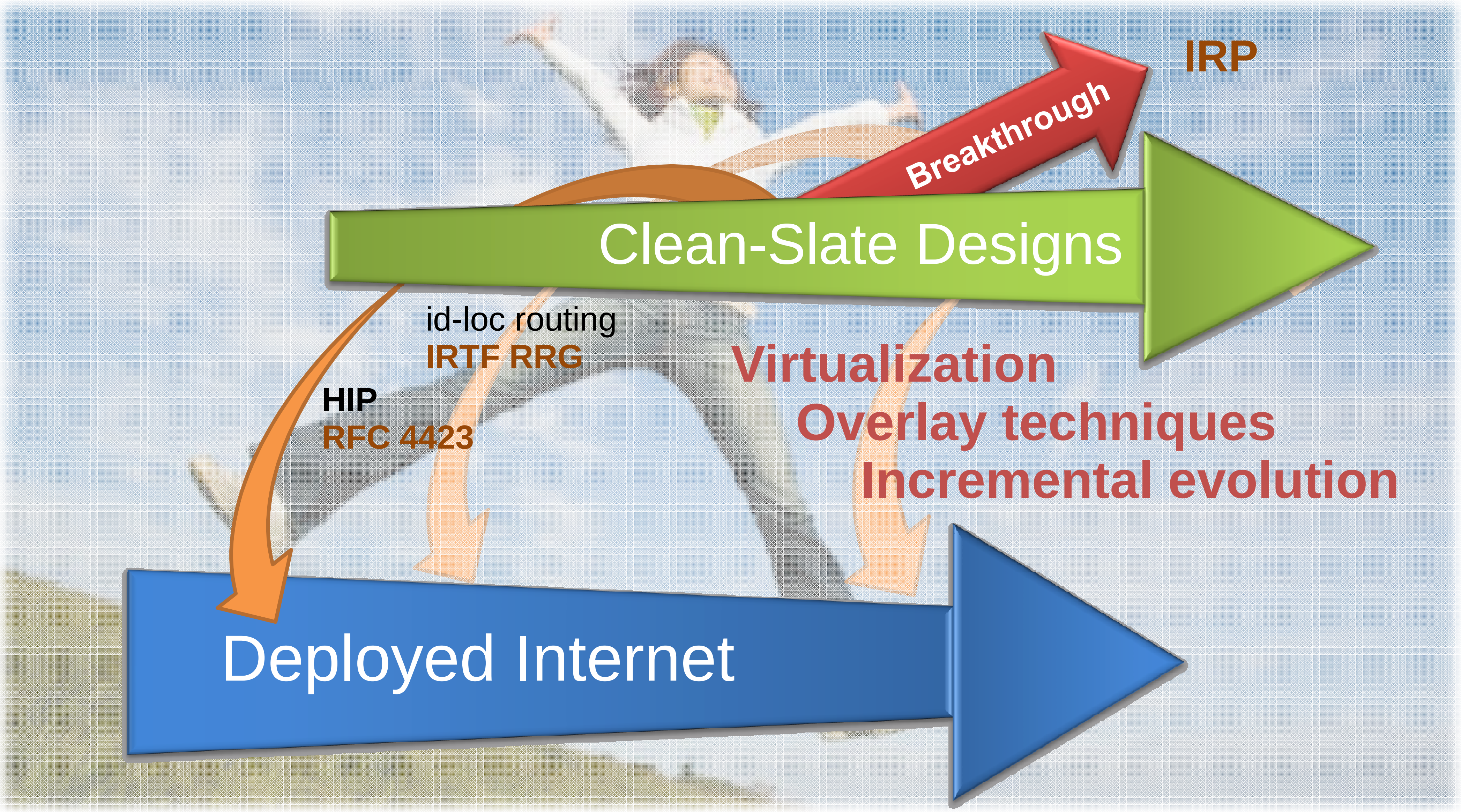
- Information is all you care about
  - You might care who created information
- Semantics is all you care about
  - Determines the collections and networks you build
- You can publish information (with labels)
- You can subscribe to information (through labels)
- You can group labels into other labels (building networks)
- Location only matters when it is information…
  - …not for the delivery of information per se!
  - …but you might care who delivers something

A lot like social networking, really!

*Credit: D. Trossen

Closing the research loop:
Late binding to reality





Clean-Slate Designs

Breakthrough

IRP

id-loc routing
IRTF RRG

HIP
RFC 4423

Virtualization
Overlay techniques
Incremental evolution

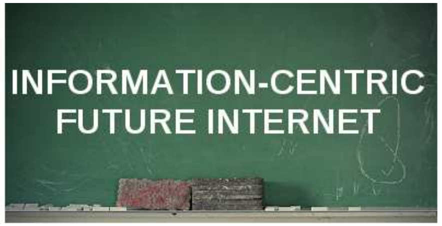
Deployed Internet

Closing the research loop: Data Center Networking

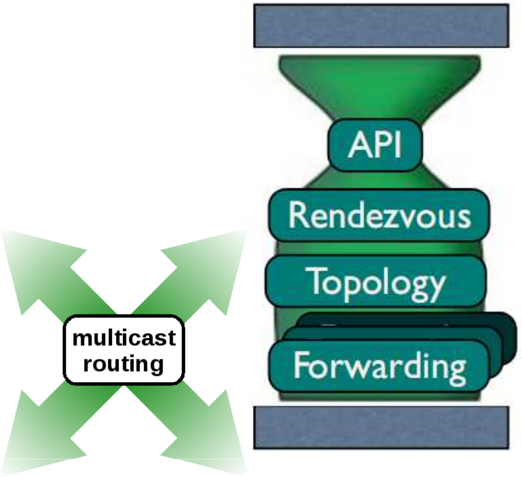
- Instantiate the forwarding mechanisms in a realistic data center environment
  - Scalable L2 flat architecture (cost-driven)
  - Source routing (e.g., middlebox concatenation)
  - Stateless multicast
  - Resource pooling:
    - Load-balanced oblivious routing exploiting multi-path & id/loc capabilities
  - DDoS-resistant architecture
- Control Plane based on OpenFlow

Take Aways

We are building an *information-centric* network based on the *publish / subscribe* paradigm

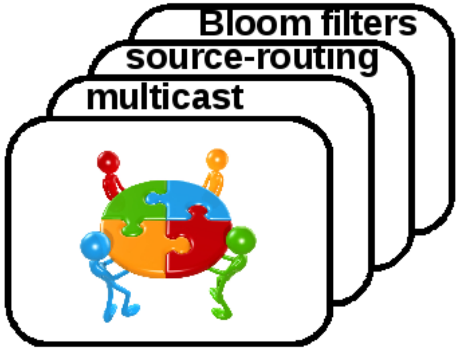


We are re-thinking the forwarding plane with *native multicast* departing from host-centric designs



To meet the *scalability* requirements, we explore the trade-off between *transport efficiency* and network state via

1) *Bloom-filter-based* forwarding decisions
2) approximate *delivery trees*
3) hierarchical/horizontal *division*

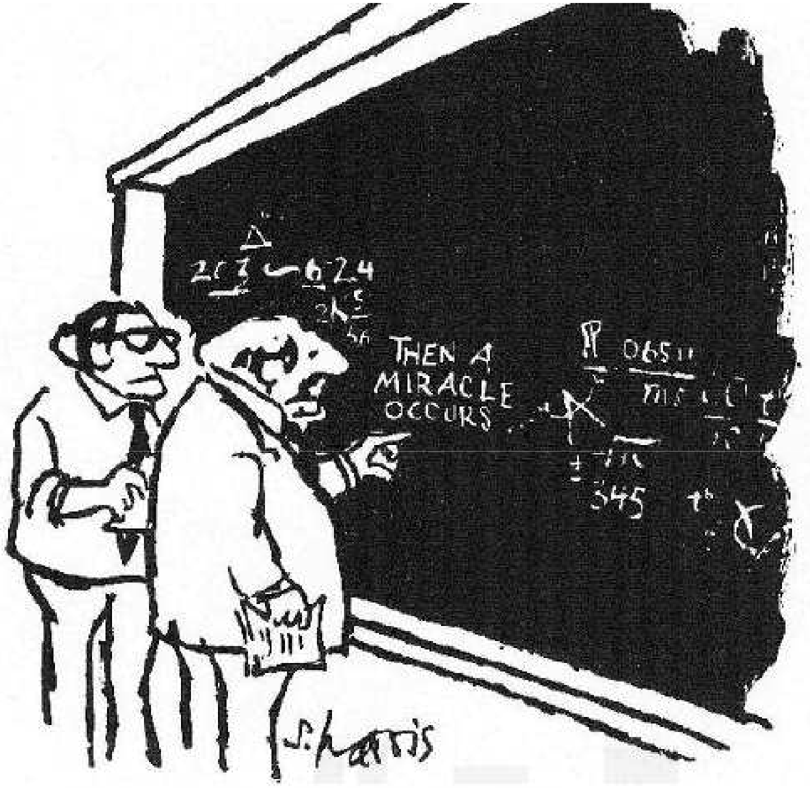


We have a flexible design for routing & forwarding, with component enablers allowing:

stateless and *stateful* operations

balance state : packet *headers* <·> netw. *nodes*

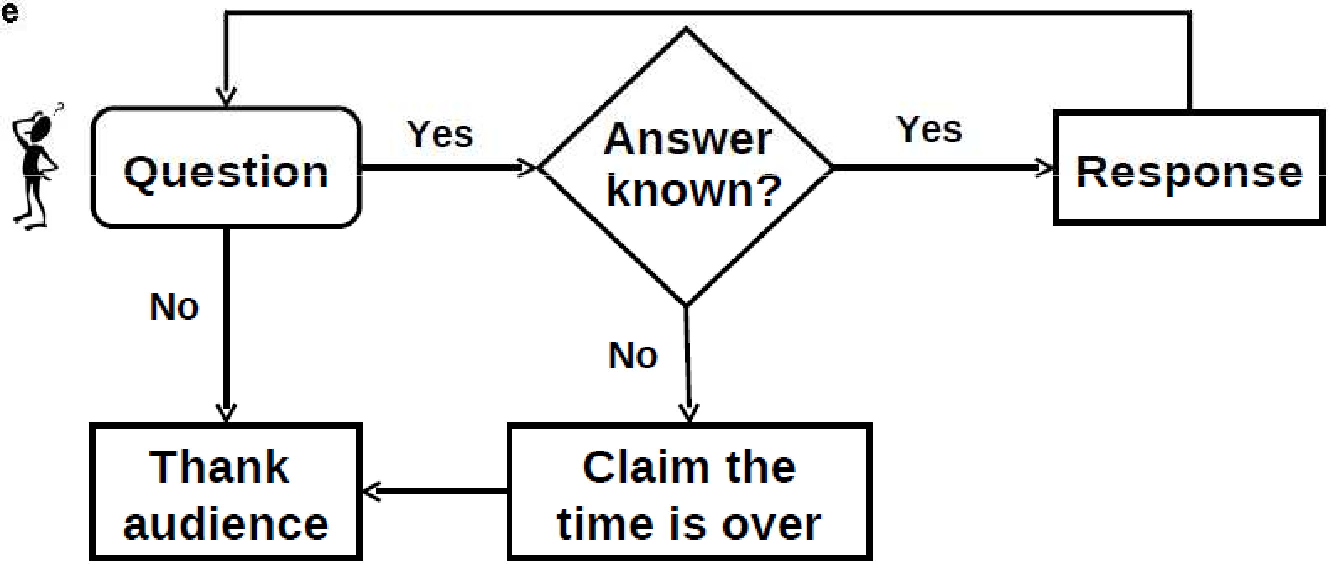
Obrigado!



“I think you should be more explicit here in step two”

questions?

Thank you!



Question

Yes

Answer known?

Yes

Response

No

No

Thank audience

Claim the time is over

References

- [1] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, “Network information flow,” IEEE Trans. on Information Theory, vol. 46, pp. 1204–1216, 2000.
- [2] A. Anand, A. Gupta, A. Akella, S. Seshan, and S. Shenker. Packet caches on routers: The implications of universal redundant traffic elimination. In ACM SIGCOMM, 2008.
- **[3] C. Esteve Rothenberg, Fábio Verdi and Maurício Magalhães. "Towards a new generation of information-oriented internetworking architectures" ACM CoNext, First Workshop on Re-Architecting the Internet (Re-Arch’08). Dec. 2008, Madrid, Spain.**
- [4] V. Jacobson. If a clean slate is the solution what was the problem? Stanford “Clean Slate” Sem., Feb 2006.
- **[5] P. Jokela, A. Zahemszky, C. Esteve, S. Arianfar, and P. Nikander. “LIPSIN: Line speed Publish/Subscribe Inter-Networking”. In ACM SIGCOMM 2009, Barcelona, Spain.**
- [6] Särelä M, Rinta-aho T, Tarkoma S. RTFM: Publish/Subscribe Internetworking Architecture. ICT-MobileSummit 2008.
- [7] D. Trossen (ed.), “Conceptual Architecture of PSIRP Including Subcomponent Descriptions (D2.2),” June 2008. [Online:] *http://psirp.org/publications*
- **[8] A. Zahemszky, C. Esteve, A. Csaszar and P.Nikander (LMF). “Exploring the Pub/Sub Routing & Forwarding Space”. In IEEE ICC, Workshop on the Network of The Future, Jun. 2009, Dresden, Germany.**

Credits

- D. Trossen and P. Nikander, EU FP7 PSIRP project, http://psirp.org
- Van Jacobson, http://video.google.com/videoplay?docid=-6972678839686672840
- Ericsson Research
- …

Images

- Jonathan Zittrain, The Future of the Internet — And How to Stop It, http://www.jz.org.
- Bert van Dijk at http://flickr.com/photos/75478114@N00/2964148062.
- Rae Brune at http://flickr.com/photo/75219074@N00/126116912
- Roy van Wijk at http://www.flickr.com/photos/royvanwijk/2974434570/
- *The Tango project at* http://commons.wikimedia.org/wiki/Smiley