May 18, 2006
Stockholm

Building scalable SIP networks

Adrian Georgescu
ag@ag-projects.com

My name is Adrian Georgescu

- Founder and CEO of AG Projects
- Co-chair of Dutch SIP SIG working group
- Member of OpenSER management board
- Member of ETSI ENUM special task force

IETF standardization for the SIP protocol is nearly finished, the SIP protocol and related components are already defined

Current IETF activities around SIP:

1. Presence standardization will be finished in June 2007
2. ENUM - IP connectivity with E.164 numbers
3. SPEERMINT - peering policies between VoIP providers
4. P2P SIP - make SIP simple to use, solve scalability problems

So, what does the end-user want?

Connectivity via any device, anywhere, at anytime:

- global mobility
- personalization
- ease of use
- fair prices
- freedom to chose additional services

What do service providers want?

- Target end-users all over the world
- Small acquisition cost per end-user
- Minimum operational expenditure
- Keep pace with innovation and remain competitive
- Business scalability

How to deliver what users want?

Telecom or Internet approach?

PSTN is a centralized network where various elements are chained to control access and perform billing

NGN proposed by ITU-T, 3GPP IMS and its ETSI TISPAN extensions aim to implement the same model using Internet protocols

The ITU-T view on the network is:

- the network is application aware
- the control resides in the network
- NGN has many QoS definitions to “guarantee” for the network services

all ITU-T networks, such as ISDN, BISDN/ATM/NGN are based on grand designs and **not based on evolution.**

The changes from TDM to ATM to IP are significant discontinuities in the ITU-T architectures.

Internet is a dumb network, the services are performed at the edge based on the end-to-end principle

Internet reached current status through natural evolution

... **the end-to-end principle** is the most fundamental Internet's architectural principles:

1. Network is unaware of the applications, this makes it scalable
2. Nothing should be done in the network that can be done in an end-system
3. Quality of Service (QoS) is a matter of bandwidth availability and not of central control

On the PSTN, voice services follow a "**create and manage bottleneck**" architecture, where arbitrary points of control are created to justify business models based on resource scarcity

On Internet, services are available on the edge, can spread virally, business become successful only if end-users find the services or applications interesting.

Internet is an "**eat all you can**" model based on flat fees

For real-time communication applications like VoIP, a client needs only:

1. **Global routable identity (DNS, ENUM)**
2. Address translation (SIP, NAT)

New business models are based on **context aware communications** real-time communications linked with subscriptions/notifications to events from personal and business spheres.

Presence framework developed by the IETF SIMPLE working group enables the publication, subscription and notification to a broad range of information

Presence enables context aware communications:
the online/offline buddy status is the simplest example.

- enables mobility and tele-working
- integration between VoIP and World Wide Web
- ticket sale start for events like entertainment
- check stock items at your supplier
- communicate efficiently based on availability and environment

Presence and SIP parallel forking **are not "channel based".** Both concepts fall out of the telephone switch design and all the NGN components that have been built around managing channels (like Session Border Controllers)

- 1 to N messaging model
- Reach end-points in parallel on different networks
- Frequent updates to multiple recipients

"Presence is the dial-tone of the 21-st century"

Henry Sinnreich, the godfather of SIP

To roll out new business models you need:

1. New strategy, marketing and sales vision
2. Scalable solutions for SIP Proxy, Presence and NAT traversal
3. Possibility to combine hardware and software SIP applications
4. **Understand what end-users want**

New business models target global Internet, are designed to scale to millions of subscribers (like Google, Yahoo, MSN and of course Skype)

any technical bottleneck can hurt new business models

Some are building IMS (Internet Multimedia Subsystem)

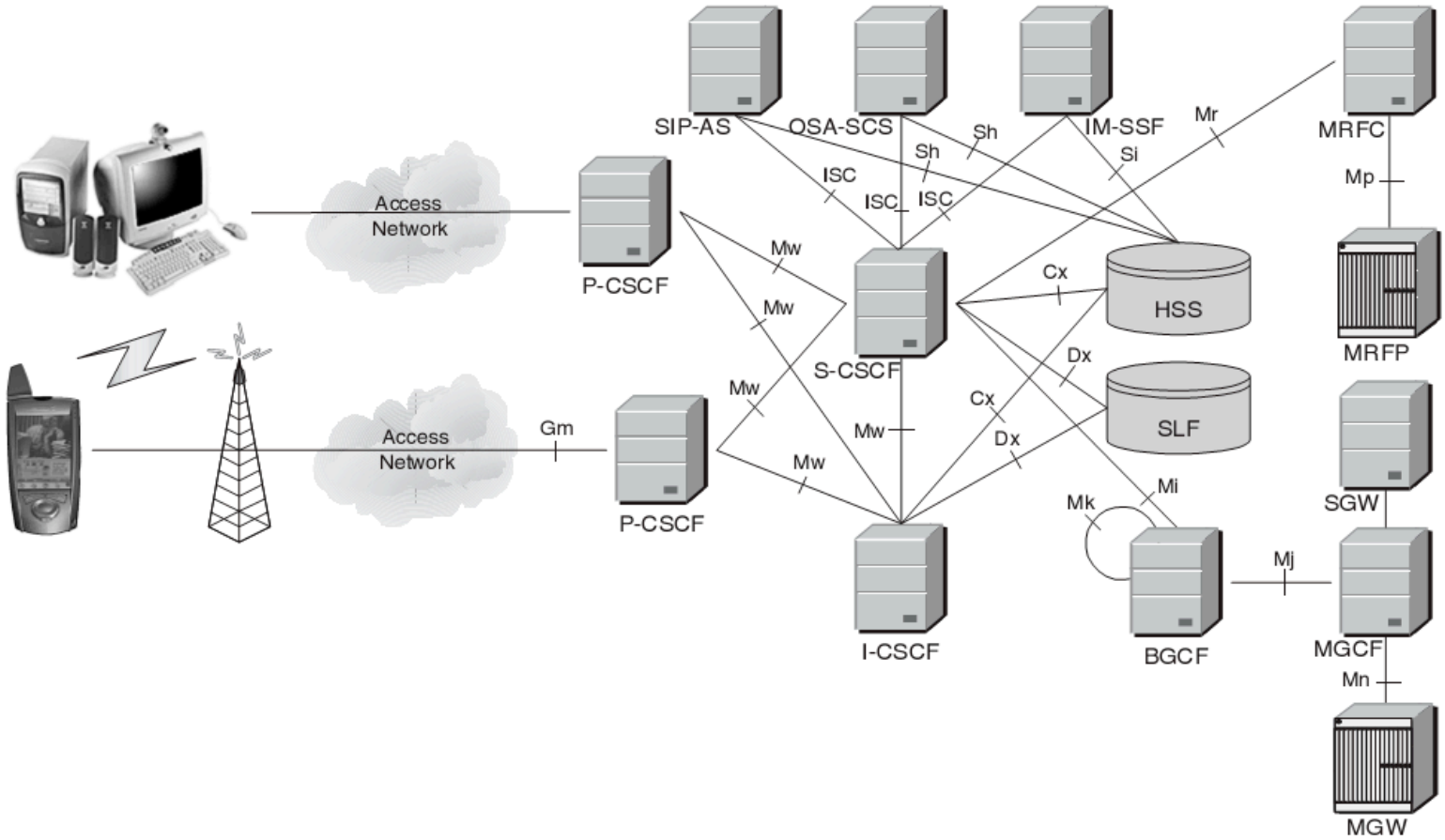
What is IMS?

IMS architecture - a crash course reality tour

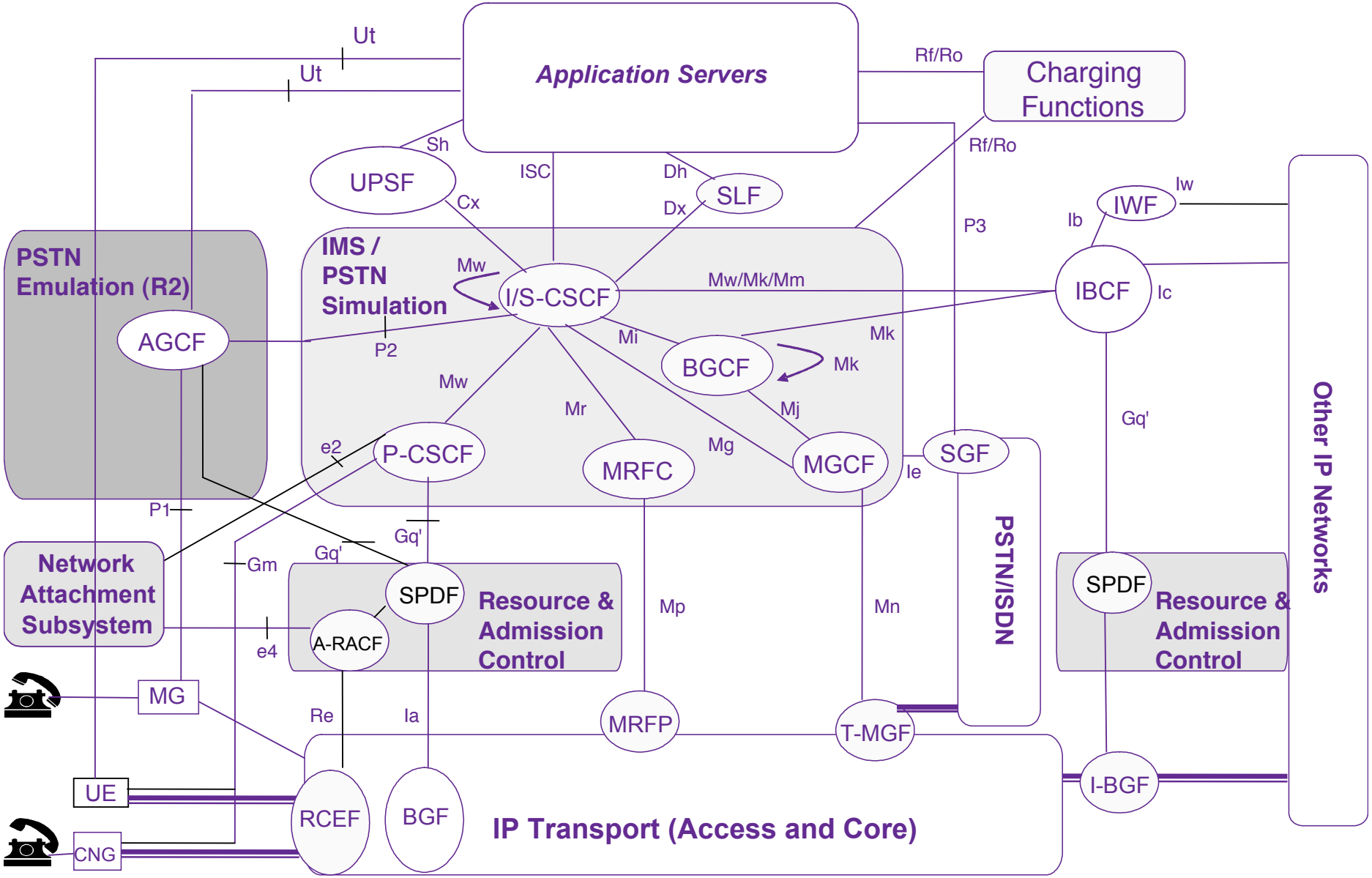
- Classic telephony design, enforcing resource scarcity
- End-to-end communication not possible between end-points
- At least 12 components with at least 22 interfaces, a forest of black-boxes that introduce delays, bottlenecks and interoperability problems
- SIP technology artificially incompatible between vendors
- Upgrade of all components is needed to allow new applications in the end-points, because IMS network is aware of applications
- Must be kept in a walled garden, software is not Internet proof
- **Last but not least, difficult to scale**

Try to build this:

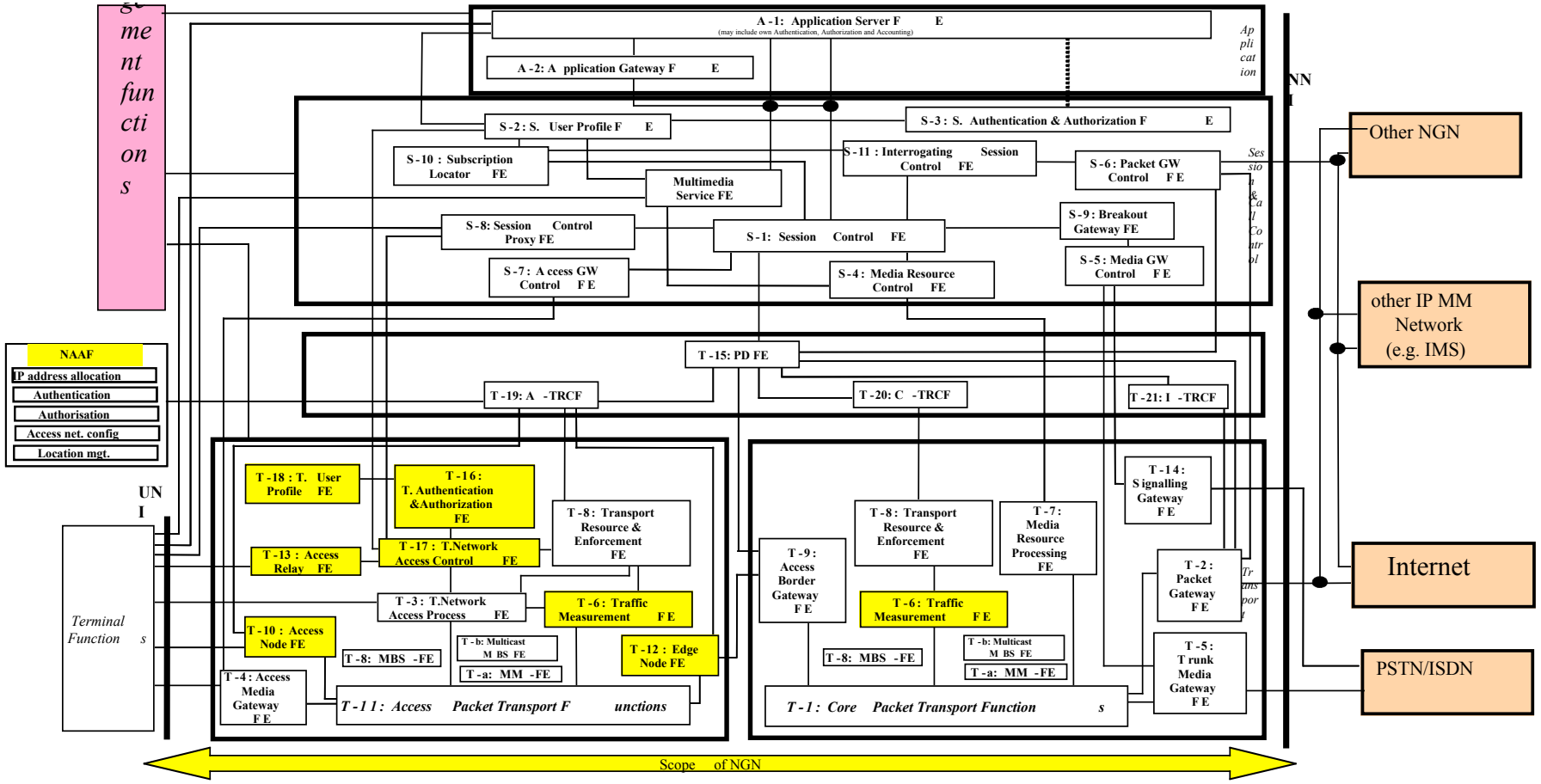
3GPP IMS Architecture – all subsystems



ETSI TISPAN IMS Architecture



both based on ITU-T NGN System Architecture



More than seven years old, IMS is still not deployed, while its release train lengthens with changes and expansions (IPTV is next). IMS complexity results from:

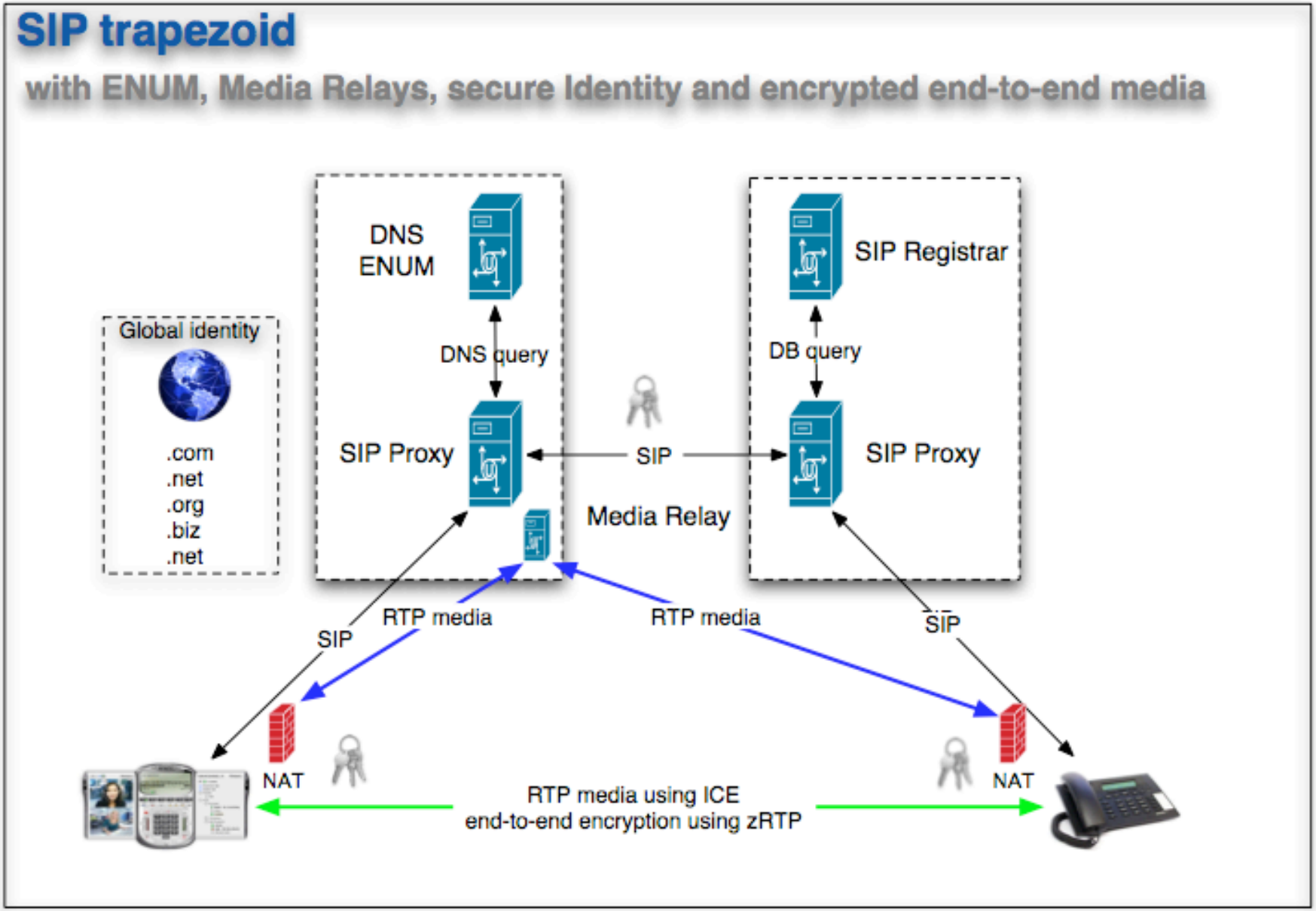
- decomposing devices into most granular functions and links
- tracking and controlling user behavior

The proliferation of boxes and protocols for the state management required for data tracking **lead to cognitive overload but add little value.**

When ready, IMS will provide less services than people got from Skype yesterday, **but at a huge cost.**

Are your customers willing to pay for this?

Many operators like Cable companies and Internet service providers have chosen a common-sense and cost effective solution, the basic IETF model:


SIP trapezoid
with ENUM, Media Relays, secure Identity and encrypted end-to-end media
DNS
ENUM
DNS query
SIP Registrar
DB query
Global identity
.com
.net
.org
.biz
.net
SIP Proxy
SIP
SIP Proxy
Media Relay
SIP
RTP media
RTP media
SIP
NAT
NAT
RTP media using ICE
end-to-end encryption using zRTP

IETF model is simple to understand and build. You only need to address:

- Business model
- Secure identity
- NAT traversal
- **How to scale**

Business model?

- Basic VoIP functionality is free
- Pay for the SIP device and get a free software for computers
- Pay for identity (Domain name or SIM card)
- Pay for value-added services

How to solve the scalability problem?

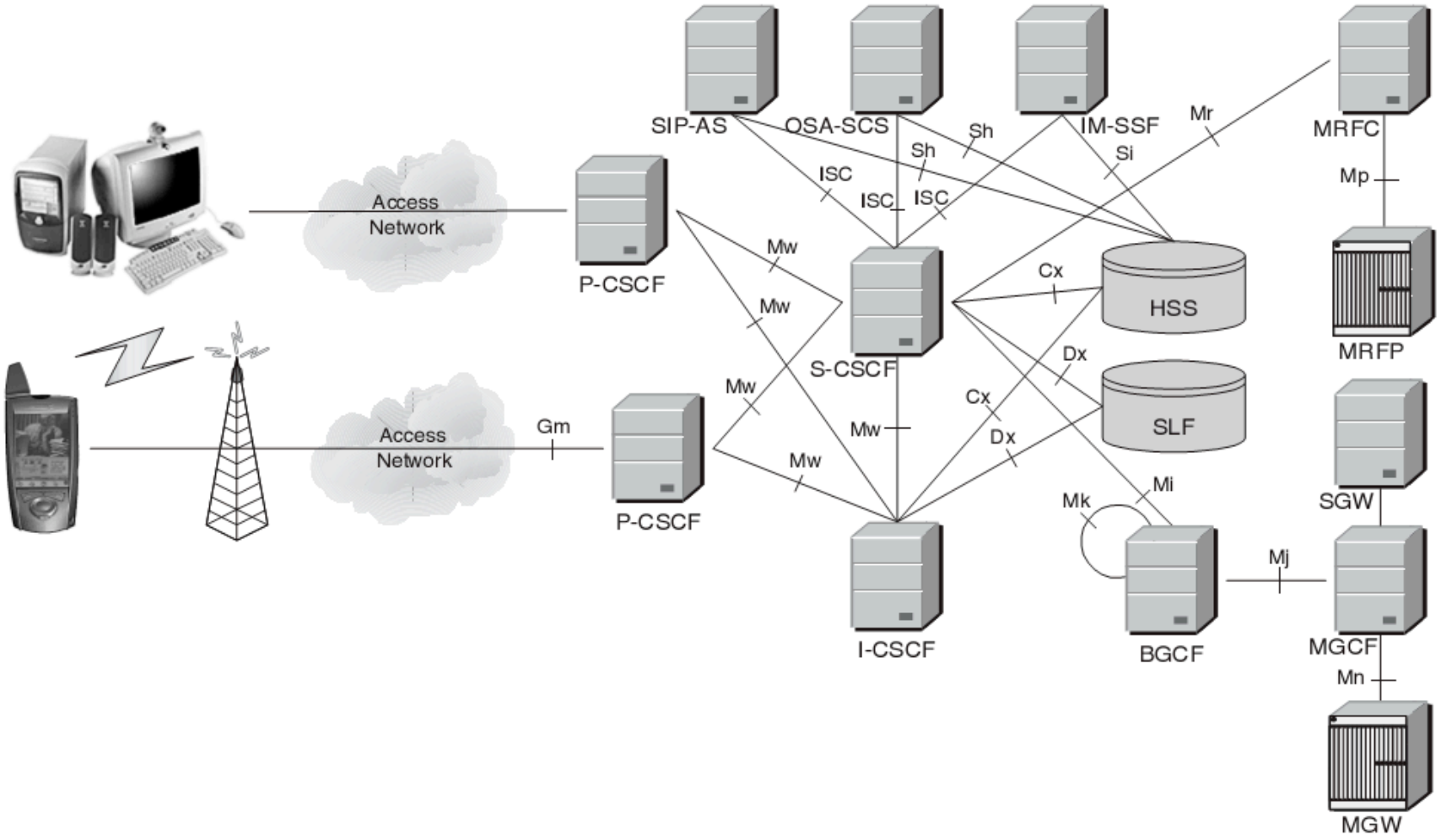
How to avoid scaling up your cost of operations while connecting more and more end-users?

Today we have enough computing power to run on a single server:

1. SIP Proxy and Registrar
2. Media relay
3. DNS and ENUM
4. Presence and XCAP
5. User profile

A single server can handle 10K users and 200 media sessions!

Take only the IMS functions that you really need

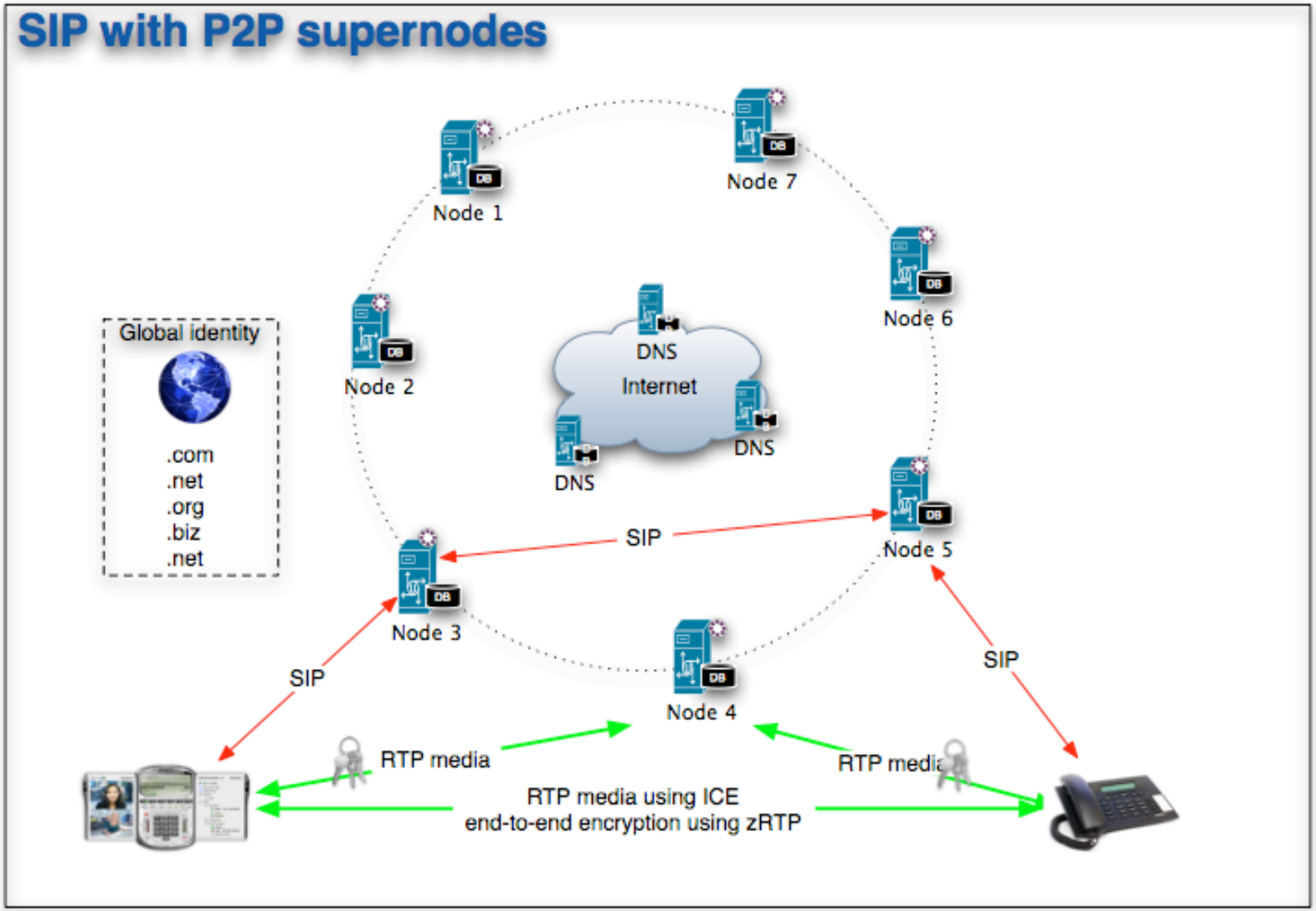


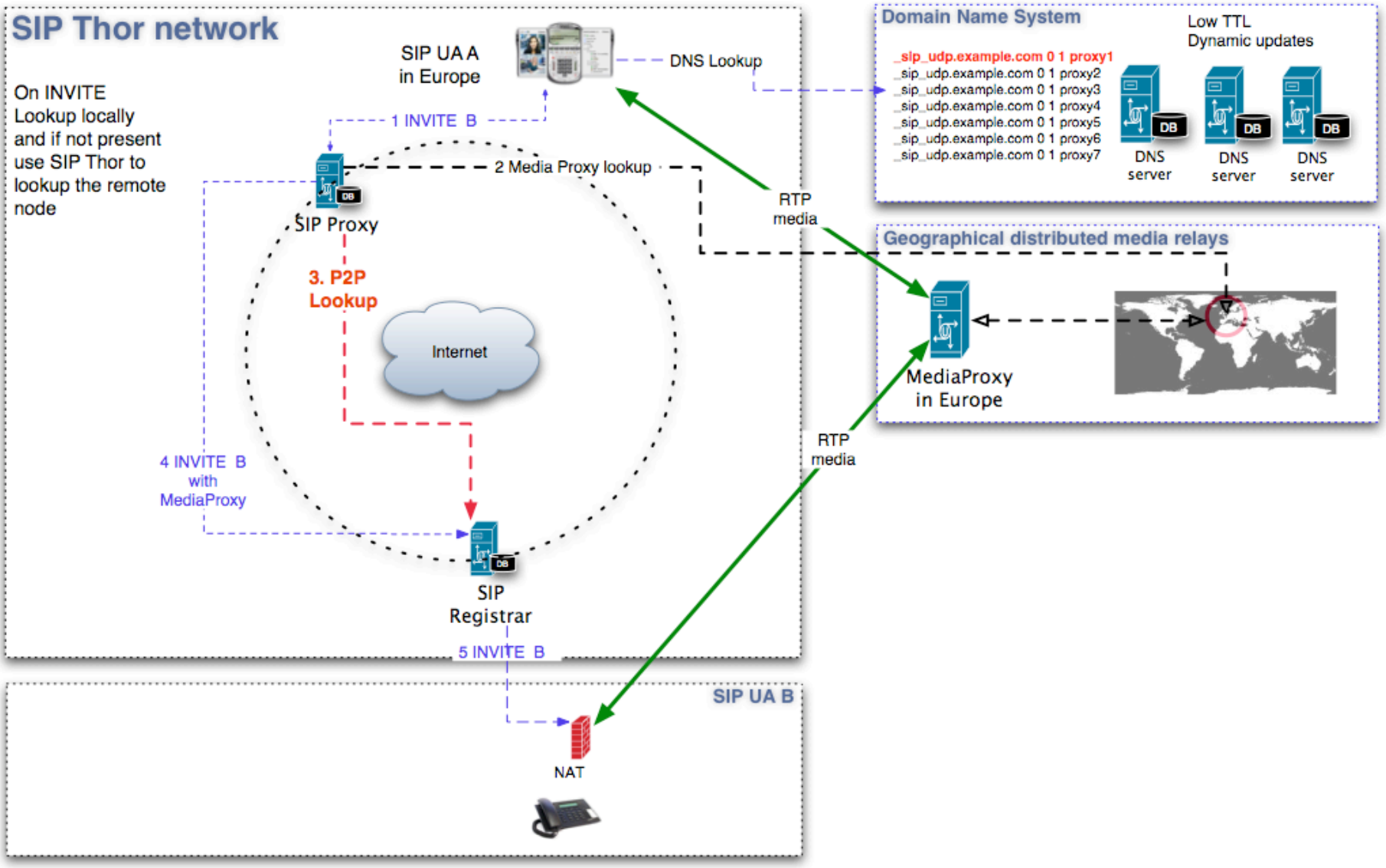
Fold all functions into a single box:

IMS
in a box

Play the hat-trick:

Scale the "**IMS in a Box**" using a self-organizing overlay network by using Peer-to-Peer technology





Thank you,

Adrian Georgescu
ag@ag-projects.com
http://ag-projects.com