# THE FUTURE MOBILE INFRASTRUCTURE: CHALLENGES AND OPPORTUNITIES

FLAVIO BONOMI, CISCO SYSTEMS

## INTRODUCTION

Facing an explosion in traffic demand, driven by the evolution of smart mobile devices exploiting a flat charging model, the infrastructure supporting mobile devices is presented with formidable challenges, but also creative opportunities, which will help manifest the potential of present and future mobile networks. In this short column we explore, at a high level, a number of the challenges and opportunities associated with the mobile infrastructure. In our broad definition, we consider here as part of the mobile infrastructure all the wired networking, storage, and computing elements required to deliver a modern user experience to smart mobile devices (e.g., from base station processing elements, to the high-end IP routers in the core, to the centers supporting content delivery and application processing).

## KEY REQUIREMENTS FOR THE FUTURE MOBILE INFRASTRUCTURE

As we consider our recent, extraordinary experience with smart mobile devices, as well as our frustrations with their performance and unrealized potential, we can identify a number of areas where much technical focus should be concentrated. However, progress in the areas we highlight here require more than pure technical breakthroughs, but also advances at the business and regulatory levels.

A short list of interesting areas of investigation and progress, implying new and more sophisticated requirement for the future mobile infrastructure, includes the following:

- Support for broader (Internet) **mobility**
- More effective **traffic offloading**
- More advanced and flexible **identity management**
- Lower **latency**
- More distributed **application awareness**
- More distributed **application processing**

Most of these requirements are motivated by the need for an improved traditional mobile user experience, but also by the evolving demands of sensor networks at the fringes of the current mobile network and the foreseen expansion of vehicular connectivity. In the next subsections we explore the requirements listed above in further detail.

### MOBILITY SUPPORT

The current generation of mobile devices, together with the current infrastructure, are far from supporting a broader notion of mobility, including seamless transition across non-homogeneous wireless technologies, such as third generation (3G) and WiMA, sometimes defined as true Internet mobility.

For a number of reasons, the current architectures support a somewhat seamless connectivity and some dimensions of application continuity only for cellular networks, via a somewhat centralized set of data centers, and through the notion of roaming. Enterprise networks also support, through their wireless controllers, a similar notion of mobility (at layer 2 of the open systems interconnection [OSI] stack). Furthermore, Session Initiation Protocol (SIP) supports a notion of mobility at the application level.

However, seamless movement from cellular to other networks is not happening, even in the case of femtocells. Thus, the achievement of the promise of pervasive seamless mobility across non-homogeneous networks still requires much work.

We see promise in the Internet Engineering Task Force (IETF) work on Mobile IP (in particular, Mobile IPv6, Dual Stack Mobile IP) [1, 2], and the work on the Mobile LISP Protocol [3].

A key functionality required in this area relates to connection management, which involves features residing at both the client and infrastructure levels. An intelligent connection manager can ensure connection continuity and optimization, thus greatly improving user experience.

A strong connection manager is especially needed to satisfy future requirements in the connected vehicle space.

### EFFECTIVE TRAFFIC OFFLOADING

Without doubt, the intense growth in traffic demand we have witnessed in recent years will continue into the future. It will be difficult for the mobile infrastructure to meet such explosive demand, even with the adoption of fourth-generation (4G) technologies and usage-based charging. For this reason, various approaches to traffic offloading will need to be explored. These will include, besides the femtocell approach, more widespread use of WiFi access in the home, enterprise, as well as public spaces. In particular, future roadside access points may provide efficient offloading for connected vehicles. Other wireless technologies may play a role complementary or alternative to WiFi (e.g., 60 GHz wireless).

The extent of service provider control over mobile user traffic as it spans multiple wireless networks is an interesting open issue.

Another very important offloading approach will involve more local communications, such as peer-to-peer communication at the periphery of the network (Nokia Paper) or more successful exploration of ad hoc networking.

### ADVANCED AND FLEXIBLE IDENTITY MANAGEMENT

The mobile infrastructure should try to expand its support for identity beyond the current well tried models, based on both hardware (i.e., SIM cards) or software (e.g., in code-division multiple access [CDMA]) approaches. There need to be better, more flexible ways to associate an identity with a user who may, for example, want to use the same identity to operate, interchangeably, a smart phone, vehicular phone, mobile tablet device, and body sensor. This more flexible identity management will open powerful opportunities, for example, in the evolving world of machine-to-machine communications. Further requirements in the direction of innovative identity management come from connected vehicles and the developing world.

### LOWER LATENCY

More than increased bandwidth, long latency is a fundamental challenge for the mobile infrastructure, as explored by Victor Bahl *et al.* [4]. Long latency is caused by a number of technical factors, including:

- Networking delays to and from a sparse set of mobile data centers
- A backhaul infrastructure based on a rather cumbersome set of tunneling techniques and a large number of hops, involving long queuing and packet processing delays
- Delays involved in loss correction (i.e., retransmissions) in the radio access network
- Data center processing delays

Long latencies in the mobile core inevitably lead to poor utilization of the available bandwidth, particularly for TCP connections, and, eventually, to a frustrating mobile user experience in many congested urban areas. It also limits the types of applications on smart devices that can make efficient use of "cloud" resources, thus limiting the scope of supported applications, and even cloud processing support aiming at extending mobile battery life. This issue is well articulated by Victor Bahl *et al.* in [5].

In order to face the all important latency issue, the mobile infrastructure should move quickly toward a more efficient network architecture, with fewer gateways between different networking technologies, tunneling levels, and hops. More effective latency-aware congestion management needs to be adopted, together with a distribution of smaller mobile data centers closer to the edge of the network. The adoption of more distributed, smaller data centers may also have positive cost implications [6].

More efficient transport models need to be explored, with the goal of avoiding as much as possible the inefficiencies of long end-to-end TCP connections. Advances in transport optimization will always be needed in order to take full advantage of the available bandwidth in the radio access network.

## DISTRIBUTED APPLICATION AWARENESS

The mobile network needs to become more content- and application-aware, at various levels of its infrastructure (e.g., mobile device, base station, gateway general packet radio service service node [GGSN], data center) in order to achieve higher functionality and efficiency. Content inspection and application classification can enable a higher degree of security, context-aware policies, quality of service, selective traffic optimization, and load balancing. Video traffic, which constitutes an ever growing percentage of mobile traffic, will need to be recognized throughout the network and specifically handled.

Finally, application processing, as we describe in the next subsection, may take place in a more distributed fashion. In order to achieve this more pervasive processing, the mobile network will need to become more application-aware.

## DISTRIBUTED APPLICATION PROCESSING

Many are the motivations for the exploration of a much more distributed architecture for application processing in future mobile networks. Among those motivations, we would include reduction of latency, more effective handling of sensor and emergency data, enabling new applications such as interactive gaming, more optimized handling of video traffic, and the potential for closer support of mobile devices, with the important goal of extending their battery life [5]. Moving in this direction will definitely not be an easy process, but may prove extremely rewarding. More distributed mobile application processing will leverage virtualization, require new middleware evolution, and definitely be facilitated by the diffusion of processors with homogeneous instruction sets from the cloud servers space to the embedded and handheld space.

## CONCLUSIONS

In this brief note we have listed and discussed a number of broad challenges and opportunities facing the future mobile network infrastructure. Meeting these challenges will require a deeply technical as well as regulatory and business effort. However, the results of progress on many of the items discussed here will open new vistas and enable a deeper, even more positive impact of technology on society and the environment. A close partnership between academia and industry will be required for more successful exploration of the issues discussed here.

## REFERENCES

[1] http://www.ietf.org/rfc/rfc3775.txt and related docs.

[2] http://tools.ietf.org/html/draft-ietf-mip6-dsmip-problem-01 and related docs.

[3] http://tools.ietf.org/html/draft-meyer-lisp-mn-03 and related docs.

[4] J. Huang *et al.*, "Anatomizing Application Performance Differences on Smartphones," *ACM MobiSys 2010*, San Francisco, California, June 2010.

[5] E. Cuervoy *et al.*, "MAUI: Making Smartphones Last Longer with Code Offload," *ACM MobiSys 2010*, San Francisco, California, June 2010.

[6] K. Church, A. Greenberg, and J. Hamilton, "On Delivering Embarrassingly Distributed Cloud Services," *ACM SIGCOMM Hotnets VII*, Calgary Alberta, Oct.. 2008.