Directions for
Future Cellular Mobile Network Architecture

Byoung-Jo "J" Kim, Paul S. Henry

Abstract— **Despite the extraordinary success of the cellular mobile telecommunications industry, many of the underlying design strategies and service assumptions that have served us arguably well over the past four decades may benefit from a fresh new look. Even today's LTE (Long-Term Evolution), which is designed to meet the demands of modern broadband Internet packet connectivity, nonetheless draws heavily on the legacy of cellular's circuit-oriented origins. Its heavy reliance on fine-grained tunnels and hard-state signaling protocols, for example, imposes performance penalties and cost burdens that may not be inevitable if hard-earned lessons are incorporated in the coming years. In this paper we describe a fresh approach to cellular network architecture. Inspired by past and present ideas and experiences by others and ours, we propose fundamental principles to guide the development of efficient and flexible network architecture, able to serve the still-unknown needs and preferences of future users. We offer an example of a network built on those principles and suggest how we can manage the evolution from today's networks to an architecture better suited to the decades ahead.**

I. INTRODUCTION

Current cellular mobile networks and the devices that connect to them are based on specifications from 3GPP (3rd Generation Partnership Project). Over the last few decades, 3GPP along with GSMA (GSM Association) has produced specifications that govern all aspects of cellular mobile network implementation, from radio interface up through the higher-layer operations, such as mobility management, accounting, QoS, roaming, IMS, and many applications. Despite the enormous success of the cellular mobile industry worldwide, the current architecture shows many signs of fundamental problems.

For example, the difficulty and delay in producing a working specification for IPv6, particularly IPv4/v6 coexistence over a cellular network, is a striking example of the high complexity of the current network architecture [1]. Even now, several of the essential features of IPv6 such as network renumbering and multi-homing are not possible in cellular networks and the essential Neighbor Discovery and Duplicate address detection [2] processes require special interventions not needed in most other networking technologies such as Ethernet. The root cause of this phenomenon is the lack of layer 2 link identity and absence of link-local multicast, due to the legacy layer 2 link model that relies on point-to-point tunnel links.

Similar difficulties are encountered when session continuity is desired between WCDMA/HSPA and LTE cellular data networks, and between cellular and Wi-Fi networks. Local traffic off-loading, desired for reducing traffic load in cellular core networks and accessing local network resources, is a relatively complicated update to specifications that touches many layers of both the network and the user device. New standards had to be hastily created to support femtocell networks and enterprise-hosted cellular networks, instead of developing them as different implementations of a single flexible architecture. These features may be supported in the end, but with a great deal of complexity, delay, and cost that may not be truly necessary.

Over the last few years, the cellular network industry has been struggling to cope with the increasing data demands of new devices like Smartphones and tablets. However, it has not been able to fully take advantage of the higher transmission rates available from the new radio interfaces and the rapid performance/cost improvements of the wireline transmission technologies that connect the rest of the cellular network beyond radio links. Among the key reasons for this shortfall is the high latency caused by cellular networks' reliance on highly centralized custom routers and other packet processing devices. They run memory-intensive and computationally heavy protocols, ultimately leading to high latency and high cost.

In this paper, we will discuss the basic architectural structure and design process of cellular network standards that manifest themselves in these and other symptoms[1]. Then we will propose a set of design principles to avoid the similar pitfalls and a high-level example that follows from these principles. The implications of the new directions and some recommendations for achieving them will be discussed.

II. ANALYSIS OF CURRENT AND FUTURE 3GPP CELLULAR NETWORK ARCHITECTURE

Starting from GPRS (General Packet Radio Service) and including LTE (Long-Term Evolution) [3] [4], cellular data network architecture has employed the same basic approach to transporting user data traffic: tunneling over diverse lower-layer transport protocols to and from a centralized gateway. Details vary depending on the technology generation, but the basic operation remains the same. In the downstream direction from the Internet to a user device, user IP packets are fragmented as needed and encapsulated in GTP (GPRS Tunneling Protocol) tunnels over UDP/IP from a gateway. (In the case of the U.S., usually placed at a handful of data centers (DC) around the country.)

In LTE as shown in Figure 1, these encapsulated packets travel between a gateway (PDN-GW: Packet Data Network-GateWay) and a base station (eNodeB, a collection of which is called UTRAN in figure 1). The Serving GW may be combined with the PDN-GW under non-roaming scenarios. Otherwise it acts as forwarder to home networks for roaming UE's (User Equipment: 3GPP terminology for devices used by subscribers). In WCDMA/HSPA, in addition to a gateway (GGSN: Gateway GPRS Support Node) and a base station (NodeB), the encapsulation is translated into lower layer fragmentation packet formats at SGSN (Serving GPRS Support Node) and RNC (Radio Network Controller) that relay between GGSN and NodeB. (In this case, UTRAN is a collection of SGSN's, RNC's, and NodeB's.)

The extensive use of tunneling was perhaps necessary in the early designs of cellular data networks, due to lack of extensive wide-area packet transport at the time. Also, the designers were familiar with the circuit protocols used for cellular voice, and built the data network as an overlay. Cellular data, after all, was an unproven add-on to highly successful voice services at very low rates of less than 20 kbps. The wired Internet was still not widely popular at the time. Thus, user level tunnels became the foundation of cellular data networks, being used for mobility, policy control, routing, QoS, service redirection, and numerous other functions.

Tunneling, in itself, is not inherently a bad design choice, given the capabilities of modern router/switch systems, and is often employed in other networking services, e.g., VPN, MPLS, provider Ethernet, etc. The issues with tunneling in 3GPP networks arise from the very fine tunnel granularity[2], potentially very long transit distances, massive concentration of tunnels to a small number of gateways, overloading the protocol with multiple functions such as QoS, policy, charging, and roaming. Particularly, the high number of states that must be maintained throughout the cellular network, as we will see later, becomes a barrier to scaling and innovation. Normally, tunneling protocols concern only the two end-points, and other network elements in the path are supposed to be stateless with respect to tunnels. However, this is not the case for cellular networks due to the overloading of user-level tunnels with many extra functions that are statefully processed by most intermediate nodes[3]. This set of protocols is addressed by a single all-purpose namespace APN (Access Point Name) [4], using more than 200 different messages for many different functionalities.

---

[1] For an interesting discussion on the future of the PHY layer of wireless communications, refer to [5].

[2] Multiple tunnels per user device to different gateways.

[3] The high number of network elements handling user packets may aggravate the issue of Buffer-Bloat [9] [28]that manifests as large latency under mild congestion.

Beyond the above direct observation of the downside of the tunnel-centric architecture, there are more subtle, yet equally important issues that arise from the current architecture and the associated standards-setting process.

A. Mismatch of architectural assumptions and reality

With the advent of Smartphones, it is not surprising that data traffic demand has risen substantially. In addition, an unexpected side-effect has presented itself due to the interactions between cellular network architecture and the need for constant connectivity by many applications on Smartphones. Apple Push Notification (not to be confused with the Access Point Name) for example, and other push services over IP have led to Smartphones maintaining persistent TCP connections to one or more servers in the Internet outside cellular networks[4]. The OS and application developers applied their experience with wired Internet services of similar nature, and they did not have an alternative reachability mechanism due to the extensive use of NAPTs (Network Address & Port Translator) and the use of private addresses inside cellular networks. These largely idle connections from a traffic perspective are now straining the gateways and other network elements such as firewalls that must keep track of APN/PDP/GTP/IP connections at all times. The tunnels meant for active-session mobility are now frequently used to maintain idle connections for mostly stationary devices. Under the current architecture, new devices such as tags, monitoring devices, meters, etc., would need permanent APN connections at all times or go through expensive network entry signaling to send a few bytes each time: a large departure from the past design assumptions of cellular networks[5] and a significant barrier to scaling the network to handle a far higher number of devices.

This is one of many symptoms of the mismatch between reality and the architectural assumptions/decisions made long ago that cannot easily be updated. Other examples include IMS signaling issues through firewalls, the complicated QoS architecture that are often not deployed, the difficulty of

---

[4] Also, recently popular AJAX web apps (such as Gmail, & twitter) tend to have long-lived connections with irregular exchanges of small packets. They can reconnect after IP network handoffs like push notification connections.

[5] that often assumed a small portion of subscribers are active at any given moment for durations that follow an exponential distribution. From the tunnel maintenance perspective, this must be now replaced with almost all subscribers having active tunnels almost at all times.

achieving local breakout routing, and handovers between different generations of cellular radio interfaces or Wi-Fi hotspots under separate administrative control.

B. Route inefficiency

It is not difficult to imagine that tunneled traffic may take less efficient routes than normally routed traffic. As a generic example, depending on the location of the GGSN/PDN-GW that terminates a particular APN, a Smartphone in New York City connecting to a web server in Atlanta may connect via a gateway in Chicago. In parts of many cellular networks in the U.S., the regional and backbone networks connect base stations to a MTSO (Mobile Telecom Switching Office), the MTSO connects to large data centers (DC), and finally the user traffic leaves the DC to enter the general Internet via an IP backbone network. At each leg, the traffic enters and leaves regional/backbone networks, as the cellular network has been using regional/backbone networks as a customer of the networks. This is caused by the combination of the tunneled nature of cellular data traffic and the history of gradual builds, acquisitions, and transitions. Ideally, the user traffic should enter and leave the IP backbone network only once.

Another issue of particular importance is the placement of CDN's (Content Distribution Networks) in mobile networks. The role of the CDN in the Internet is now essential[6] and will continue to be so [23]. However, its effectiveness has been limited in cellular networks, since they cannot be closer to user devices than PDN-GW/GGSN which is already quite far away from most end users. The centralized handling of user transport tunnels makes it difficult to benefit from CDN's, since all traffic must arrive at a DC regardless of its ultimate destination. Also, due to the use of point-to-point user tunnels in cellular networks, it is impossible to take advantage of IP multicast over the wired part of cellular RAN[7] (Radio Access Network) for potential applications for which multicast over wired network may be highly effective [23].

---

[6] Particularly for video distributions.

[7] Multicast over wireless links for user packets is often not useful due to vastly different channel conditions experienced by different mobile devices. Depending on the number of devices in a multicast group and their geographic distributions, multi-cell simulcast or unicast to individual devices may be more efficient.

C. *Requirement creep during standardization*

3GPP and many other standards organizations adopt similar processes for creating new or modified standards, typically involving the following steps: identifying usage models, agreeing on requirements for the usage models, creating a high level architecture that satisfies the requirements, and finally specifying detailed protocols. This approach has seen many successes in technologies with near-term horizons, well-understood markets, and limited complexities and interactions. Examples include Wi-Fi interface standards and PC component standards. However, cellular network architecture standards have been far harder to update and replace, since they tend to stay in service for long periods of time due to cost and build-out delays. Consequently, they demand backward compatibility and coexistence from any new ideas. Thus, any new features that are needed to respond to changing market conditions are burdened with compatibility requirements for numerous functions that have come before. Also, many new features with unproven benefit to most users are continually added to the overall architecture, in part due to the lack of realistic ways to prove their utility before standards are in fact set for them. This requirement creep is one of the major sources of complexity and the slow pace of evolution of cellular network standards, since there is no empirical way to accept or reject them based on their feasibility and utility.

D. *High complexity and its impact on competition, cost, scalability, and security*

Along with the well-known observation, “Complexity is the enemy of security”, the high complexity of cellular network standards, especially the backhaul/core networking and higher layers above the radio links, has been a barrier to competition in the network infrastructure market. The lack of competition, along with the high complexity, naturally leads to high cost of gear compared to similar devices used in wired networking. For example, a GGSN device costs a few orders of magnitude more than a normal router based on the same hardware and software platform, while handling far fewer packets and flows per second. Combined with the recent rapid growth of mobile wireless traffic, it is becoming increasingly difficult to scale the cellular network in a cost-effective manner. The security implication is even direr, as poorly-tested implementations of highly complex protocols and service logic are exposed to increasing interest from hacker communities, thanks to the rising popularity of wireless internet and Smartphone ecosystems [7] [26][27]. Particularly, the large number of states maintained in the network for each mobile device could become a target for DoS attacks using a small number of compromised devices.

III. DESIGN PRINCIPLES

Based on the history of mobile cellular networks and the Internet, the following set of design principles should be kept in mind for future network evolution.

A. *Alignment with the Internet architecture*

It can be argued that the current cellular data network still employs circuit-like transport for user traffic, albeit running on packet transports, through the use of constructs such as APN and GTP. Future network architecture should be a truly packet network that does not involve semi-permanent per-service tunnels and heavy reliance on network intelligence. Any protocol or service that is not proven necessary for the operation of a commercially-viable IP network should have to satisfy a very high threshold for its necessity in order to be included in the basic architecture.

B. *Endpoint-centric protocols*

With multiple radio interfaces available on user devices and the reality of many different generations of radio networks under different administrative control, it should be clear that most network protocols should be controlled primarily by the endpoint devices, since they have the most accurate view of their radio connectivity and characteristics of applications and services on them. The current architecture strives very hard to handle this reality from the network side and is sometimes failing to meet real needs in terms of scale, cost, performance, and manageability. An example is the prolonged activity around creating network-based selective IP flow mobility between cellular and Wi-Fi networks, i.e., Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO) [8]. With the advent of Smartphones, the increased complexity of endpoint devices is a foregone conclusion; thus the cellular mobile network should follow the principle that was proven successful in the Internet: most intelligence should be at the network edges and host devices.

C. Simplicity

The current cellular network architecture employs many protocols that involve 4 or more parties and numerous round trips. The initial entry and connection establishment involves a mobile device, base station, RNC, SGSN, GGSN, and HLR, as described in the standards. In addition, firewalls, NAPTs, accounting systems, and QoS policy servers are often involved. There are 10s of roundtrips depending on previous states required to complete this process. It is doubful that this is an ideal or unavoidable situation, yet some aspects appear inherent in the current architecture. Also, the messages carried in these signaling flows are complex, overloaded and/or nested in multiple layers, so that the correctness of any implementation is doubtful at best. It also presents a rich target for hacking or DoS attacks. Thus, simplicity of protocols should be one of the prime objectives of any new protocol design, along with modularity of protocols and soft-failure under errors.

D. Designing for uncertainty

Most of 3GPPs recent activities, such as latency reduction, local traffic offloading, flattening the network architecture, and peak rate improvements, have been in reaction to seemingly unexpected growth of demands from reality: web centric content, Smartphones, push applications, etc. It should be obvious that the basic network architecture standards should be designed for uncertainty and flexibility [24], rather than specific service scenarios. Also, considering the long delays in responding to market demands through the standard-setting process, the basic standards should be independent of the details of features and services, even if such a separation sacrifices some of the benefits of tight integration. The use of host-centric protocols is also consistent with this principle, since new services or features can typically be implemented on host devices and servers, and should not involve changes deep inside networks.

IV. LONG-TERM ARCHITECTURAL DIRECTIONS

A. *Engineering for low latency*

With the recent increase of the aggregate throughput and the peak rates of radio links, the high latency of cellular data networks has become a major performance issue. The use of layer 2 tunnels, connection-oriented centralized routing, protocol encapsulations and translations, and complex protocols have all contributed to the high latency that is now making the improvements in bandwidth less useful and noticeable. Due to high latency, most short TCP connections (that comprise most interactive connections) cannot fully utilize the higher peak rates available with HSPA and LTE. Under high latency, the impact of dropped packets (while very rare in cellular networks by design) is very severe, congestion collapse is more likely, TCP congestion control becomes less effective, and temporary performance disparity among users is more pronounced. VoIP and other real-time applications show noticeably poor performance compared to the wired Internet, even when the cellular throughput for bulk transfers approaches that of some wired broadband access technologies. Any aspect of standards, protocol design and implementation, or network architecture/layout that may increase latency must be scrutinized with the utmost skepticism in order to be included in the next-generation architecture.

B. *Elimination of ubiquitous usage of layer-2 user plane tunnels*

Layer 2 PDP tunnels for User Equipment (UE) traffic, running over providers' IP networks, span multi-state distances to reach a single point deep in the network, regardless of their eventual destinations. Considered necessary in the past for large-area mobility, operator control, and roaming, they have increasingly become a source of inefficiency, poor scaling, and complexity. Alternative approaches as suggested below may better address the needs of present and future cellular networks.

C. *Direct use of IP routing and Metro-Ethernet switching*

Instead of tunnels, direct use of Ethernet switching in small areas and plain IP routing over bigger areas, similar to the rest of the Internet, could harness the innovations and competition existing in the wired networking industry. The basic functions that tunnels used to provide, such as mobility, transport over circuit networks, and route control, are now unnecessary, or can be provided by other means as described below. The base stations would handle localized fast mobility in a distributed manner, without centralized control. The LTE EPC specification creates a separate entity called MME (Mobility Management Entity) specifically to handle only signaling for RAN (Radio Access Network) mobility. This is a step in the right direction, but having layer 2 mobility managed by base stations, rather than the MME, may be the ultimate solution.

Also, new layer-2 and layer-3 protocols such as TRILL [11], IEEE 802.1aq [12], LISP [13] and Serval [29] may be used to reduce the need to develop mobility-specific protocols and devices for cellular networks [20]. With the inevitable trend toward using a large number of small cells in the future, more distributed protocols and technologies should be favored, even if they are not needed at present.

D. Elimination of private layer-3 networks and private IP addresses

Current cellular core networks are largely private networks, with user devices typically given dynamic private IP addresses. They access operator-provided services within private networks and access the rest of the Internet via multiple layers of NAPTs. In future architectures, operator-provided services should be placed on the public Internet (beyond the NAPT as NAPT for IPv4 may still be in use for some time.) and thus be accessible from anywhere without the use of tunnels and special arrangements between roaming partners. Most of the widely-believed benefits, such as the alleged greater security of using private networks, are either untrue, not worth the price, or can be achieved via different means[9].

E. Separation of basic network layer and service overlay

Many decisions made about cellular network architecture in the early days have remained in effect and influenced many subsequent changes and additions with numerous unintended interactions and limitations.

---

[9] For example, the majority of host security compromises are via application/OS vulnerabilities that being on a private network does not prevent. NAPT traversal and Push techniques allow hosts to be contacted from outside their private networks. Access control to special services is best done by application-level authentication, rather than identifying source networks of user traffic.

Considering our general inability to predict technology and economic changes for any significant stretch of time, the basic network architecture should not be designed based on any set of specific usage models or requirements that might be asserted at design time. Instead, the architecture should be designed to allow quick addition and removal of overlay of protocols and services for diverse applications[10].

F. *Host-based on-demand IP mobility, instead of network-based always-on IP mobility*

At present, the network handles most aspects of mobility and does so at all times for all traffic, at great cost and complexity. However, it is becoming clear that not all traffic flows need such services, and with multiple radio interfaces like public/home Wi-Fi in addition to cellular radio, the host/UE is clearly becoming the right place to determine and manage its IP mobility needs. Also, many devices use the network in a mostly static context, including some applications of M2M services. Treating every device uniformly for high speed mobility and always-on connectivity is clearly not optimal, considering the overhead needed for such features. The network need only provide IP mobility anchoring services as hosts request them on demand.

In principle, IP applications on user devices require two fundamental functions from the cellular networks they attach to: Session Continuity and Reachability. Session continuity refers to the survival of layer 2 and layer 3 connectivity across handoffs between base stations, while there is active traffic to and from user devices, such as streaming video, ftp, etc. This is what the cellular network provides by using the user-plane tunnels that combine layer 2 and 3 and follow wherever user devices go. In effect, user devices never see changes in their IP subnet over large geographical areas, unless they have to switch between different gateways, roaming partner networks or different wireless link technologies. Reachability is provided when a mobile device is reachable by its communication peers even when idle. For voice calls, this is achieved by paging, but for data apps, the current cellular data networks do not provide this functionality, because they use dynamic private addresses. Dynamic and/or private IP addresses have also been a reality in the wired Internet for most consumer connections, so the application developers have

---

[10] Which we argue is more plausible with endpoint-centric architecture, as evidenced by the success of the Internet.

typically solved this problem by maintaining persistent connections initiated by clients to a rendezvous server such as a chat server, push server, or AJAX server. This approach was later adopted by mobile OS and application providers and resulted in Microsoft ActiveSync, Apple Push Notification Services and other push services. This technique provides the appearance of reachability while the actual connections are initiated by clients and maintained through firewalls and NAPT by both communicating parties. Given this reality, and the use of persistent SIP/TCP connections for VoIP signaling, the reachability issue need no longer be solved in mobile networks through the use of Mobile IP or user plane layer 2 tunnels. It can be also argued that reachability is in fact better solved at the application/OS layer, since that is where the actual needs are understood, instead of at layer 3 or layer 2 network services. Given this reality, a new architecture should not suffer from excessive network control signaling triggered by these long-lived connections coming in and out of idle states, as the current architecture does [7]. One possible approach is for the network to be largely stateless beyond base stations with respect to user devices.

In light of this development, IP mobility should be per-flow on-demand rather than always on, and also locally distributed to allow flexible network deployment choices, considering various utilities of caching, CDN's, and peer-to-peer communications.

G. Substitute hard-state rigid protocols with soft-state, soft-fail protocols

As mentioned earlier, 3GPP standards rely heavily on hard-state multi-party protocols for most operations, such as mobility, paging, network entry, sleep/idle modes, and routing. These protocols often require that three or more parties remain synchronized in their protocol states for correct operation. Perhaps not surprisingly, these complex protocols are rarely verified for correctness and race conditions. They fail hard, recover slowly, require large memories and processing power, and are difficult to interoperate and debug. Most of them could well be replaced with soft-state, soft-fail protocols with better results.

H. Separation of air interface and mobility core network evolution

Each new generation of the cellular air interface has been accompanied by a large overhaul of the wired core and RAN. That approach appears no longer sustainable for keeping pace with different rates of innovation in wired networking, mobile computing, and wireless link technologies. Arguably, innovations and changes in the network and application layers are more rapid and unpredictable, as evidenced by the rapid rise of the Smartphone ecosystem, compared to wireless link layer evolution that proceeds more slowly and requires industry-wide coordination. This separation also widens the market for air interface technology beyond cellular mobile technologies into fixed wireless access, indoor private networks, etc. without being burdened by the necessary network elements dictated by the current standards. The current tight integration of standards for the air interface and the wired network should be separated, with clear interfaces to allow their separate evolution.

V. AN EXAMPLE ARCHITECTURE

Based on the preceding discussion, an example architecture for a future cellular network can be imagined in a straightforward manner. This of course is not the only possible solution, nor entirely original, but serves as a concrete example for discussion.

A. RAN connectivity

As more and more cellular base station sites are connected with high-speed backhaul to handle increasing demand, Ethernet is becoming the de facto choice. For this example, we assume all base stations are connected via wide-area Ethernet services commonly called Metro Ethernet[11]. They are connected via Metro LAN service with a topology that fits the local conditions rather than point-to-point logical connections between base stations and MTSO as done in the present architecture. On this layer 2 network there are one or more IP routers with a small set of mobility-related features, such as dynamic distributed IP

---

[11] Inevitable exceptions to using Ethernet transport can be handled via emulations such as pseudowire. Point-to-point microwave links are increasingly migrating to Ethernet transport, and most wireline access technologies now support Ethernet transport, if not Ethernet LAN services.

mobility, to be discussed later. This familiar IP subnet forms a unit of RAN and connects to the Internet directly, instead of connecting back to a cellular core network as is done currently. There is no cellular core network for user traffic. Instead of near-permanent tunnels for each user device, tunnels or VPN's are sparingly used for groups of devices, and only for special functions such as lawful intercept, hot-lining for emergency access, etc. This unit of RAN can be as big as a metropolitan area or as small as one base station. The same logical unit of RAN can be used to build a large city network, a venue network, or a small hotspot.

The current air interface technologies do not have MAC layer addresses for user devices that can be directly used to switch user packets in the RAN Ethernet connecting base stations. They can be added by user devices or by base stations, and there is little risk of address spoofing since the identities of user devices are verified cryptographically per existing air interface security measures[12].

New LAN technologies such as TRILL [11] and Shortest-Path Bridging [12] can be used to enhance the scalability, the efficiency, and the manageability of such RAN networks. They may even assist layer 2 mobility directly [20]. The handoffs between base stations are handled directly between base stations using Ethernet switching along with light-weight L2 (such as MACinMAC) temporary tunnels when needed for forwarding buffered packets.

It is commonly envisioned that there will be increased direct communication between base stations for various reasons such as multipoint coordinated transmission [21], or plain local user-to-user traffic. The tunnel-less dynamic auto-configuring layer 2 network described above would make transition to such scenarios much simpler.

B. *Dynamic distributed mobility*

Some user devices will cross the coverage boundary between two RAN units while still engaged in active data flow at the IP layer. Some will cross with data flows, such as push notification connections, that are

---

[12] Although it may not be absolutely necessary, the binding between MAC address and the IP address can be also be verified by base stations.

designed to survive IP network changes by reconnecting in the background. Some will cross while idle and some, such as meter readers and stationary modems, will never cross. The actual mixture of these different types will keep changing and the best place to understand the need is at the devices. In light of this observation, the most sensible and scalable IP mobility protocol should be host-based, on-demand, distributed mobility. (For a survey of IP mobility protocols, refer to [5].) One possible implementation is depicted in Figure 2-1 where 3 RAN units, each with its own router, are serving a geographic area. These RAN units can each be covering a city and its surrounding areas. A mobile device always uses the locally valid IP address for any new IP flows.

If there is an IP flow that must survive the subnet change as determined by the user device, the initial router A (called mAR A, mobility Access Router A, for convenience) acts as the HA (Home Agent [5]) for the flows to forward the IP packets belonging to the flow to the new RAN unit[13], RAN unit B as shown Figure 2-2. For subsequent handoffs (to RAN unit C for example) within the lifetime of the flow, mAR A remains as the anchor for the flow. This flow will remain anchored at the origin router A until the flow is terminated by the user device (and reinitiated if needed as a new flow in RAN unit B), or when a maximum life span is reached as determined by mAR A[14]. While this flow is anchored at the previous router A, any new flow originated by the user device uses a locally valid IP address corresponding to the new network B, thus, as shown in Figure 2-2, it is not tunneled. If this flow also requires session continuity while the user device moves to a new RAN unit C or a Wi-Fi hotspot, mAR B becomes the HA for this particular flow[15] as shown in Figure 2-3.

The key assumption for this particular example is that most long-lived user flows, if not all, are reachability flows that can reconnect after attaching to a new subnet. Other flows only need temporary session continuity (minutes at most) or not at all. Even an hour-long SIP VoIP call can be seamlessly

---

[13] We limit the description to IPv6. IPv4 can be handled similarly with MIPv4 via the use of co-located CoA mode or FA.

transferred to the new network using SIP mobility techniques [14]. Temporary session continuity maintains the call while necessary signaling is performed to transfer the call to the new IP address[16].

Observe that this distributed mobility management also uses tunnels, even per-flow tunnels in some cases, so it appears to be not much different from the current architecture at first glance. The key difference is that these tunnels are temporary, localized, and on-demand, instead of being permanent and global for every device. Even for highly mobile user devices, mobility tunnels are used temporarily only when they cross boundaries between potentially large RAN units[17].

One of the main sources of handoff delay across IP subnets is IP stack configuration in addition to security-related signaling, particularly when the UE can only communicate with one network at a time. There are many approaches to reduce this delay such as the use of Signal Forwarding Function [25] in mAR's and other radio interface-specific optimizations.

C. *Application layer service overlay*

Some features only available to the lower layers of cellular radio links, such as paging, may be of great use for some applications, if they can be safely exposed to the higher layers. For example, even when many devices handle their own reachability issues like most Smartphones, it may be desirable to keep some devices idle with no layer 3 connectivity and rely on layer 2 radio paging to wake them up for infrequent activities. An application on a user device can negotiate with a server that has backend connections with the cellular paging system and provide authorization for another trusted party from the Internet to trigger cellular paging. Details of such a system are omitted here for space concerns, but this service can bridge the application layer reachability and cellular radio paging capabilities to enable efficient operation for certain

---

[14] This would also force reconnection if necessary by the user device. Application and OS developers should be encouraged, by stopping forwarding after a reasonable delay, to design their protocol behaviors so that there are no long-running active flows after a subnet change. They should reconnect using their new local IP after the handoff tunnels bridge the transition.

[15] In IETF, there are early activities approaching similar ideas [15].

[16] Many optimizations for rapid IP layer handoff are possible using layer 2 indication mechanism such as IEEE 802.21, RFC 5184, and other layer 2 & 3 interaction approaches. Also, two neighboring RAN units can share base stations along their boundaries that handle subnet changes smoothly.

[17] For a small number of devices that require permanent mobility tunnels for some reason, a conventional centralized mobility anchor may be provided over IP, or over layer 2 VPN. They would need to understand the performance impacts from suboptimal routes, lack of local CDN access, etc.

applications. This approach to network features where end hosts and the network interact on the application layer over the normal IP network can be used to deploy and experiment with new features rapidly with low risk.

D. *Roaming*

Currently, data roaming is handled by a visited network forwarding APN-indicated tunnels to a home network. The current cellular architecture is in fact designed with roaming in its foundations, which may not be ideal for some network operators with low roaming activities. The current architecture generally assumes that all the traffic of a visiting roamer (except for circuit or packet voice) must go back to its home network[18]. This is done for accessing private operator-specific services and accurate accounting purposes when the fees for roaming traffic are very high and/or the cost of fraud is substantial. Due to the resulting inefficiency and complexity, such an assumption is not longer appropriate for many operators, especially those with small amount of roaming business and those experiencing strong pressure to lower roaming fees. In the future, roaming should be handled via local accounting, while all user traffic is routed locally and efficiently without tunneling back to home networks. This can be augmented with rapid accounting information exchange[19], device self-reporting, third-party audit, or micro-transactions directly between roamers and local visited networks. Thus, the basic architecture should reflect this trend and the old roaming model may be accommodated as special cases only for very high value traffic.

E. *What about other network services?*

The main function of a mobile network is to handle a large quantity of packet traffic efficiently and cheaply while user devices move across its coverage area. Other features such as Mobile Broadcast Multimedia Service [22] or layer 2 VPN for enterprise networks are important for some market segments, but should not dictate how the basic architecture is devised. Most of these services can be implemented at the IP layer, sometimes with some loss of efficiency from the perspective of those specific services. (For

---

[18] It is not uncommon to route a roaming user's traffic via series of tunnels from London through a roaming exchange network in Amsterdam, to his/her home network in the US, and then to a web server back in London over the Internet.

example, emulating layer 2 VPN over user IP layer over radio links is less efficient than forming layer 2 VPN over a transport network connected to a radio bearer.) However, that is not a reason enough to justify changing the basic architectural goal, which is to cover the majority of users and services efficiently. This principled approach should be maintained in the face of many competing interests from subsets of users and services.

VI. TRANSITION AND COEXISTENCE

Fortunately, there appear to be several voices [7][10][15] [16][18][19][20] in industry and academia that generally agree with this paper on the problems of current cellular architecture and the general directions for solutions. It is a valuable exercise to design a new green-field architecture unencumbered by existing networks and user devices such as the one described in the previous section. However, it is obviously necessary to examine how to transition from the current architecture to the new, and what kind of compromise is necessary. The major issue is how to support mobile devices that operate on the existing network during the transition.

A. *Layer 2 overlay in the RAN*

Once most base stations are connected by Ethernet technologies as the current trend continues, a new RAN transport network would be enabled on the Ethernet RAN as a separate Ethernet overlay[20]. The base station software would be updated so that traffic from new user devices would be put on the new Ethernet RAN overlay, while existing devices would continue to use the conventional RAN.

B. *Emulation of legacy tunnel protocol termination at base stations*

As more and more user devices are replaced or software-updated to use the new architecture, it becomes feasible to emulate the basic PDN-GW/GGSN functions in software at the base stations [19] or at the first RAN router. During this transition, a network-based IP mobility protocol, such as PMIP [17] can be used to handle IP layer mobility for these devices transparently.

---

[19] For example, www.RoamEX.com, a Near Real Time Roaming Data Exchange (NRTRDE).

VII. CONCLUSION

When the RAN and core network architecture for LTE was first discussed, there was a proposal [16] within 3GPP that shared some aspects of the architecture discussed in this paper. It was criticized as unrealistic and too simple, and dismissed by the majority of the 3GPP members. It did not help that this was before the recent remarkable success of Smartphones and the resulting increase in data traffic and active PDP sessions. In hindsight, it was indeed a golden opportunity to update the cellular network architecture, since there were both an acknowledgement that some aspects of the network architecture were no longer viable, and a willingness to adopt changes. Indeed, the LTE network architecture called Evolved Packet Core (EPC), does eliminate a few network elements, and simplifies some of the RAN architecture. Although, it was a change in the right direction, the result appears to provide somewhat constrained enhancements in terms of reduction in complexity and improvement in flexibility.

The future of the mobile cellular network is difficult to envision with specifics beyond a few general observations: There will be far more devices, orders of magnitude more base stations connecting them, and numerous different applications - ever changing - running over the network. The high-level architecture sketched in this paper can be the first step to meeting the challenges of this inevitability.

ACKNOWLEDGMENT

The authors would like to thank Thomas Anschutz, William Rembert, Peter Magil and Charles Kalmanek for their helpful discussions and comments.

REFERENCES

[1] IPv6 in 3GPP Evolved Packet System, http://tools.ietf.org/id/draft-ietf-v6ops-3gpp-eps-08.txt

[2] Neighbor Discovery protocol, http://tools.ietf.org/html/rfc4861

---

[20] TRILL also provides mechanisms to separate layer 2 network.

[3] Chapter 14, Tero Ojanpera, Ramjee Prasad, "WCDMA: Towards IP Mobility and Mobile Internet," Artech House (February 2001)

[4] Erik Dahlman, et al, "3G Evolution," Academic Press; 2 edition 2008

[5] Dohler, M. Heath, R.W. Lozano, A. Papadias, C.B. Valenzuela, R.A. "Is the PHY layer dead?, " IEEE Communications Magazine, April 2011, Volume: 49 Issue: 4, On page(s): 159 - 165

[6] A Survey of Mobility Support in the Internet, http://tools.ietf.org/html/rfc6301

[7] Joe Eggleston, et al., "3G Mobile/Packet Core Security and Engineering Challenges," http://www.nanog.org/meetings/nanog51/abstracts.php?pt=MTcwNSZuYW5vZzUx&nm=nanog51

[8] Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO), http://www.3gpp.org/ftp/Specs/html-info/23829.htm

[9] Jim Gettys, "Bufferbloat: Dark Buffers in the Internet," IEEE Internet Computing, vol. 15, no. 3, pp. 96, 95, May/June 2011, doi:10.1109/MIC.2011.56

[10] Problem statement for distributed and dynamic mobility management: http://www.ietf.org/id/draft-chan-distributed-mobility-ps-05.txt

[11] TRILL Workgroup, IETF: http://datatracker.ietf.org/wg/trill/charter/

[12] Shortest-Path Bridging, IEEE 802.1aq: http://www.ieee802.org/1/pages/802.1aq.html

[13] LISP Workgroup, IETF: https://datatracker.ietf.org/wg/lisp/charter/

[14] Henning Schulzrinne , Elin Wedlund, "Application Layer Mobility using SIP," Mobile Computing and Communications Review, 2000

[15] DMM (Distributed Mobility Management) working group, https://datatracker.ietf.org/wg/dmm/charter/

[16] New Architecture proposal for LTE, 2005, 3GPP, http://www.3gpp.org/ftp/tsg_sa/WG2_Arch/TSGS2_46_Athens/Docs/S2-051260.zip

[17] Proxy Mobile IPv6, http://tools.ietf.org/html/rfc5213

[18] P. Bosch, L. Samuel, S. Mullender, P. Polakos, G. Rittenhouse, G., “Flat Cellular (UMTS) Networks,” Wireless Communications and Networking Conference, 2007.WCNC 2007. IEEE

[19] M. Bauer, et. al., “The UMTS Base Station Router,” Bell Labs Technical Journal, Volume 11, Issue 4, MAR 2007

[20] Nuutti Varis, Jukka Manner, Johanna Heinonen, "A Layer-2 Approach for Mobility and Transport in the Mobile Backhaul". IEEE ITST 2011

[21] M. Sawahashi, et al., “Coordinated Multipoint Transmissions and Receptions for LTE-Advanced,” IEEE Wireless Communications, June 2010, Volume: 17 Issue: 3, page(s): 26 – 34

[22] Mobile Broadcast Multimedia Service, http://www.3gpp.org/ftp/Specs/html-info/22146.htm

[23] R. Doverspike and A. Gerber, “Traffic Types and Growth in Backbone Networks,” in Proc. of OFC/NFOEC, Invited Paper, 2011

[24] David D. Clark, et. al, “Tussle in Cyberspace: Defining Tomorrow’s Internet, ” IEEE/ACM Transactions on Networking (TON), Volume 13 Issue 3, June 2005

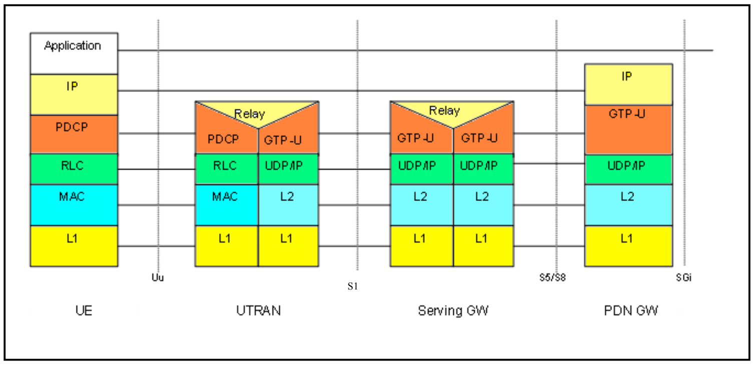
[25] SFF Extensions for Mobile IPv6: http://tools.ietf.org/html/draft-perkins-mext-sffexts-01

[26] Zhiyun Qian, Z. Morley Mao, ” Off-Path TCP Sequence Number Inference Attack: How Firewall Middleboxes Reduce Security,” IEEE Symposium on Privacy and Security, May 2012

[27] Z. Wang, et. al., "An untold story of middleboxes in cellular networks," Proceedings of the ACM SIGCOMM 2011, pp. 374~385

[28] Haiqing Jiang, Zeyu Liu, Yaogong Wang, Kyunghan Lee, Injong Rhee, "Understanding Bufferbloat in Cellular Networks," ACM SIGCOMM Workshop on Cellular Networks: Operations, Challenges and Future Design, Helsinki, Finland, Aug. 2012

[29] Serval Architecture: http://www.serval-arch.org



UE: User Equipment, UTRAN: Universal Terrestrial Radio Access Network

GTP-U: GPRS Tunneling Protocol-User plane

PDCP: Packet Data Convergence Protocol

RLC: Radio Link Control

Figure 1: A simplified user-plane protocol stack for LTE

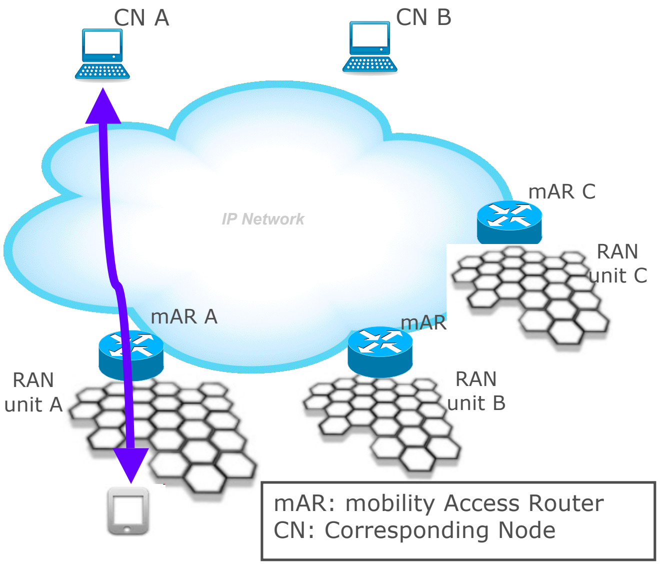


Figure 2-1

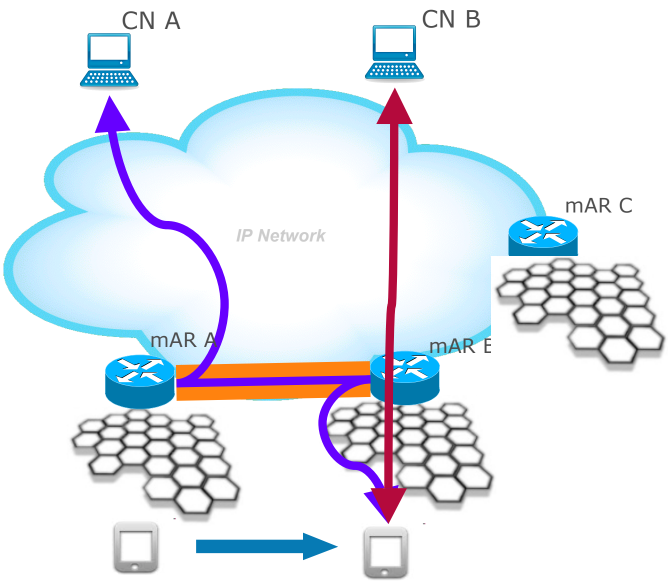


Figure 2-2: A new flow to CN B uses mAR B

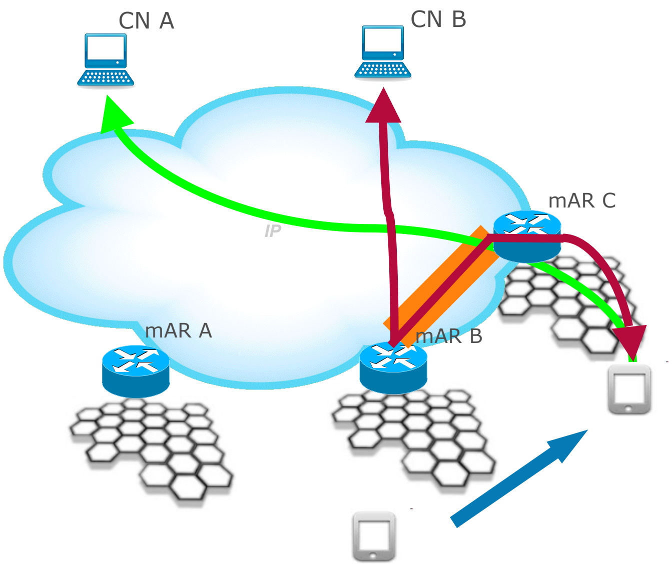


Figure 2-3: a new flow to CN A is routed without tunnel