Mitigating Strategies for Smart Phone Signaling Overload

Deepak Kataria, IP Junction Inc, www.ip-junction.com

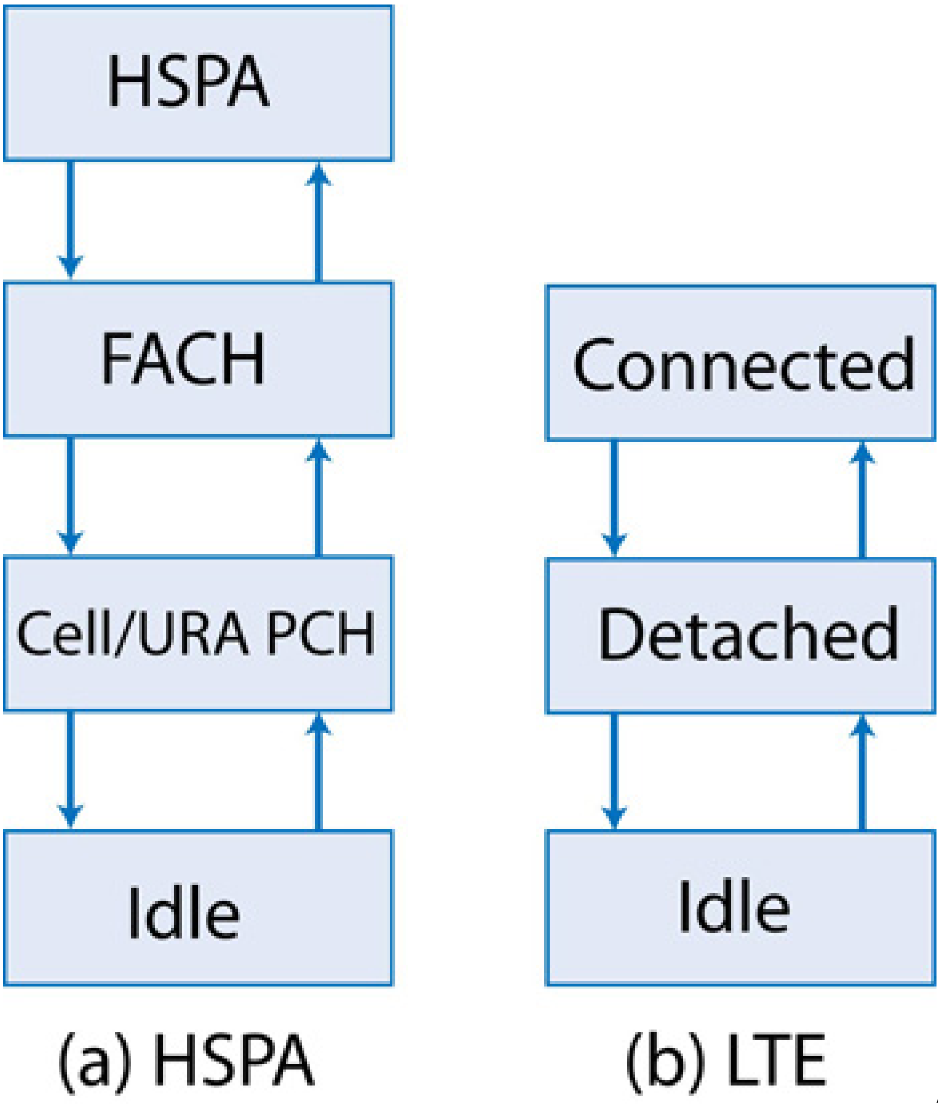
The rapid adoption of next generation mobile devices like the Apple iPhone, Google Android based smart phones as well as the increased traffic originating from these devices has taken network operators and equipment manufacturers by surprise. Mobile data traffic surpassed voice traffic at the end of 2009 and is currently growing at close to 100 percent year on year in developed countries. It is no longer about voice or SMS in a two-way communication; a single user with a single device communicates with multiplicity of applications over the mobile web. The user and application communication has escalated bandwidth demand while the device and application communication has increased signaling traffic. While everybody has had their eyes on coping with bandwidth demand, and data offload strategies have received extensive coverage, the dimension of increased signaling traffic has not received adequate treatment.

In mobile networks, signaling messages are exchanged between terminal and the network to set up and tear down connections. There are also messages exchanged to notify the network of terminal location and new calls as well as notifying terminals of network ability and new calls. These signaling messages use separate channels from those used by data traffic but affect the overall bandwidth and compute resources.

With smart phones, the situation is exacerbated. For instance, smart phones have an impressive repertoire of functions and features and support applications such as IM, push email, social networking. Such applications have the proclivity to maintain “heart-beat” connectivity with the network by sending messages as often as every 30 or 60 seconds. These messages cause terminal state changes for which signaling messages are exchanged with the network. Indeed, as many as 30 signaling messages are generated equivalent to what would be needed to set up a voice call. Imagine the impact, when such messages are exchanged every minute or so to maintain the refreshed state required by these applications. In such scenarios, it is the signaling traffic which dominates as opposed to the data traffic. From figures published by network operators, the profiled traffic showed on the average smart phones generated 8x the signaling traffic than a PC/laptop dongle. This is understandable, given the fact that with PC/laptop dongle usage, signaling traffic is only generated with each new page request, and further PC/laptops have bigger batteries allowing them to keep connections open for a longer duration.

Operator networks powered up by conventional base station and packet core network element configurations and designs are now recognized as increasingly unable to handle the traffic patterns that are emerging from these devices. The dimensioning of network resources was based on requirements for PC/laptop dongle users. In addition, the Long Term Evolution (LTE) of 3G as defined and standardized in 3GPP is emerging as a global standard for mobile broadband embraced by operators in North America, Europe and Japan. LTE promotes a flat architecture where base stations, or eNodeBs as they are called, are directly connected to a gateway node eliminating the need for a base station controller node. Further, a standalone Mobility Management Entity (MME) has been defined dedicated to handle signaling traffic in LTE networks. The absence of a controller node adds more processing burden on the eNodeBs and MMEs. This situation will be further amplified by the proliferation of Machine-to-Machine (M2M) devices which are connected through a web of WPANs, WLANs to cellular WAN networks across to various segments of the industry such as transport, energy, and heath care. These intelligent M2M devices provide a host of functions such as energy monitoring, home automation, security, and fleet management, and they support e-health and m-health functions. As new apps proliferate and leverage network connectivity and location aspects for the expanding m-life, signaling issues will need to be dealt with carefully.

Let us consider the case of M2M or mobile user device connected to the cellular WAN using an HSPA module.



An HSPA module may exist in one of the states as shown in Figure 1 (a). State 1 HSPA: In this state the mobile device is allocated a dedicated transport channel in the downlink and in the uplink along with a requisite number of physical channels, depending on the required bandwidth. When a mobile device is in Cell_DCH it is consuming the most network resources, both RNC processing and air interface resources, while the drain on the battery is also at its highest level.

State 2 FACH: In Cell_FACH (Cell Forward Access Channel) the mobile device is communicating with the network and the network (RNC) knows where the mobile device is located, thus the mobile device is consuming network resources – both in terms of air interface capacity as well as with respect to RNC processing power. In the current implementation of HSPA, small bits of data can be transmitted while in Cell_FACH state at a relatively low data rate, or up to 64 kbps in the downlink and 8-16 kbps in the uplink.

State 3 Cell/URA PCH: In URA the network (Radio Network Controller or RNC) knows that the mobile device exists, but other than having that basic knowledge the RNC/network does not dedicate any resources to the phone.

State 4 Idle: When in Idle mode the mobile device is basically dormant and not communicating with the network although it does listen for certain broadcast messages. In this state the device is not consuming any network resources, and it consumes the least amount of power.

Figure 1(b) shows the state diagram for an LTE mobile device.

All changes between states involve signaling in order to prepare the networks and device for the state change and then also perform the change. This signaling increases the signaling load in the network and has to be handled by appropriate dimensioning of the signaling capacity.

Thus, with respect to power usage, data throughput and use of maximum network resources, the drain on device battery is at its peak in the HSPA state. Since the device stays dormant in the Idle state and no data transmission is involved, the device battery is at its minimum usage.

Device manufacturers with a penchant for differentiation implemented features such as fast dormancy without realizing network consequences. With fast dormancy, the device signals the RNC for connection release as soon as the transaction is over (instead of waiting for timers to expire before going to lower states), so that the device can return to Idle state sooner and preserve battery life. While this extended battery life, what was lost on the device manufacturers was the fact that the device needed to traverse the higher states more often every time there was a request to connect to the network. This action resulted in increased signaling traffic between the network and devices. Soon after the negative impact on network was realized due to increased signaling traffic, fast dormancy as a feature was no longer supported.

A well thought out strategy working across networks, systems, and devices is needed to combat signaling overload while meeting the goals of network optimization, efficient use of system resources and increased battery life for the end devices.

At the network level, approaches such as 3G Direct Tunnel have been employed for 3G networks, where SGSN is bypassed for direct connectivity to GGSN for data traffic. This provides opportunities for scaling signaling capacity in SGSN to deal with smart phones. Further opportunities for scaling include pooling of SGSNs such that a shared compute farm with load balancing can be created, providing the needed scalability, agility and elastic computing resources for handling signaling messages. In LTE networks, MME is a dedicated entity for handling signaling traffic. In co-existent 3G/LTE networks, SGSN-MME pools can be organized to efficiently handle signaling traffic. I-HSPA is another approach, where the device connects directly to the Internet with only the base station in between. This approach works well where no mobility is expected but portable PC/laptop dongles can be used for Internet access, leaving more capacity for the mobility-prone smart phones to use extra capacity traversing the path through the RNC and SGSN. RNCs can be designed where the signaling capacity is available as a shared pool and can be flexibly used as opposed to being hard partitioned across various Node Bs.

Femto cells provide an attractive alternative to offload mobile core network from the impacts of increased signaling traffic by carrying signaling traffic over the wire line network for indoor environments.

Device management is another area of focus where the device is configured with appropriate settings to prevent unnecessary pinging from devices. For example, some social apps tend to poll the servers every 30 seconds or so to find new social data, contributing to a large number of signaling messages. Other apps such as those for patient monitoring and financial trading require periodic connectivity to the network. By optimizing protocol behaviors across the server and clients, the impact of signaling can be mitigated to some extent. For instance, instead of polling every 30 seconds for social updates, a push approach may be employed for every configurable threshold level of event updates. Further, by controlling the states of transition, the number of signaling messages can be reduced. For example, it has been shown that by keeping the device in Cell URA/PCH state, as opposed to going to Idle state, allows for a connection to be restored to active state with significantly fewer signaling messages. Further, coordination between device OS and radio layers can enable a handshake whereby the network is made aware of application characteristics and can use this information to select the most appropriate state. For instance, periodic unattended status updates can be treated differently relative to user involved updates. The network can move the former to lower states after the transaction while keeping the latter on higher states, knowing this to be an on-going session.

At the processor level, design approaches that can distinguish control plane traffic and user plane traffic and can scale to handle each type will be needed. Signaling messages (control plane) require a lower number of processing operations per event and have lower throughput; but they require high scalability for signaling process. User plane traffic, on the other hand, has very high throughput requirements and in many cases requires deep packet inspection to handle application flows with given policy and connection attributes. Therefore, the processors not only need to provide for fast path offload engines with high speed cores but also require higher compute performance per core. In other words, standard multicore processors or standard network processors will experience a signaling “bias” for the same number of subscribers and data traffic per subscriber if the capacity is not flexibly tunable in these two dimensions. A design approach that carefully calibrates performance to match the requirements of the traffic mix is needed. Higher signaling traffic leverages the capacity of higher performance compute engines, while the higher throughput data traffic leverages fast path processing engines to effectively deliver applications that require both multicore and high performance per core capabilities. These types of capabilities are usually required in network elements such as eNodeB, RNC, SGSN to support both user plane traffic and control plane for smart phones.