The ISP Column
A monthly column on things Internet

September 2015
Geoff Huston

The Changing Mobile World

Today's Internet is undoubtedly the mobile Internet. Sales of all other forms of personal computers are in decline and the market focus is now squarely on tablets, "smart" phones and wearable peripherals. In 2014 these providers sold 1.5 billion such devices into the global consumer market, and doubtless 2015's numbers will be greater. Half of all Internet-visible devices are now mobile devices and they generate 75% of all access provider revenues.

You might think that such significant volumes and major revenue streams would underpin a highly competitive and diverse industry base, but you'd be wrong. In 2014 84% of all of these mobile smart devices were using Google's Android platform, and a further 12% were using Apple's iOS system. The remaining 3% were mostly Lumia models using Windows Phone, and a light dusting of Blackberrys. Rather than a large pool of potential suppliers who are strongly motivated to compete with each other by building to the specifications provided by the cellular access providers, the concentration of supply has lead to a different dynamic where the platform providers are able to effectively dictate their terms and conditions to the mobile access network operators. Interestingly this, in turn, has emboldened some of the larger application providers to flex their market muscles and in turn set forth in their own direction.

As usual, it's scarcity that is driving much of these changes, but in this particular case it's not the scarcity of IPv4 addresses. It's access to useable radio spectrum.

There are two types of radio spectrum. The first is the traditional medium of exclusive use spectrum licenses, where the mobile network operator pays the government a license fee for exclusive access to a certain spectrum allocation in a particular geographic locale. The model of distribution of this spectrum has shifted from an administrative allocation model to one of open auctions, and the auction price of these licenses has, from time to time, reflected an irrational rush of blood to the collective heads of these network operators. The high cost of spectrum access implies that the network operator starts with a non-trivial cost element for exclusive access to the spectrum, and on top of this the operator must also invest in physical plant and business management operations. Typically, the spectrum actions are constructed in a manner that there is no ability for a single operator to obtain a monopoly position, and most regimes ensure that there are between two and four distinct spectrum holders in the most highly populated locales. Sometimes this level of competition is enough to maintain an efficient market without price distortions, while at other times the small number of competitors and the barriers to entry by any new competitors leads to various forms of cartel-like behaviours with outcomes of price setting distortions in the mobile market, where the retail price of the service has no direct relationship to underlying costs.

But competitive pressure from other mobile network operators is not the only source of competition. The other form of spectrum use is also a factor. WiFi systems use two unmanaged shared spectrum bands, one at 2.4.Ghz and the second at 5Ghz. There are typically limits on the maximum transmission power used by devices that operate in these bands, but to all other extents the spectrum is effectively open for access. For many years WiFi had been used to support domestic and corporate access. WiFi systems typically operate within a range of up to 70m indoors and 250m outdoors. This small radius of WiFi systems, which is an outcome of the limited transmission power, has fortuitously also allowed these system to operate with extremely high capacity. They run at speeds that range from 10 to 50 Mbps (the 802.11b specification) through to speeds of up to 1.3Gbps (the 802.11ac specification), and expectation is that this can be lifted to even higher speeds in the near future.

For a while these two spectrum use models compartmentalised themselves into distinct markets. Exclusive-use spectrum was for the 'traditional' mobile network operators and shared spectrum use was for self-installed domestic and office applications. But the mobile devices did not make such a distinction. They include radio interfaces for both cellular data across licensed spectrum, and WiFi data across shared spectrum. So while the access suppliers segmented themselves into cellular or WiFi operators or various forms, the devices are able to straddle both environments.

We are now seeing some operators of wired network access infrastructure entering into what looks like head to head competition with the incumbent mobile operators in the provision of mobile services using WiFi as the platform. Comcast's Xfinity service in the US is a good example of this approach, and it boasts of millions of WiFi hot spots which are usable by existing Comcast customers without any additional cost. Perhaps surprisingly, some of the incumbent mobile operators have followed this lead and are also offering free WiFi access services to their customer base. For example AT&T has a WiFi offering in the US. Part of the rationale for this may well be protecting their market share. However there is also the ever-present issue of congestion in the licensed radio spectrum space. 3G and 4G data services are opportunistic, and scavenge otherwise uncommitted access capacity. However, in places of intense use, such as high density urban centres, the challenge of providing high capacity data services in licensed spectrum becomes an extremely tough challenge. One approach is to use WiFi access points as a relief mechanism for these areas of otherwise high congestion. The observation that this move on the part of the traditional mobile operators to use unlicensed spectrum as a pressure relief mechanism for over-committed exclusive use spectrum clashes with the business objectives of the mobile operator's WiFi-based competition is perhaps a convenient bonus in such a scenario!

But there are some issues with WiFi that are not as obvious in the traditional mobile service space. The mobile industry has supported base-station handover since its inception, so that an active data stream to a device can be supported even as this device is in motion, being handed off from one radio access point to another. WiFi handoff is not a as cleanly supported. The issues surface when the WiFi access points reside in different IP networks, so that a handoff from one WiFi access point to another implies a change of the device's IP address. Conventionally, a change of IP address equates to a disruptive change to all of the device's active connections.

If you were able to manage the collection of cellular base stations and the WiFi access points in a single managed domain then it might be possible for the network itself to perform the necessary feats of mobility support. An approach could potentially be borrowed from the 3G environment and connect the mobile device through a persistent PPP tunnel whose endpoint could be migrated across cellular and WiFi access points as the device itself moved. Of course all of these WiFi access points would need to participate in the PPP signalling environment, and that means using customised functionality in the WiFi access points. The handset interface drivers would also need to change some of their behaviours to allow the handset to switch its local IP protocol stack between the cell and WiFi interfaces. Part of the attraction of WiFi is its existing deployment base and requiring additional functions in the base station requires new deployments. Handsets would also need to use modified drivers, so that the existing set of handsets could not make use of such a seamless network switch. The attractions of using WiFi as a seamless adjunct to cellular data quickly fades in the front of such challenges.

So it's not surprising to observe other approaches to WiFi handoff being explored by this industry. Lets look at a few here.

The "traditional" model of the internal architecture of the mobile device is no different to any conventional computer architecture, as shown in Figure 1. The operating system on the mobile platform supports a number of devices and access to the associated access network services.

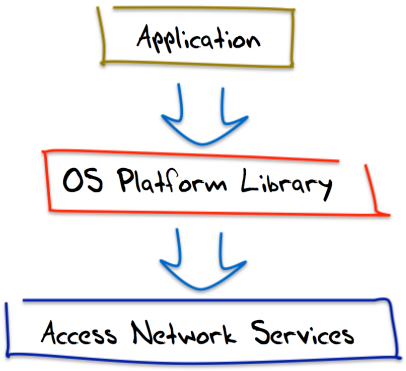

Figure 1 – The Original Mobile Platform Architecture

The application sits on top of the operating system, and interacts with it by performing a number of quite conventional procedure calls (such as the Unix socket abstraction, that allows the application to view a network connection as a serial I/O device, for example). The OS platform also includes a number of device drivers for the onboard communications ports, including cellular data and WiFi for example.

When a device is within range of both cellular data services and WiFi services what happens? Both Android and iOS use a local preference setting, and they prefer WiFi over cellular data. But its not a complete switch. Connections that were active across the cellular interface will not switch over to use a WiFi interface. The device would be changing its IP address in such a switch and the remote end cannot track such a change in an active session. So when a mobile device associates with a new WiFi access point any existing cellular connections will try to remain up using the cellular interface, while all new connections will open across the WiFi interface.

This is not altogether satisfactory if you want to perform a seamless handoff from cellular to WiFi. It's possible to construct a non-disruptive form of switching, and Google's recently announced Google Fi service is a good example.

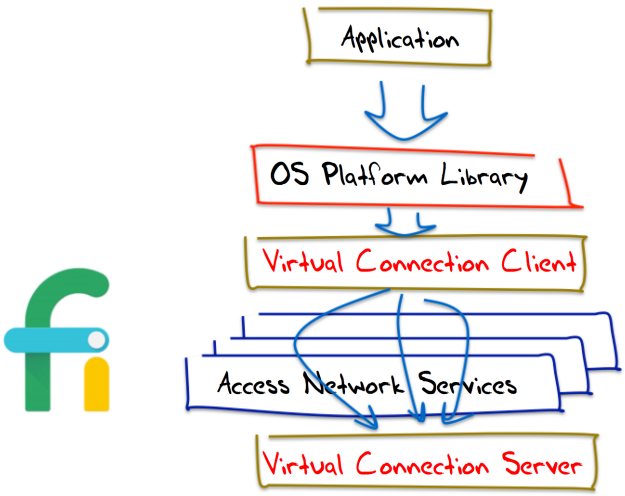

Figure 2 – Google's Fi Service Architecture

In this approach Google is a Mobile Virtual Network Operator (MVNO), and it uses the cellular data infrastructure of two mobile access networks. However, was well as roaming between these two cellular networks, a Google Fi handset will opportunistically shift to an accessible open WiFi whenever it can without disrupting any active sessions on the handset. It is likely that in constructing this service Google are using a relatively conventional Virtual Private Network tunnel between a nearby data centre and the handset, similar in some ways to 3G's use of a PPP tunnel to support switching between radio access points. A switch between access networks is implemented as a change to the outer IP wrapper of the VPN tunnel, and the interior IP connections is unaltered across any such switch. In this way applications use the same interface to the Android platform without alteration, and the platform itself controls which access networks it uses at any time.

Google say that the data is encrypted when using WiFi access. It is no doubt possible to encrypt the VPN tunnel in all cases, although there may be some regulatory constraints in a MVNO encrypting radio traffic that may prevent this encryption happening all the time. In essence the OS platform if performing a "hop-over" of the access networks, and treating them as substitutable commodity providers. Control of which access network is used passes from the network operator to the platform provider.

Apple has been experimenting with a somewhat different approach to handover. The current version of iOS includes support for MultiPath TCP in its TCP library. Apple's Siri application has made use of this form of connection.

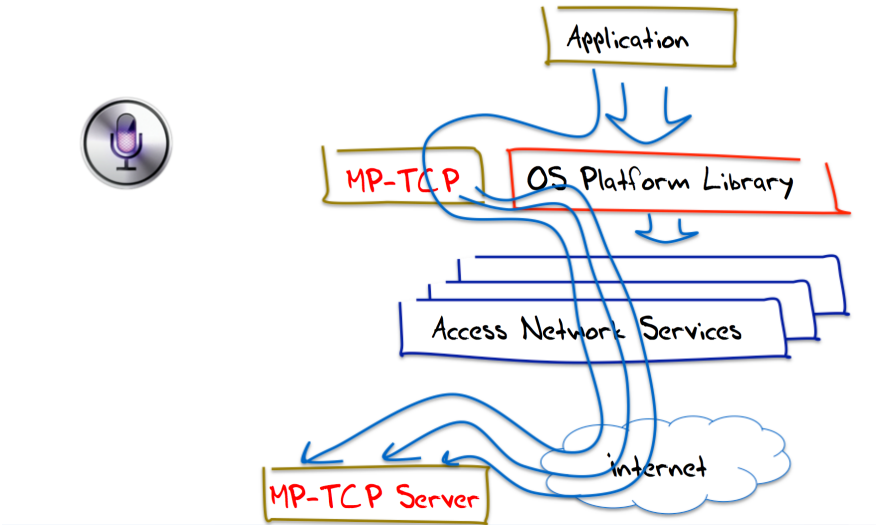

Figure 3 – Apple's Multi-Path Service Architecture used with Siri

MultiPath TCP does not necessarily switch between access networks, but allows the TCP session in the device to make use of all available access networks simultaneously. The TCP session is multiplexed across a number of access networks, and fragments of the TCP session are passed across each access network in what looks like a distinct TCP session per access interface. The remote end performs the re-stitching of the fragments into a coherent single TCP stream again.

This is provided by the iOS platform and is an option for applications who are able to negotiate a MultiPath session with a MultiPath enabled server. At this point the defined interface is relatively simple,

allowing an application to use the cellular and WiFi interfaces simultaneously, but its not a big stretch to add additional signalling to the application interface, allowing the application full control over which active interfaces are used by the application at any time.

This model of allowing the application to exercise its own decisions as to how it uses access networks is also evident in the architecture of the Facebook application (Figure 4).

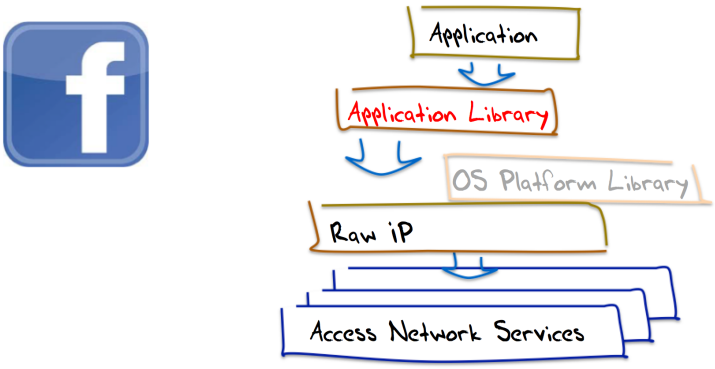

Figure 4 – Facebook's Application Service Architecture

The application does not use the connection libraries provided by the host platform, but uses a TCP connection protocol that is bundled into the application. Not only does this provide Facebook with control over how its application behaves when talking to Facebook servers, it has the potential to hide the application's behaviour from the OS platform.

This is not a novel concept by any means. Browsers have used libraries to undertake DNS name resolution for many years in order to improve upon the "click-to-load" times for browsers. Some applications have gone further and direct DNS queries to their own chosen DNS resolver, eschewing the use of the "default" resolver provided by the access network. Google's QUIC is another example of this approach of using the application, where the Chrome browser was configured to use a somewhat different reliable flow control data protocol, layered on top of UDP, in an effort that allowed the browser application to take control of the data flow and hide it from TCP-intercepting middleware.

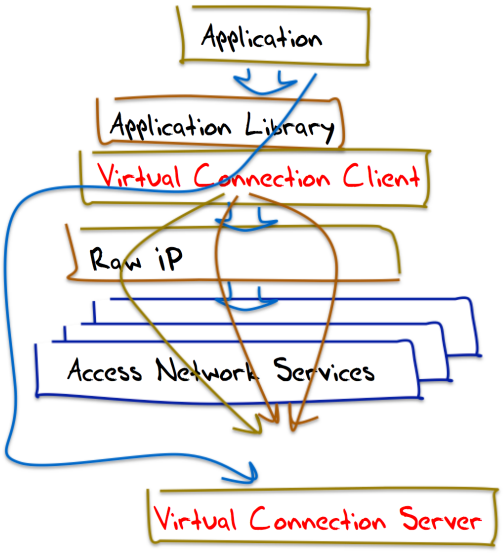

Figure 5 – The Paranoid Application Service Architecture

It's evident that applications are attempting to exert ever greater levels of control over their connections, and also becoming a little more paranoid in terms of what other parts of the environment they are prepared to trust.

What we are seeing is that the mobile device is no longer exclusively tethered to a mobile network operator, and the device is able to react opportunistically to use the "best" network, whether it's the greatest available capacity or the lowest incremental cost to the consumer. From the device's perspective the mobile network is just one possible supplier of transmission services, and other options, including WiFi, Bluetooth and USB ports can also be used, and the device is able to make independent choices based on its own preferences.

This has profound implications. While the device was locked into the mobile network, the mobile network could position itself as an expensive premium service, with attendant high prices and high revenue margins. The only form of competition in this model was that provided by similarly positioned mobile service operators. The limited number of spectrum licenses often mean that the players established informal cartels and prices remained high. Once the device itself is able to access other access services, then the mobile data network operators find it hard to maintain a price premium for their service. The result is that mobile service sector is being inexorably pushed into a raw commodity service model. The premium product of mobile voice is now just another undistinguished digital data stream, and the margins for mobile network operators are under constant erosive pressure. The unlicensed spectrum open WiFi operators are able to exert significant levels of commercial pressure on the mobile incumbents in the mobile service environment. This means that the prices paid for exclusive use spectrum licenses are exerting margin pressures on operators whose revenues are increasingly coming from commodity utility data services.

So who is winning here?

The cellular data access providers appear to be losing their historical control of the mobile world, so I can't really see that they are winning here.

The OS platform providers are trying to assert a greater role in terms of deciding how mobile devices communicate, and Google's Fi is a good example of this effort. But neither Apple nor Google are having a clear run, so its not clear that they will gain the ascendency here.

The applications themselves are now also changing, wanting to both achieve a higher level of direct control over their communications and assert a higher level of secrecy about the extent to which application information is visible to both other applications and to the OS platform and the access provider. It's likely that this will continue, and each application may be further motivated to look after its own DNS requirements, its own session transport control, its own security framework and its own way in which it make use of the underlying connectivity services.

Whichever way this plays out in the coming months it will drive further changes in the mobile world and that necessarily implies further changes in the Internet itself.

Author

Geoff Huston B.Sc., M.Sc., is the Chief Scientist at APNIC, the Regional Internet Registry serving the Asia Pacific region. He has been closely involved with the development of the Internet for many years, particularly within Australia, where he was responsible for building the Internet within the Australian academic and research sector in the early 1990's. He is author of a number of Internet-related books, and was a member of the Internet Architecture Board from 1999 until 2005, and served on the Board of Trustees of the Internet Society from 1992 until 2001. He has worked as a an Internet researcher, as a ISP systems architect and a network operator at various times.

www.potaroo.net

Disclaimer