White Paper

Internet Offload for Mobile Operators

Prepared by

Gabriel Brown
Senior Analyst, *Heavy Reading*
www.heavyreading.com

on behalf of

radisys

www.radisys.com

and

STOKE

www.stoke.com

September 2011

Introduction: What Is Internet Offload?

Mobile networks are now dominated by data. Smartphones are driving revenue growth in virtually all advanced markets, and they are now at the heart of operators' network evolution strategies. It's a glorious thing – but is it a victim of its own success?

The idea that the classic 2G/3G mobile network architecture cannot cope with the increase in data and signaling traffic has gained currency, and is driving the emerging concept of the "Internet offload." The argument is that if data and signaling traffic could be offloaded to a lower-cost network, operator economics would improve and services could be provided at lower cost. This is turn would drive data growth.

Taken to its logical conclusion, however, the argument is that cellular technology *per se* is too expensive to support true broadband. This is something of an extreme view that is championed outside the industry mainstream by individuals and organizations that are intent on disrupting the value chain. *Heavy Reading* does not believe it is useful to frame the discussion in such terms; we prefer a more measured analysis.

Yet there is an element of truth in the argument: Mobile broadband is in a transitional phase, usage is growing, and demand for high-performance, low-cost services is economically challenging. The "classic" network architecture was designed in another era and must evolve. As such, taking "Internet offload" ideas and integrating them into the mainstream offers potentially transformative opportunities for mobile broadband network design.

Internet Offload Is Services-Led

The very name "offload" has negative connotations. In order to be truly integrated with the mainstream of mobile network technology, Internet offload cannot only be about cost of production; it must also be about improved service capability. Therefore, it is useful to recast "Internet offload" as a services-led initiative, which can be characterized as follows:

- Internet offload is fundamentally a response to growth in mobile data and evolving smartphone usage patterns that are characterized by frequent, short sessions, low latency and high throughput.
- Internet offload, if done correctly, can enable new and better end-user services, such as better indoor coverage with femtocells, enhanced performance for rich media services over WiFi, or the ability to support larger monthly data quotas at lower prices using "offload gateways."
- Innovation in Internet offload is coming from the "edge," but over time it will merge with the industry mainstream, as its value and utility is more clearly demonstrated.

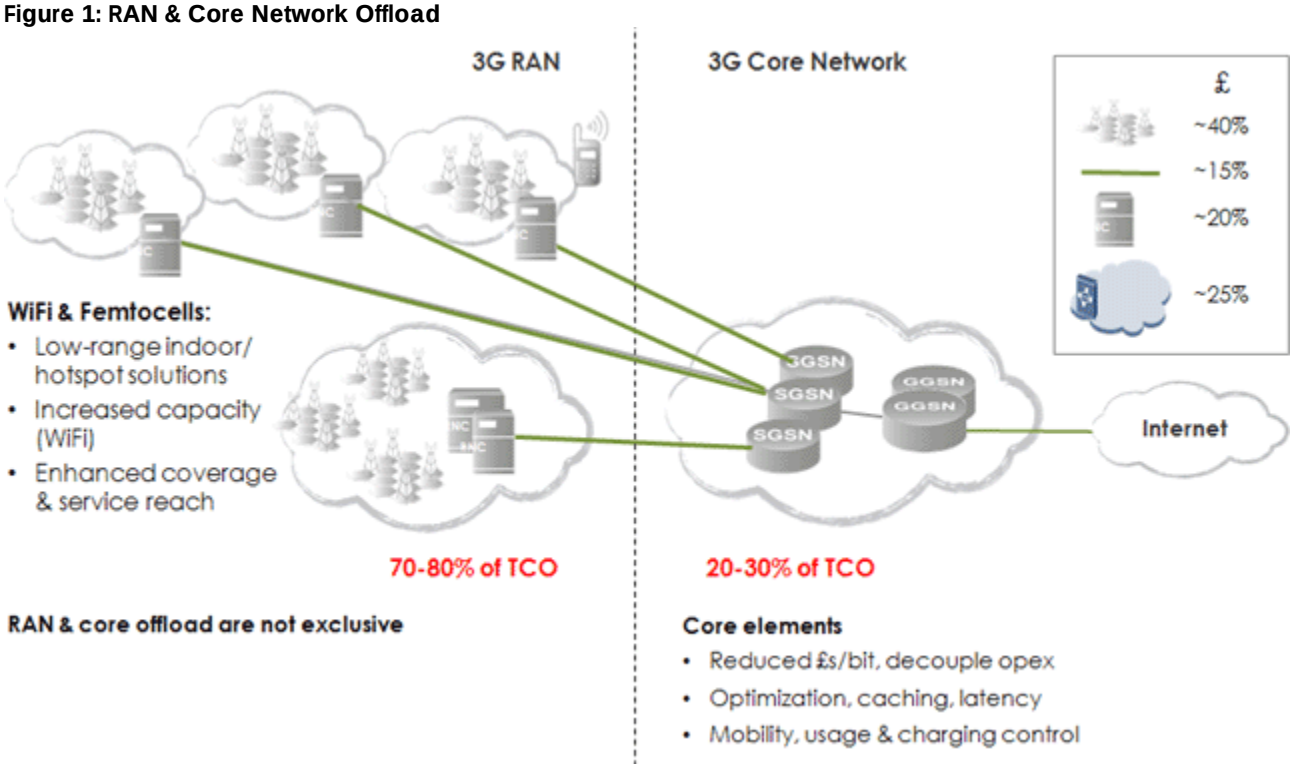
Innovation in networking has often started out outside the industry mainstream, only to become essential in later years: Ethernet, IP and arguably WiFi are all examples of this. In this vein, the concepts encompassed within the Internet offload should not be dismissed as uninteresting simply because they are not well supported by the classic network architecture today. It is logical that new services will require new approaches to network design, and *vice versa*.

Types of Offload

Broadly speaking, we identify two types of offload for mobile operators: radio access network (RAN) offload and core network offload. These are distinct in that they target different parts of the network and impact users differently. They are not mutually exclusive, however, and there are some important links between them.

Figure 1 shows a high-level estimate of the total cost of ownership (TCO) allocation in mobile networks. RAN is obviously the larger part of this, since it incorporates the network coverage operators need to offer a service.

Figure 1: RAN & Core Network Offload



Source: Stoke

RAN offload cannot realistically aim to replace this spending, since investment in coverage is fundamental to an operator's success, but through the use of WiFi and femtocells it can moderate the rate of investment, provide a location-specific solution to capacity thresholds, drive coverage deeper indoors and provide residential service where macro networks are currently uneconomic.

The core network accounts for relatively less TCO, but is still a substantial area of investment and is growing fairly rapidly as mobile data takes hold. Again, there is no suggestion that this investment is unnecessary, simply that offload techniques can be used to ensure core network equipment is used more productively. For example, if most of the traffic an operator is delivering is destined for the best-effort Internet and will not generate value-added revenue or attractive margins, why pay to process it through expensive core network platforms? By offloading traffic to the Internet operators can potentially reduce core network TCO.

3GPP Offload Initiatives

The mobile industry has, in fact, a long history of embracing technologies and approaches to network integration that could be categorized as "offload." Through the 3GPP standards process, work has been done to take offload concepts and formalize them for use in the mobile operator environment. Some examples are shown in **Figure 2**.

Figure 2: Examples of "Offload" Specification Work in 3GPP

INITIATIVE	3GPP TS	RELEASE	DESCRIPTION
I-WLAN	TS 22.234	Rel-6	Integrate WiFi Access to mobile packet core
UMA/GAN	TS 43.318	Rel-6	Run CS and PS services over WiFi Access
Direct Tunnel	TR 23.919	Rel-7	SGSN bypass for user-plane traffic
ANDSF	TS 24.312	Rel-8	Core network function to ensure device selects best available WiFi
Femtocell	TS 22.220	Rel-9	Residential small cell operating in licensed spectrum
LIPA & SIPTO	TR 23.829	Rel-10	IP offload specifications at BTS and the packet core
IFOM	TS 23.261	Rel-10	Ability to split flows across access networks (e.g. WiFi & cellular) on the same device

Source: Heavy Reading

3GPP Release 6, for example, included two approaches to integrating WiFi technology into the carrier network. At the time the WiFi attach-rate into smartphones was still low (and even in laptops, not universal), so in this sense both WLAN Interworking (I-WLAN) and UMA/GAN were ahead of their time. Both have provided useful pointers, however. The I-WLAN specifications for example, remain at the heart of a renewed push toward SIM-based authentication in service provider WiFi networks.

Direct Tunnel (Release 7) is limited in scope but has also been widely adopted across the industry and was the first move toward core network offload. As such, it opened the door for some newer, more ambitious techniques that have followed, such as Iu-PS offload and Selected IP Traffic Offload (SIPTO).

Femtocell is a quirk in that, like UMA/GAN, the technology was commercialized ahead of industry standardization (although in both cases industry forums hammered out the details in advance of submission to 3GPP). Now, however, work on Home NodeB and Home eNodeB is very much ongoing within 3GPP.

More recently, work in 3GPP Release 10 on Local IP Access (LIPA), SIPTO and IP Flow Mobility (IFOM) is very much a formalized response to the issue of Internet offload, and an explicit recognition that it has a role to play in future mobile broadband networks.

WiFi Offload

In a cellular networking context, WiFi is the comeback story of the decade. From a position in 2005/2006 where operators lobbied handset makers not to include WiFi or cripple it in firmware, because they were worried it would cannibalize nascent 3G services, the industry has now embraced WiFi. Virtually every smartphone sold includes WiFi, and operators actively encourage their subscribers to use it.

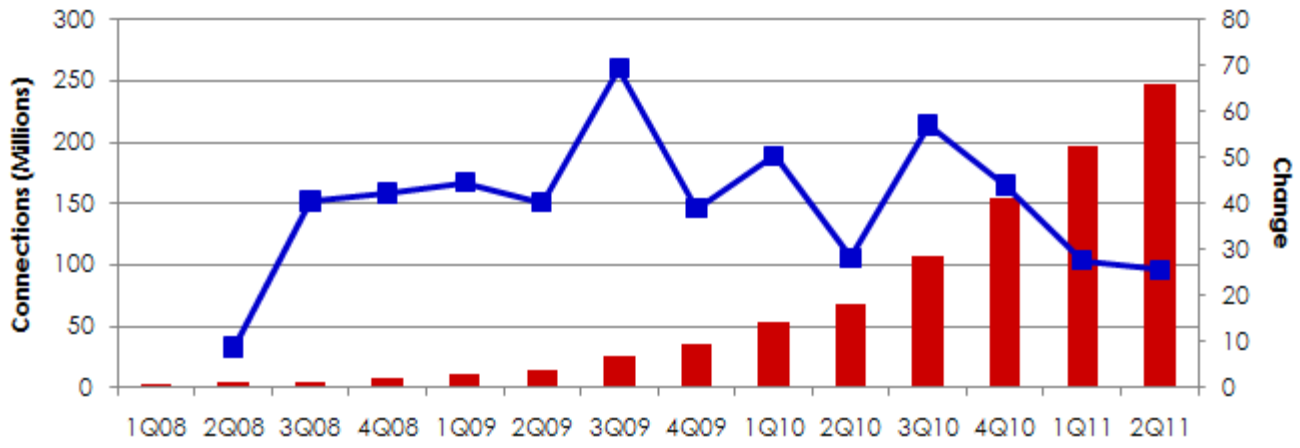
Increased use of WiFi has been user-driven, because WiFi provides a better data experience than 3G in certain types of indoor locations, such as homes and offices. The iPhone, in particular, broke the back of carrier resistance to WiFi and there have since been a number of other proof-points that underline the renaissance of mobile operator WiFi:

- AT&T's acquisition of hotspot operator Wayport in 2008 for $275 million. This underpinned a strategic expansion of AT&T's public-access WiFi footprint.
- Qualcomm's acquisition of Atheros for $3 billion in 2011. Perhaps more than anything, this move by the chipset market leader confirmed that every smartphone would have WiFi.
- Embrace of SIM-based authentication by operators worldwide. T-Mobile, Orange, SK Telecom, Korea Telecom, PCCW, China Mobile, and China Telecom all either offer SIM-based WiFi access or have plans to do so.

WiFi Usage Is Booming

Evidence that WiFi usage by smartphone users is booming is easy to find. In South Korea, for example, an academic study has found that about half of all smartphone usage (measured by time and data load) is via WiFi, while in the U.K. statistics from Bango, a mobile content analytics firm, show that by the start of 2011 about half of smartphone Web sessions occurred over WiFi. There is also strong growth at WiFi hotspots, as shown by the AT&T data in **Figure 3** (more than 70 percent of connections are smartphone-generated).

Figure 3: Connections to AT&T WiFi Hotspots by Smartphone Users



Source: AT&T

Integration With Operator Service Portfolios

The question for mobile operators now is how do they want to play WiFi? Should they continue to allow the user driven-model to prevail, since they already benefit from *de facto* offload and may have secured the bulk of the gains without having to do very much? Or should they take a more active role in offering managed offload solutions where they have more control over how and when traffic is offloaded, perhaps in return for greater ease-of-use and better overall service?

While not a near-term priority there is also the strategic question of how can an operator run a viable value-added services (VAS) business when half of their subscriber's usage is, essentially, off-network? Typically an operator's VAS strategy would seek to leverage the unique attributes of the network, so losing customers to WiFi could be problematic.

For reasons related to a desire to provide better service and retain usage in its own infrastructure, operators are looking toward tighter integration of WiFi with cellular services – which might be termed "managed WiFi offload" or "service provider WiFi." We expect this to play out over a number of phases and over several years. The three major phases, shown in **Figure 4**, are defined as follows:

- **Hard Offload.** This is essentially a user-driven phenomenon, where devices are manually configured to work with private WiFi. Public WiFi is bundled with smartphone subscriptions and perhaps will have some form of app or automated login support; however, this will typically be non-secure from a user-plane perspective.
- **Optimized Offload.** This phase will bring SIM-based authentication and secure public WiFi connections using 802.1X. The Next-Gen Hotspot initiative will provide for better service discovery and WiFi roaming. Public/private WiFi hotspots could also take off (a.k.a. the "Fon model").
- **Integrated WiFi.** WiFi access will be fully part of the mobile network and integrated into the core. It will support session mobility and be transparent to end users, with most services available and functional over either cellular or WiFi access.

Figure 4: WiFi Network Integration

Source: Heavy Reading

Core Network Integration

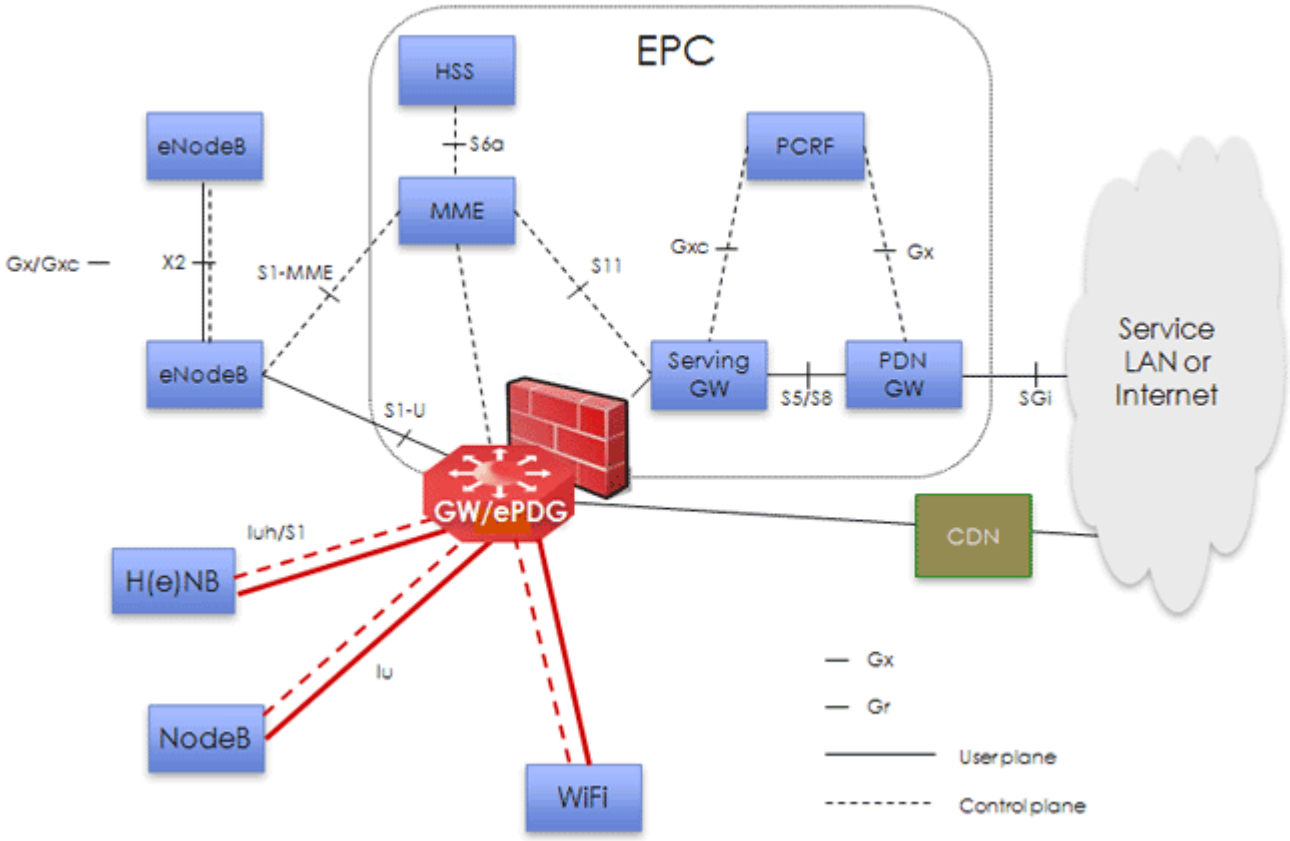
Until now mobile operators have been reluctant to pursue tight integration of WiFi with the core network. This has been a logical strategy so far, but without core network integration, WiFi will remain a user-driven, *ad hoc* type service.

Evolved Packet Core (EPC) could prove a turning point. Because EPC is designed to be access-independent, it can support common service delivery and session mobility across 3G/4G and WiFi. This will enable more sophisticated traffic management, managed offload techniques, and policy-based use cases – for example, if the device and network can determine the best form of connectivity according to the particular application or the prevailing load on each type of network, the user experience and network efficiency will be improved.

EPC integration of WiFi access could look somewhat similar to LTE small cells. In both cases untrusted IP backhaul is likely to be used, which implies the need for IPsec termination. Also like LTE small cells, operators will need a way to aggregate the high number of incoming tunnels/connections and therefore, an aggregation device installed at the edge of the EPC, perhaps running on an EPC hardware platform, is likely to be needed. The approach is shown in **Figure 5**.

It is probably one or two years from being implemented at even the more progressive carriers; however this looks today to be the best way for operators to integrate WiFi with their networks. It will make provision of a common service portfolio much easier and will allow operators to make better use of technologies such as ANDSF and policy to determine when and where a device should connect to WiFi.

Figure 5: EPC-Based Core Network Integration



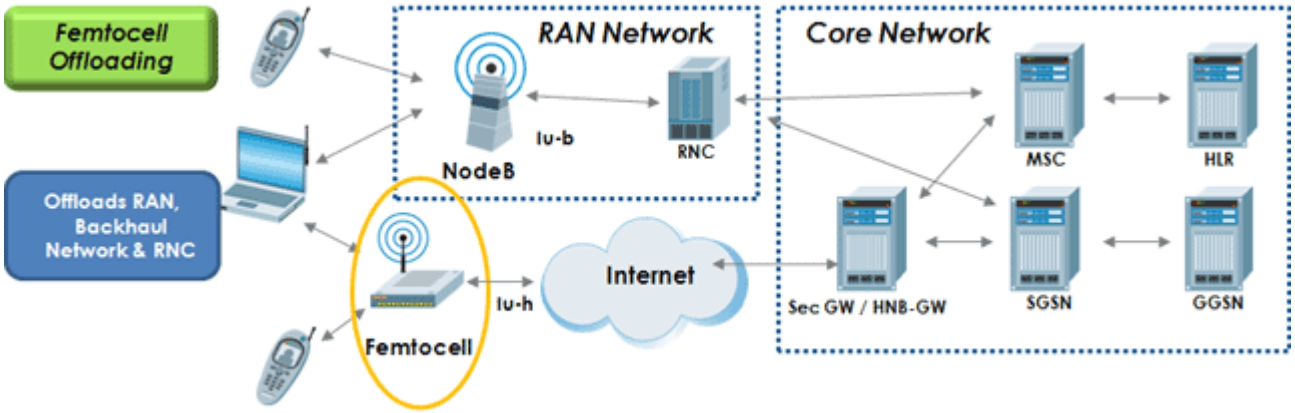
Source: Stoke

Femtocells

Femtocells, also known as Home NodeBs, are low power, mostly unmanaged small cells that are deployed in homes and offices to provide better coverage and capacity. Typically they are 3G devices today, although LTE femtocells, known as Home eNodeBs, are also under development. They can be thought of as a form of RAN offload in that subscribers may use them as an alternative to the macro radio network because the signal is stronger at their point of use.

The two major differences with WiFi are that they are deployed in licensed spectrum (e.g., the 2.1GHz 3G band) and they are fully integrated with the carrier network. Both points are important, but perhaps surprisingly, it is the core network integration that makes the real difference because it means femtocells are transparent to all operator services and are simply an extension of the network. The femtocell network architecture is shown in **Figure 6**.

Figure 6: Femtocell Network Architecture



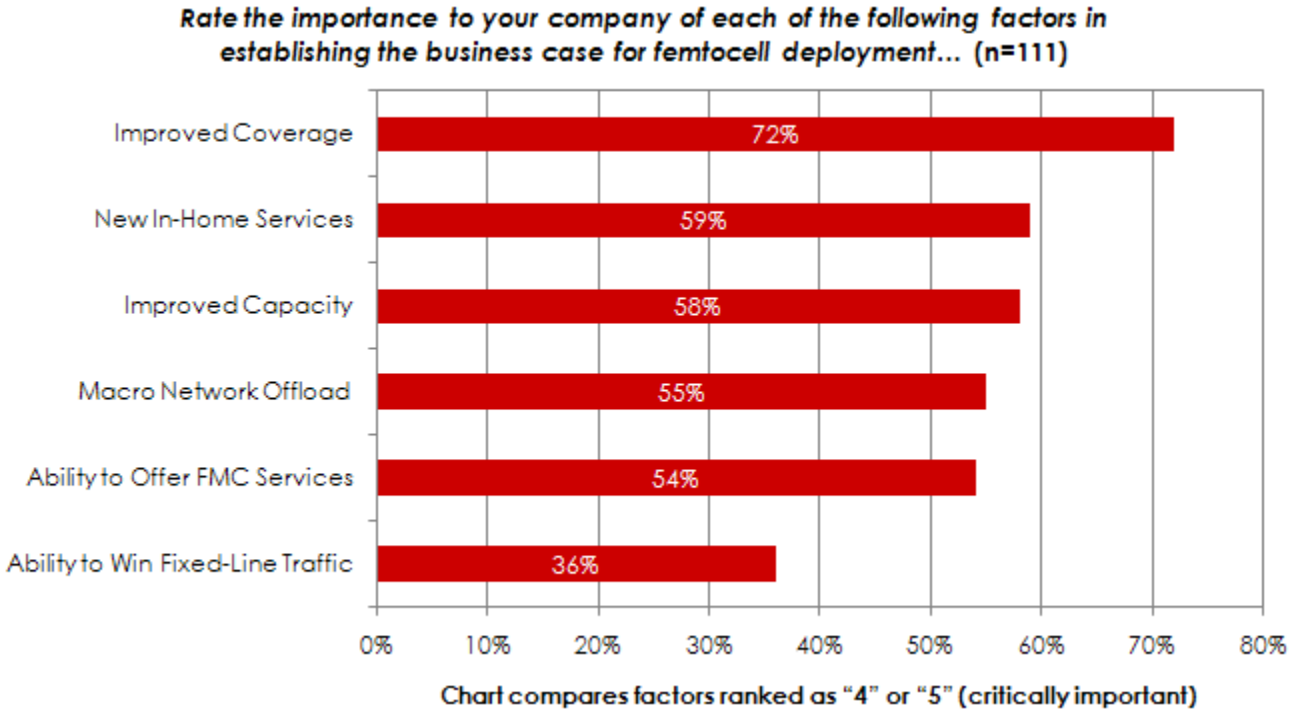
Source: Radisys

It is debatable to what extent femtocell should be considered an "offload" technology. There a number of complex variables related to spectrum bands and installation densities that determine how much of an efficiency gain femtocells really produce. Some of the technical analysis is disputed and there are entrenched positions on both sides

This is in some ways beside the point, however. Femtocells are typically deployed as a way to improve coverage, rather than as an offload solution, and the fact that they provide full voice service makes them better to suited to this role than WiFi. Services such as Vodafone's SureSignal are marketed to end-users as coverage extension, and that is generally how operators view the technology.

In a survey of over 100 mobile operator professionals *Heavy Reading* carried out back in 2009, before femtocell services had been commercialized, it was already clear that coverage is more important than offload in a femtocell solution. **Figure 7** is an excerpt from that study. It shows that coverage was ranked by 72 percent of respondents as a critical business case driver, versus 55 percent for macro-network offload.

Figure 7: Coverage Extension is the Driving Force for Femtocells



Source: Heavy Reading

Core Network Offload

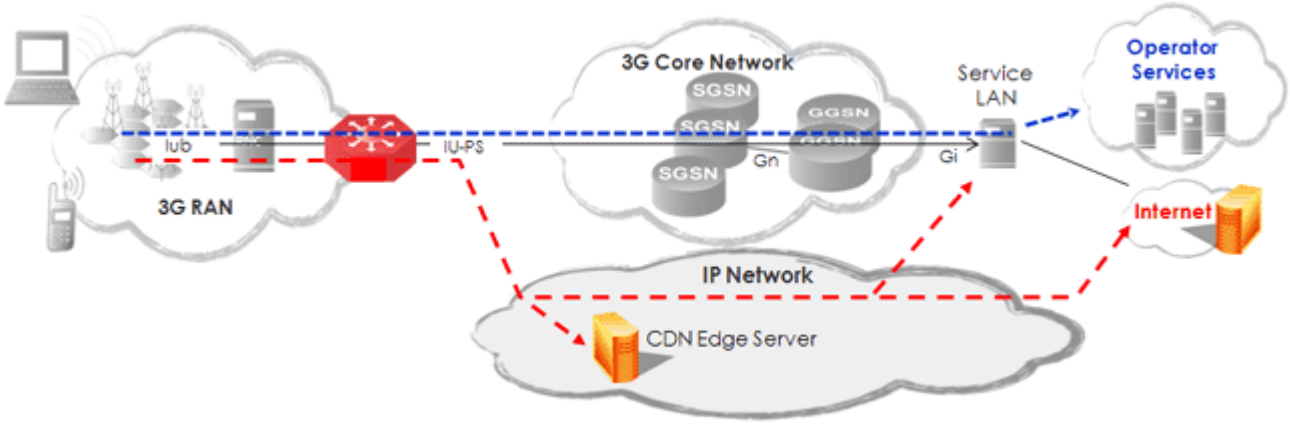
Core network offload, also sometimes called "Internet breakout," refers to the idea that because a majority of 3G traffic is destined for the best-effort Internet, it would be less expensive to break out this traffic from the mobile network and offload it to the Internet as soon as possible. The operator would not have to pay to process this traffic through the mobile packet core, and this would moderate investment even where traffic increases very rapidly. A number of core network offload options are discussed in the LIPA/SIPTO work in 3GPP, Release 10 (see document TR 23.829).

Packet-Switch Core Offload

Packet-switch (PS) core offload involves deployment of Internet offload gateways behind an RNC, or group of RNCs, to split out traffic bound for the Internet (the majority of traffic by volume) from traffic bound for the operator's core network (which includes signaling). This is also known as "Iu-PS offload."

It is argued that this will save operators money in two ways: by reducing the incremental investment in GGSNs and SGSNs needed to support broadband traffic growth; and the potential to use lower cost transport and distributed Internet peering to reduce costs. The concept is illustrated in **Figure 8**.

Figure 8: Offload Gateway at RNC Site



Source: Stoke

The success of this concept largely rests on the belief that GGSN and SGSN equipment was designed before the mobile broadband era really began and is optimized for control-plane functions related to mobility management, authentication, and billing, rather than high throughput. There is some foundation to this view – historically it is the case that some products, and especially those with smaller form-factors, were not designed primarily with throughput in mind. The move toward SGSN bypass with Direct Tunnel is evidence of this.

This view that SGSN and GGSN equipment is, by nature, not suited to high-traffic load and is inherently expensive is not sustainable, however. Vendors continue to introduce new hardware platforms that can comfortably meet throughput

demands from 3G access and operators are tending to refresh legacy PS cores to take advantage of this.

An Internet offload gateway should in theory be less expensive than a new GGSN, which typically will have additional software costs (although this is negotiable, according to vendor pricing strategies) and be designed to manage a greater signaling/transaction load. An offload gateway, by contrast, is optimized for data throughput, with fewer user sessions and lower transaction rate.

A challenge to the Internet offload gateway is the impact on the operational environment. Introducing new elements into the network, testing and maintaining them, and so on, is costly, and so the business case must be compelling for operators to contemplate it.

Content Distribution & Optimization

As the offload gateway concept has gained currency the industry has started to think about how to leverage this distributed architecture to push content caches and content optimization closer to the end user. The argument being that the user will experience faster load times and benefit from having content delivered in the most appropriate format for the device and connection speed. The operator will make more efficient use of radio resources by not delivering unnecessarily large streams of files and save some money on core transport costs.

Figure 8 above shows that using an offload gateway content can be served either direct from the Internet (e.g., an Akamai server), or can be served from a local server (a.k.a. a "Mobile CDN" or "CDN Edge Server") collocated with the gateway. In both cases there are potential gains over having to push traffic through the mobile core.

Figure 9 shows the impact on load times of popular websites using the three different methods. These figures have been provided by Stoke and have not been confirmed independently by *Heavy Reading*. Our insight into operator test results, however, suggests there is at least some user benefit from distributed caching.

Figure 9: Impact on Load Times of Mobile CDN (Distributed)

WEBSITE	PUBLIC CDN*	PUBLIC CDN WITH 3G CN OFFLOAD**	GAIN	MOBILE CDN***	TOTAL GAIN
trailers.apple.com	10.5s	7.2	31%	4	62%
nytimes.com	13s	10.1	22%	8.8	32%
facebook home page	7.3s	4.8	34%	3.1	57%
facebook photo album	8.4s	7.4	12%	5.2	39%
trailers.apple.com	10.5s	7.2	31%	4	62%
nytimes.com	13s	10.1	22%	8.8	32%

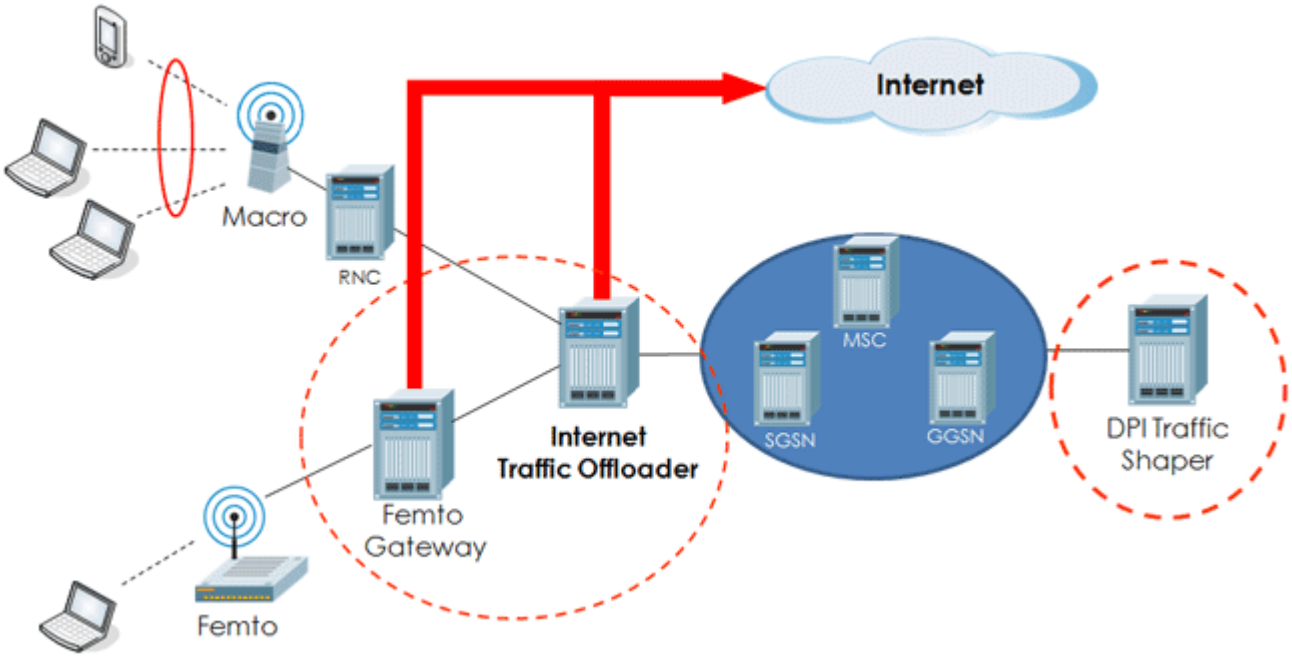
***** *non-offloaded traffic (50ms 3G CN latency) using Akamai commercial network*
*** offloaded traffic (3G CN bypassed) using Akamai commercial network*
**** offloaded traffic using Akamai local private EdgeServer (collocated with offload gateway)*
Source: Stoke

What started as a spin-off benefit to the cost-reduction objective of offload gateways has now arguably become the driving rationale for them. Distributed caching certainly aligns with the central thesis of this paper, which is that Internet offload should be services-led. A word of warning, however: The overall argument for distributed caching and content optimization is far broader than can be discussed in depth here.

Offload & Traffic Management

An unintended consequence of core network offload is that by diverting traffic from the core it becomes harder for the operator to meter usage, bill for traffic and apply traffic management techniques, since these functions all reside in the core. This is also an issue for the content caching concept discussed in the previous section. This problem is illustrated in **Figure 10**.

Figure 10: Internet Offload & Traffic Management Conflict



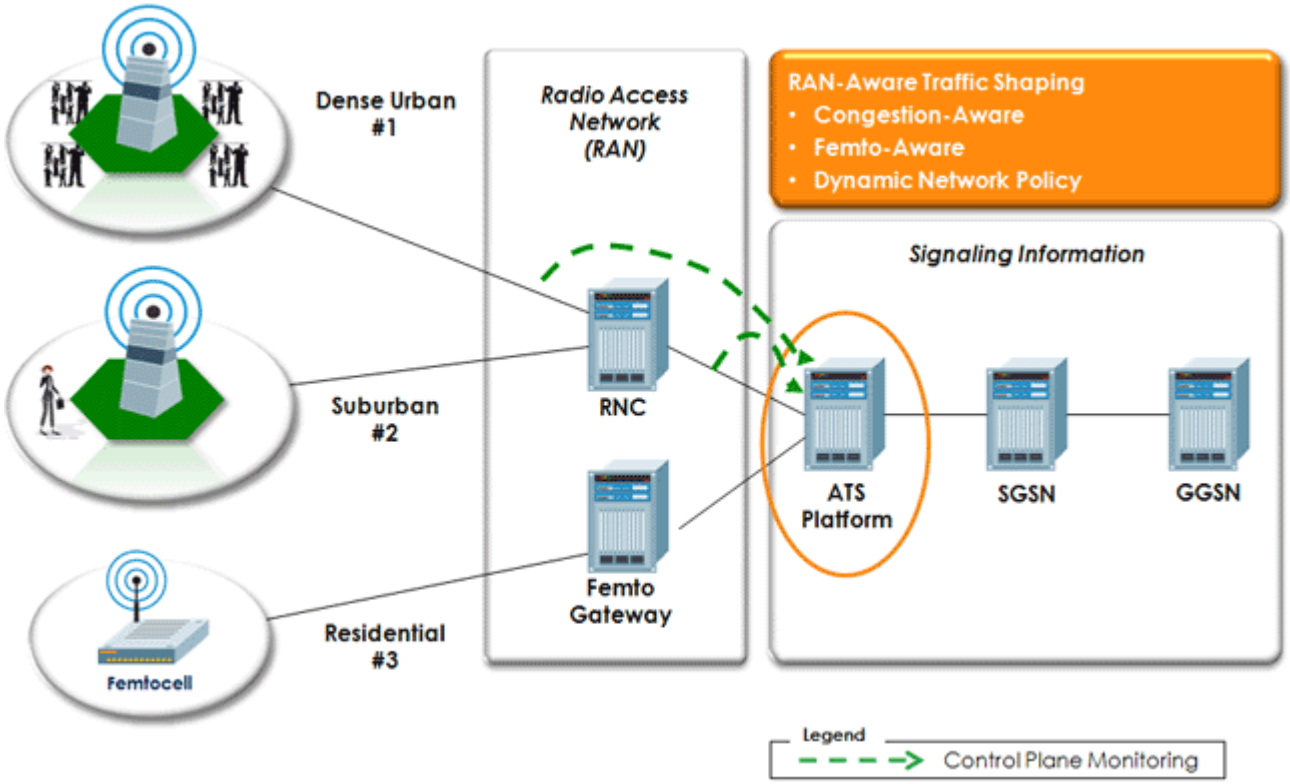
Source: Radisys

A proposed solution is to make the offload gateway also function as a traffic management device. This effectively means it should integrate DPI and policy enforcement capability that mirrors capability in the core network. Moreover, because traffic management must increasingly be RAN-aware, it is logical to take advantage of the gateway's distributed location close to the RAN. This is shown in **Figure 11**.

The implication of this solution is that the offload gateway should have access to standardized RAN and policy management interfaces that enable it to act dynamically according to prevailing load conditions – which at present is subject to ongoing technical development and standards work. This is an interesting area

to explore, but again, a word of warning: The issue of distributed policy enforcement and RAN-aware traffic shaping is a broader one than can be discussed in depth in this paper.

Figure 11: Traffic Shaping & Mobile Offload



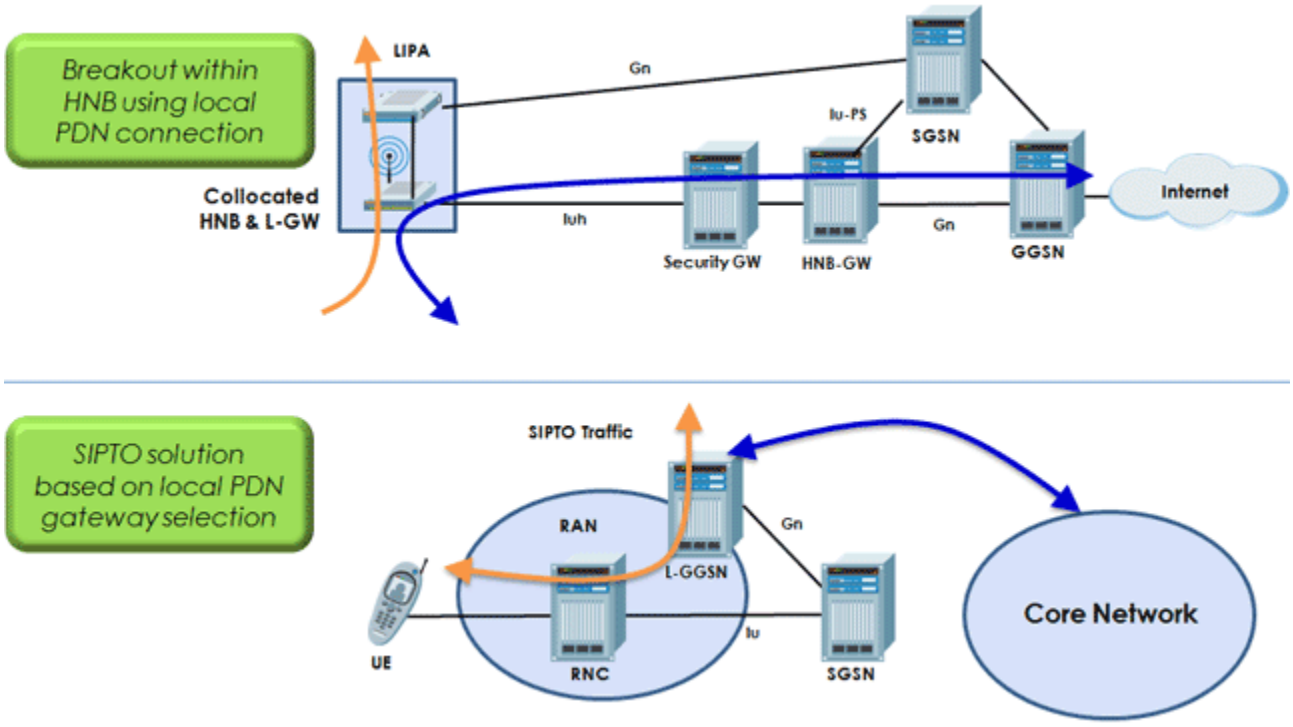
Source: Radisys

SIPTO & LIPA

Selected IP Traffic Offload (SIPTO) and Local IP Access (LIPA) are two other offload technologies already under discussion in 3GPP TR 23.829. They are often discussed in the same context but are different in scope and intent. LIPA is really part of the femtocell architecture and is designed for use cases such as the residential user who wants to access local services, but doesn't need to access the wide-area network to do so – for example, to access a film or photo album on home media server from a smartphone.

SIPTO is more directly related to the core network offload. In this case, the intent is to distribute packet gateways (GGSNs, in effect) so that traffic is not concentrated on a small number of core nodes. This requires some changes to SGSNs, HLRs and MMEs, and likely local routing infrastructure at the distribution site, but in theory should be reasonably straightforward to implement. As with everything, there will challenges, particularly as relates to arguments for and against the distribution of GGSN and P-GWs.

Figure 12: Selected IP Traffic Offload (SIPTO)



Source: Radisys

Conclusion

Internet offload for mobile operators is a response to strong growth of mobile broadband and new smartphone usage patterns. It should be thought of as a services-led concept that will help further drive data penetration.

There are a number of innovations both in the RAN offload and core offload concepts that are of potential value to mobile operators. WiFi, femtocells, and Internet offload gateways provide a toolbox of solutions that will suit different circumstances and objectives.

There is a sense that Internet offload is promoted by individuals and organizations intent on disrupting the industry. However, many of these innovations are being gradually integrated into the mainstream through formal standards development work and best practices emerging from actual deployments.

This paper has discussed a wide range of offload mechanisms that are not necessarily similar in nature. This makes it difficult to talk generically about a single technology and highlights the need for broad view of Internet offload. If there is a unifying theme, however, it is the transition to the Evolved Packet Core (EPC), which impacts all the offload technologies discussed from WiFi through femtocells and Internet offload gateways.

The adoption of EPC as a common core for multiple access networks – including 4G LTE, 3G, WiFi and licensed small cells – makes it central to RAN offload. The potential to distribute EPC elements such as S/P-GWs toward the edge of the network aligns closely with the core offload concept and the various forms of offload gateway discussed in this paper.