White Paper

Scaling Up Policy: Balancing Cost & Functionality in the LTE Era

Prepared by

Graham Finnie
Chief Analyst, *Heavy Reading*
www.heavyreading.com

with input from

Wireless 20|20

on behalf of

BroadHop

www.broadhop.com

September 2011

Executive Summary

In the past two years, policy management caught fire as network operators sought better ways to manage the way bandwidth is allocated and congestion is handled. Now, many are looking to move on from these early deployments, and seeking to put policy at the heart of their traffic management and service development strategies.

But as policy deployments scale up, it raises major new issues for operators. Can policy servers cope as new use cases are added? What will it cost? And can the new business case really stack up?

The purpose of this paper is to examine these issues. In it, we explain why policy servers must scale up massively as mobile operators move into the LTE era. We look in detail at a key measure of power – transactions per second (TPS) – and model the impact of new policy use cases on TPS requirements. And we analyze how this affects the cost of policy deployments, using real cost data. We conclude that, so long as operators choose a next-generation policy solution capable of scaling to meet demand, at a predictable and manageable cost, the business case looks viable in principle.

In the first stage of dealing with the new challenges thrown up by the growth of mobile broadband and data, network operators have been focused on traffic and congestion management. Many identified policy servers as the key means to apply more rational policies to applications and customers when networks became congested, usually by dividing customers into tiers with different data volume limits, and devising policies for what happened when limits were breached. This has become widely known as fair use management. Other applications, including "bill shock," were added to ensure that customers understood what and how much they were being charged for.

These early deployments of policy are relatively simple and placed a relatively light load on the policy servers in use. But evidence from *Heavy Reading* surveys and from discussions with operators strongly suggests that they are now shifting toward a new set of policy use cases focused on differentiation, personalization and monetization. And in this new environment, many more policies will be devised and deployed. This will be especially true as operators start to make the transition to LTE. The flat all-IP architecture of LTE makes it even more important to control QoS and customer experience in a rational way that maximizes value to customers at a manageable cost to the operator.

These shifts are likely to result in a very rapid increase in the scale of policy deployments. In a survey of network operators, we found a strong intent to add new policies and policy conditions fairly frequently, resulting in relatively complex deployments within three years of initial deployment of policy servers. Along with this, of course, the number of customers to which policies will be applied will multiply as mobile broadband spreads through the user population.

For operators, understanding both the scaling and cost implications of this transition is vitally important. Policy will become a key part of the core network architecture, and like other elements of that architecture it will normally be expected to have a lifetime of at least three years – a long time in the dynamic environment of mobile broadband. Wrong decisions early in the cycle may have highly damaging effects as policy use matures.

In order to more fully understand the implications, *Heavy Reading* has worked with BroadHop and the wireless network modeling company Wireless 20/20 to develop a detailed model of the effects of this transition. In previous work, we have identified TPS as the key scaling metric in policy, so this is the metric used in this modeling work.

We begin with a "Base Case" model – a very simple policy deployment consisting of fair use quota management and temporary pay-as-you-go passes. We then run the model with a medium-term deployment and a longer-term deployment that steadily add more policies, in each case at a rate consistent with our survey findings regarding operator intentions.

This demonstrates that a 3GPP network of 10 million mobile subscribers will scale, based on our assumptions, from handling about 2,200 TPS in the Base Case to almost 23,000 TPS in the Mature Case, which is assumed to be after three years.

> A 3GPP network of 10 million mobile subscribers will scale from about 2,200 TPS in the Base Case to almost 23,000 TPS in the Mature Case, after three years

In the second part of this exercise, we modeled the TPS impact of beginning the transition to an all-LTE network. This also assumes a network with 10 million subscribers, in which the customers are gradually transitioning to LTE, and that during this transition, wireline customers are also transferred to the same policy environment. In this case, policy scales from nearly 25,000 TPS at the end of Year 1 to more than 75,000 TPS at the end of Year 3.

Finally, through our collaboration with Wireless 20/20, we looked at the cost implications in our LTE deployment. This sets the cost of a relatively highly-loaded policy deployment running a rich range of services, including voice over LTE (VoLTE), within the cost of the overall deployment, using pricing data provided by BroadHop for its next-generation Quantum Network Suite policy platform, and pricing obtained from another legacy vendor of a policy server product.

This modeling exercise showed that the cost of policy is only a small proportion (around 2 percent) of the overall cost of a new LTE build over a five-year build period in the next-generation policy case – and given the high strategic value of policy, this suggests to us that the return on investment (ROI) is likely to be positive in the short to medium term. However, we also found that BroadHop's pricing on a TPS basis is less than half that for the legacy policy server product we modeled, so this has an impact on the business case.

We also show in this work that although cost does inevitably rise as policy use accelerates, this cost can be controllable and predictable if a next-generation policy infrastructure is deployed.

There are of course many challenges when scaling up a policy deployment; not all of them covered in this paper. But on the key measure of TPS, our modeling demonstrates that a next-generation policy platform is needed to scale to handle extremely high levels of demand and policy complexity, and meet the core need for a solution that can be deployed for the longer term, at a manageable cost.

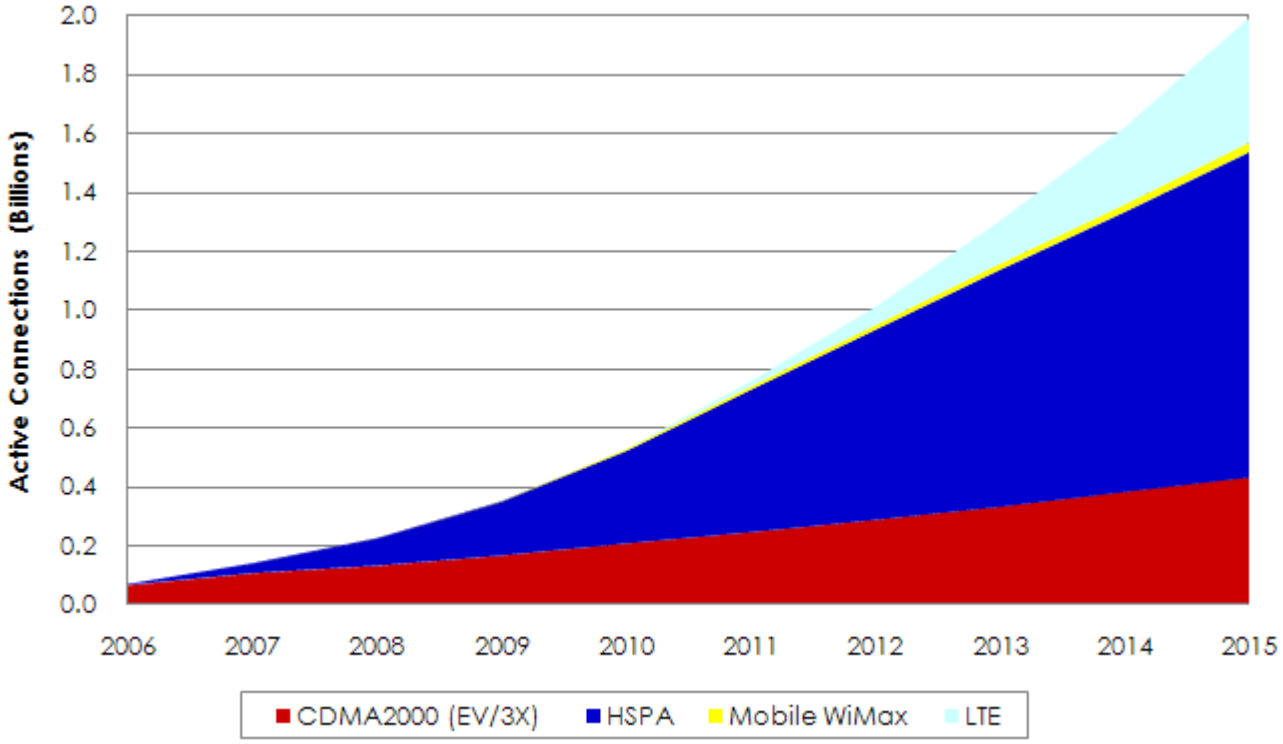
The Road to Policy Proliferation

Policy control is not new. The first true policy servers were deployed in wireline networks about six or seven years ago, and the initial ramp-up in deployments was fairly slow. In the past 18 months, however, all this has changed as a result of the huge success of mobile broadband and – in parallel – the spectacular ascent of the smartphone.

In this section, we discuss the impact of this ramp-up and how it will affect policy management deployments. We show that as the number of broadband customers, smartphones and applications grow, it will put even more pressure on operators to better manage the way bandwidth is used. And we show that operators are already preparing to scale up policy deployments to meet this need.

As **Figure 1** shows, the rapid ramp-up in current-generation mobile broadband will be followed by an equally rapid rise in LTE. According to Pyramid Research – like *Heavy Reading*, a division of the Light Reading Communications Network – worldwide broadband service connections grew from 141 million in 2007 to 531 million in 2010, and will reach nearly 2 billion by 2015. LTE – deployment of which began in 2010 – will account for more than 400 million of those users in 2015.

Figure 1: Active Mobile Broadband Connections Worldwide, 2006-2015



Source: Pyramid Research

When considering the impact on policy of this increase, it is important to consider the wider impact on usage patterns, since this creates a kind of multiplier effect that has a major impact on the load on the policy platform.

The Broadband Multiplier

More Broadband Customers

The first metric, of course, is number of broadband customers. Pyramid's forecast shows the proportion of all mobile customers subscribing to a mobile data service at 6 percent in 2006, rising to 21 percent in 2010 and 50 percent in 2015. (Naturally, proportions and growth vary by region: In North America, these figures are 35 percent, 59 percent and 78 percent, respectively.) By the end of this decade, broadband will likely be almost universal in many countries.

More (& More Powerful) Devices

Subscribing to a service does not mean that it will be heavily used; that requires suitable devices and applications. Initially the device that drove high usage was the dongle, which was effectively a wireline replacement or backup service for PCs. More important in the longer run is the smartphone, because it will be both much more widely distributed and much more nomadic. Smartphones encourage the use of IP/ Internet applications and bandwidth, creating the key circumstances for both higher broadband usage and a more central role for policy.

The total number of smartphones in use will exceed 1 billion by 2015, and these devices are becoming more capable, with better screen resolution, faster processors and multi-tasking OS

Smartphones are spreading rapidly, driven by fashion, utility and price. According to Pyramid, more than 200 million smartphones were sold worldwide in 2010, and this figure is expected to exceed 500 million by 2015. The total number of smartphones in use will exceed 1 billion by 2015, and these devices are becoming more capable, with better screen resolution, faster processors and multi-tasking OS.

On top of all this, a whole new range of devices designed to use mobile networks for machine-to-machine communications (M2M) is spreading fast, and these specialist devices put widely differing demands on the network.

In summary, three effects are reinforcing each other here:

- More customers will have smartphones
- Smartphones will be more powerful
- The range of mobile devices will proliferate as M2M spreads

More Applications

Many applications already run across mobile broadband networks, and it is reasonable to expect that the number of applications used by customers (both in parallel and in serial) will continue to increase markedly. More applications mean more policy complexity, and these applications (both the type and the range) increasingly resemble those already used in wireline services.

Although detailed analysis of usage is beyond the scope of this report, a typical palette of applications for a relatively active mobile broadband user might include VoIP (e.g., Skype), video streaming (e.g., YouTube), video downloading (various), Web searches (e.g., Google), news (e.g., BBC), gaming (various), social networking (e.g., Facebook) and location services (e.g., Google Maps). And again, the range of applications will broaden as M2M spreads, creating specialized needs in industries ranging from health to energy to logistics.

A proliferation in applications has a wide range of impacts. An important ancillary characteristic of this traffic is that patterns are less predictable than in the past, as new applications rise to prominence and others fade away. A second is that the network requirements of different applications vary widely. And a third is that customer's perception of their quality of experience will depend on the kinds of applications they use and value, which will vary widely.

More Traffic

All of the above – connections, devices, applications – create the multiplier effect on traffic that has become a central concern for so many operators. In most networks where broadband and smartphones or dongles are in wide use, traffic on mobile networks has been growing by at least 100 percent per annum. Though this figure will moderate as networks mature, it is reasonable to assume that the established pattern in wireline networks – where traffic increases by around 40-50 percent per annum year on year – will also become the norm in mobile networks.

The Policy Multiplier

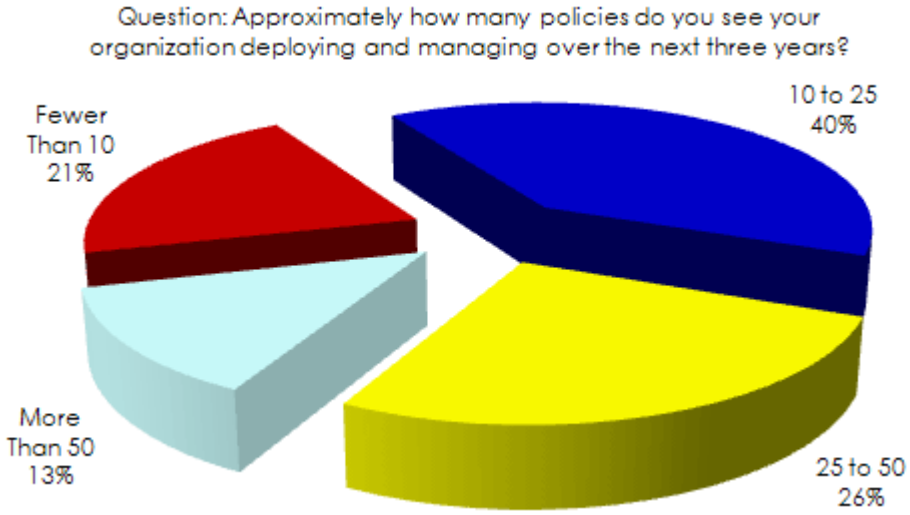
These trends have had a radical impact on mobile network operator strategies, both defensive and offensive. Defensively, operators want to retain as much control as they can over traffic, applications and customers. Offensively, they want to take maximum advantage of a unique opportunity to provide a rich, targeted set of services that better meet the needs of individual customers.

As a result, and as research by *Heavy Reading* on network operator plans for policy has consistently shown, there is now a strong desire to use policy in a wider range of contexts and to put it at the heart of both network and service strategy. Specifically, the range of use cases includes:

- **Refining how scarce resources, especially radio spectrum, are allocated**
  - Customer-centric congestion management
  - Yield management
- **Increasing the value of broadband connections**
  - Improving the performance of applications that customers value
  - "Rationing" only when it is necessary
    - Don't ration at night or off-peak periods
    - Don't restrict use of valued applications that have only marginal impact on bandwidth consumption
- **Giving customers control over their connection**
  - Information, notification, upgrade options
  - Dashboards and portals
  - Dynamic parental controls
- **Aiding service discovery**
  - Campaign & promotion management
  - Special offers and loyalty bonuses
- **Increasing market share and reach**
  - Packages suited to all types of users, including low and high income users, temporary users, enterprise users and M2M services

The effect of these plans is shown in **Figures 2** and **3**. When asked how many policies they expect to deploy in the next three years, most operators said "more than 10," with an implied average of 15 – far more than actually deployed today.

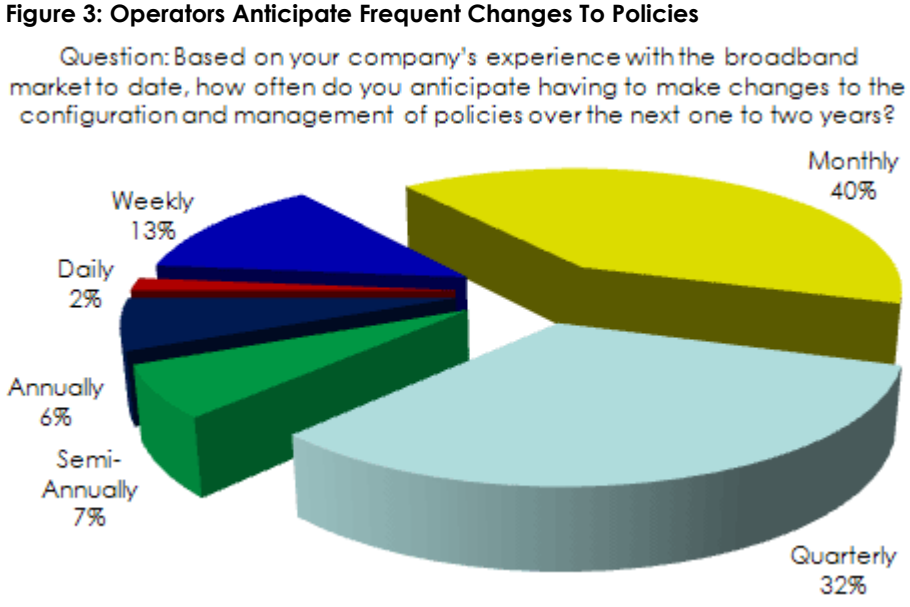
Figure 2: Many More Policies Will Be Added



Source: Heavy Reading survey of 71 network operators, June 2010

In a separate question, we asked how often operators expected to modify policies, and found that the average operator expected to change policies about once every two or three weeks.

Figure 3: Operators Anticipate Frequent Changes To Policies



Source: Heavy Reading survey of 71 network operators, June 2010

This is not simply a theoretical expectation. Pioneering operators are already putting in place relatively complex deployments and expect them to become even more complex in future. Taking two examples among many:

- An operator in Eastern Europe that justified its deployment on the basis of a single policy application, but now has four applications running and four more in planning stage. This has resulted in the operator going to tender for a new policy deployment able to handle the higher expected load.
- A Tier 1 operator with multiple national operations that deployed policy to meet an immediate need to provide a range of policy-controlled bandwidth tiers, but said that some of its local operations were already deploying further policies that offered greater granularity. The company said its marketing departments had developed long lists of new ideas, and it foresaw a "massive scaling up" in policy complexity over the next three years.

In summary, mobile broadband has opened the door to an entirely new era in mobile services, driving operators to overhaul both control layer and service layer. In both cases, policy will play an increasingly important role, and this means that policy platforms must be capable of handling much higher processing loads in the future. In the next section, we consider the technical implications of this scaling up in more detail.

Scaling Up Policy in a 3G Network

This section considers the impact of more complex policy deployments, looking specifically at the impact on TPS. We begin with a short primer that explains how policy technology works in a 3GPP context. We also show how policy use cases translate into TPS. And using the operator expectations for policy scaling set out in the last section, we show that TPS will scale by several orders of magnitude over a relatively short period.

How Policy Works

Where a policy implementation conforms to the 3GPP Policy & Charging Control (PCC) standards, the core elements in the architecture are a policy server that makes policy decisions and conforms to the Policy & Charging Rules Function (PCRF) standard, and one or more enforcement devices that conform to the Policy & Charging Enforcement Function (PCEF) standard. PCRF and PCEF devices are connected over the Gx interface.

In a 3G broadband network, when a Packet Data Protocol (PDP) session is begun, the mobile device attaches to the network and activates a PDP context. This allocates a PDP context data structure in the SGSN that the subscriber is currently visiting and the GGSN serving the subscriber's access point.

Every PDP context can be associated with an operator-defined policy, such as applying a specific level of quality to the session. This generates interactions or **transactions** among the policy elements. As more sophisticated devices are added – perhaps running several simultaneous applications or content services for which different policies must be applied – and as new policy triggers such as location are employed, a rapid rise in the number of transactions is inevitable.

In the case of 4G-LTE, this trend is likely to accelerate. In the new architecture, policy and QoS are inherent and the policy server will need to be referred to every time a session is established or modified. Among other things, this will include applying a suitable level of QoS to handle voice services and any other services that require priority treatment, such as premium video or gaming – both of which are likely to become more widespread as LTE spreads.

But exactly how many transactions must the policy platform be able to serve? In answering this question, most network operators will want to have a feel for the transaction scaling rate at least three years from initial deployment, unless the deployment is simply a short-term tactical point solution to solve a specific problem. The rest of this section considers this issue.

Modeling Policy in a 3G Network

Our initial Base Case for analyzing TPS load is as follows:

- Total mobile subscriber base of 10 million.
- 30 percent of these subscribers are data users.
- 30 percent of these data users are active in the busy hour.
- The number of PDP contexts per subscriber at busy hour is 1.2 (for most it is 1, but for a minority it may be 2).

- The only policy applications running are fair use (i.e., quota) management and "pay as you go" or casual use passes.

This generates an initial policy load, in terms of TPS, of about 2,200 TPS, of which about 57 percent is accounted for by initial session authorization.

This load, while not trivial, is within the capabilities of most policy platforms, including many first-generation legacy platforms.

In the second case, which we call the Growth Case, and which we assume is in place about 18 months after the Base Case, all metrics in the initial case have increased, generating the multiplier effect described in the last section:

- Total mobile subscriber base of 10 million.
- 40 percent of these subscribers are data users.
- 40 percent of data users are active at busy hour (the expectation is that this figure will rise as the number of applications in routine use increases).
- The number of PDP contexts per subscriber at busy hour is 1.2 (for most it is 1, but for a minority it may be 2).
- The number of applications running rises to five, including Congestion Management, Facebook Add-on Pack and Premium Video with Online Charging, in addition to the Base Case applications.*

This scenario generates a policy load of more than 7,950 TPS, increasing the policy load approximately four-fold over the Base Case. This load will strain some legacy policy servers, though most should be able to handle it.

In the third case, which we call the Mature Case, assumed to be operational after about three years, the operator has ramped up its policy deployment with a range of new policies, and the following assumptions are now made:

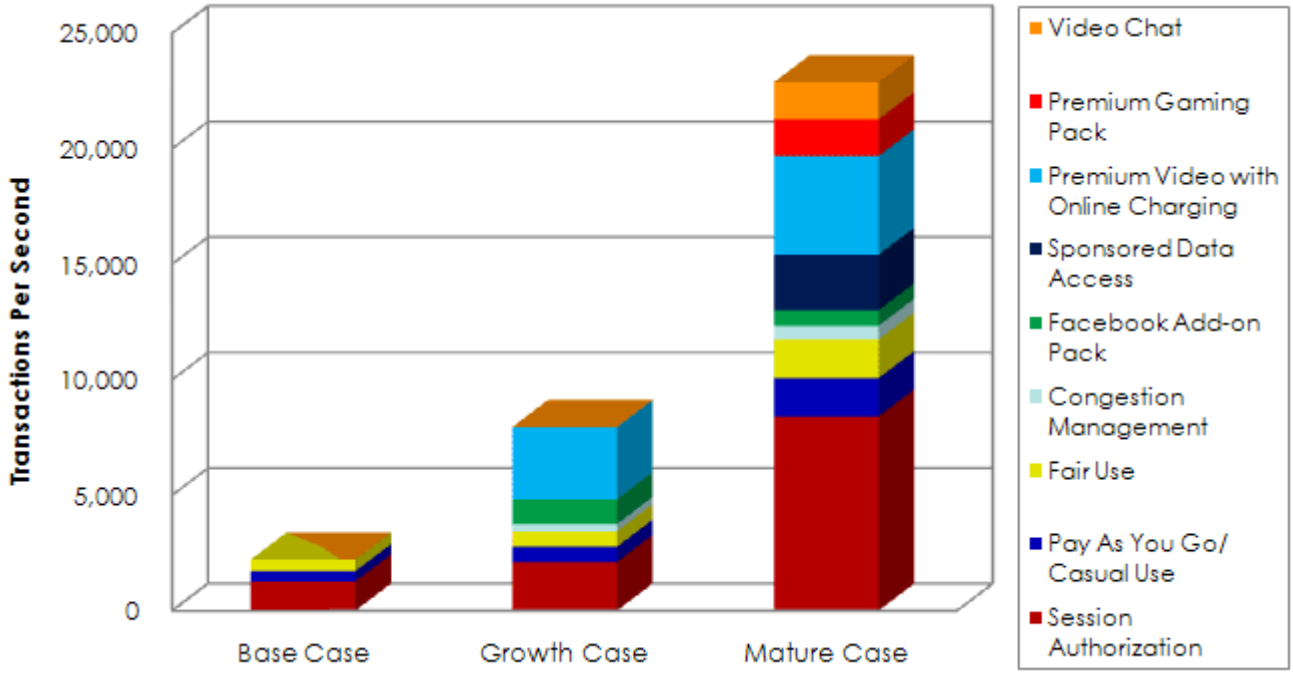
- Total mobile subscriber base of 10 million.
- 60 percent of subscribers are data users.
- 50 percent of data users are active at busy hour (the expectation is that multiple applications on smartphones will generate much higher usage at busy hour).
- The number of PDP contexts per subscriber at busy hour is 1.5.
- The number of applications rises to eight, including Sponsored Data Access, Premium Gaming Pack and Video Chat, in addition to all those listed in the Base Case and Growth Case.

In this scenario, total TPS would be almost 23,000, well in excess of the capabilities of legacy policy deployments. This would be likely to require a new deployment unless the initial deployment has been made and equipment chosen with this level of scaling in mind.

*Note that while a list of specific applications was required in order to realistically model impacts, in practice the actual applications will vary widely by operator. The point of this exercise is to model a situation in which the number of policy use cases rises, in line with network operator expectations as set out in Section 1.

Bear in mind also that this scenario is for a medium-sized deployment of 10 million mobile connections, with 6 million actual mobile **data** customers in the Mature Case. In large Tier 1 networks with 50 million customers or more, the policy load will be much larger – on the order of 100,000 TPS, far beyond the scope of the initial generation of policy server platforms.

Figure 4: 3G Policy Sample Use Cases – Impact on TPS by Application



Source: Heavy Reading

Figure 5: Summary of Three Sample Use Cases

CASE	TOTAL MOBILE DATA CUSTOMERS	ACTIVE AT BUSY HOUR	# OF POLICY USE CASES	POLICY LOAD IN TPS
Base Case [0 months]	3 million	0.9 million	2	2,210 TPS
Growth Case [18 months]	4 million	1.6 million	5	7,950 TPS
Mature Case [36 months]	6 million	3.0 million	8	22,876 TPS

Source: Heavy Reading

In summary, we have shown in this section that the multiplier effect described earlier, in conjunction with network operators' desire to differentiate and refine service offers and traffic management, is likely to lead to a very large increase in the load on policy servers.

Scaling Up Policy in an LTE Network

In this section, we examine the likely impact of the transition to LTE on policy requirements. We show that the shift to LTE will put even more pressure on policy deployments, both because policy must be applied to all mobile customers in an all-IP network infrastructure, and because the transition will require more innovation in the way service packages are created and applications handled. The result is that policy load will scale by several orders of magnitude as LTE matures.

LTE represents one of the biggest technology changes ever made in mobile networking, not least because it will accelerate the migration of primary circuit-switched voice to IP-based voice using voice over LTE (VoLTE). The key impact here is that VoLTE **requires** the use of policy management, which must be applied to all customers. The timing of this transition is, of course, debatable: Some operators will move only cautiously to VoLTE, once they have confidence that their LTE deployment is mature and stable; others may deploy VoLTE from the start. *Heavy Reading* takes the view that in most big 3G networks, VoLTE deployment will lag initial LTE deployment by one to two years, and this is taken into account in the discussion and modeling that follows.

A second core effect of LTE is that it will encourage the development of new policy use cases, though many of these are hard to predict in advance. In LTE networks, all customers will be using smartphones or similar devices, and will have access to even more bandwidth, encouraging development of new, demanding applications, especially in the area of video. More generally, though, the high investment in LTE will require operators to ensure that their portfolio of services is rich enough to maintain and even increase ARPU. The transition to LTE will encourage even more competition from third-party application providers in all areas, including communications services, and network operators must respond with an ever-richer and more personalized set of service options for customers.

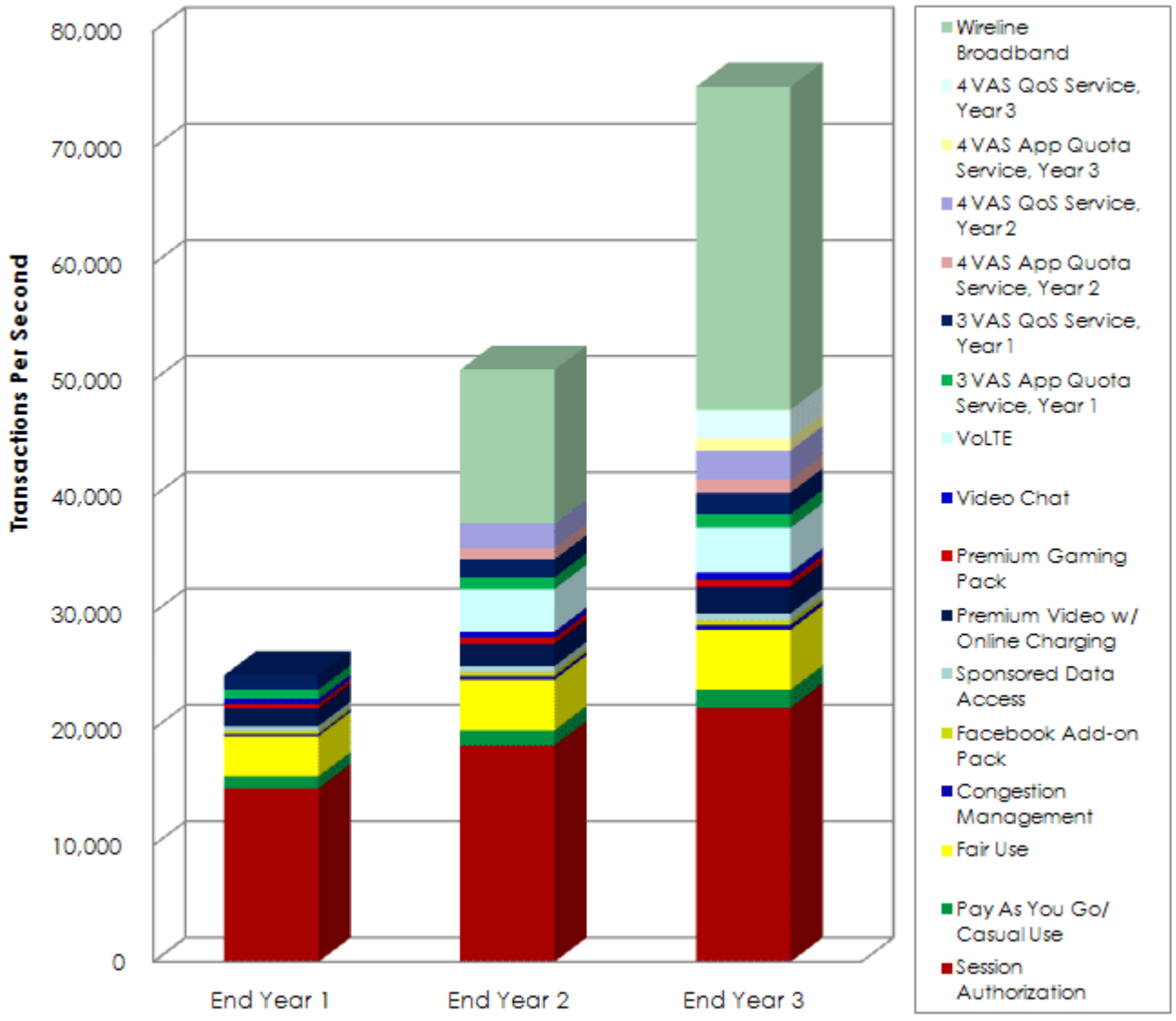
In the discussion and modeling in this section, we have assumed a specific set of use cases, but new use cases added in subsequent years are not identified. Our working assumption is that it is too early to identify these use cases, but we expect that, as with the Internet generally, new use cases will inevitably emerge.

Bearing this in mind, how quickly might we expect policy deployments to scale up? In this modeling exercise, we make the following key assumptions:

- The network supports 10 million subscribers; it is assumed that the subscriber count rises 5 percent per annum.
- At the starting point, all customers are 3G customers, and are gradually transitioned to LTE through the forecast period.
- All customers are mobile data customers.
- The number of customers active at busy hour rises from 50 percent to 80 percent. The assumption is that the wider range of applications in use means that most customers' phones will be "active" most of the time.
- The number of policy use cases grows from 14 at end of Year 1 to 30 at end of Year 3.
- VoLTE begins to have an impact in the network during Year 2.
- The policy server begins to handle wireline broadband use cases as well, from Year 2 onward.

The results of this modeling exercise are shown in **Figure 6**. This shows that total TPS scales from nearly 25,000 TPS at the end of Year 1 to about 51,000 TPS at the end of Year 2 and 75,000 TPS at the end of Year 3.

Figure 6: Scaling Up Policy in an LTE Network – A Three-Year View



Source: Heavy Reading

In this section, we have shown that, as operators make the transition to LTE, the load on servers will continue to climb quickly. This raises some major issues about the ability of servers to cope and the likely impact on cost. This is discussed in the next section.

Business Implications for a Policy Deployment

We showed in the previous section that we can expect massive scaling up of policy deployments and the TPS associated with this, given reasonable assumptions about the ramping up of policy use cases in both existing 3G and future LTE networks.

In this section we consider the business implications of this, focusing specifically on the cost of deploying policy in an LTE network. We show that, while policy management forms only a small part of the cost of an LTE build, a fully-loaded policy server running many use cases will become a significant element in core capex if a legacy server is deployed.

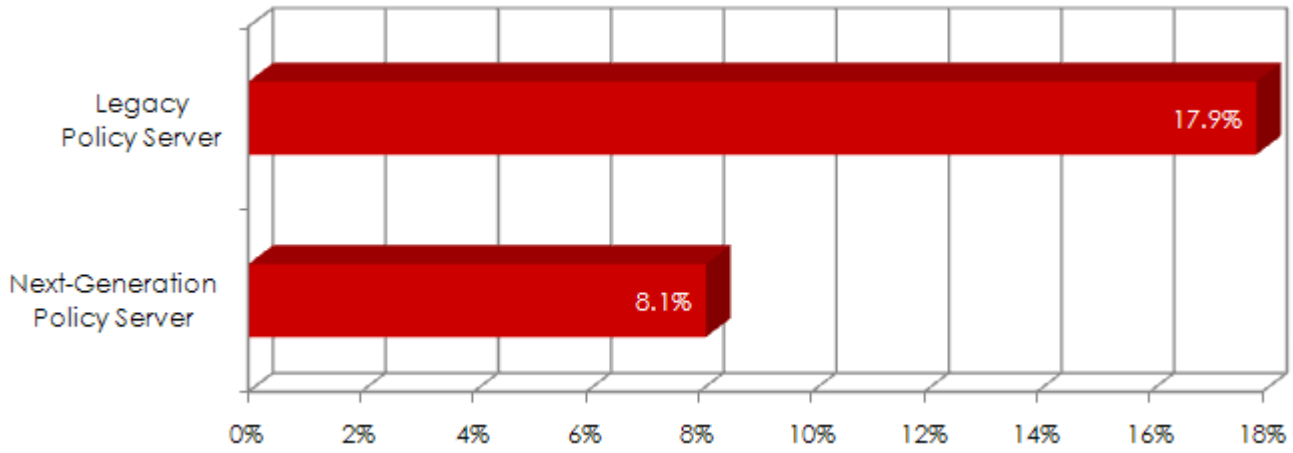
We modeled the actual cost of the deployments based on the modeling assumptions set out in **Sections 2** and **3** of this report, using pricing information provided by BroadHop and Wireless 20/20. The modeled costs include the cost of hardware, software and professional/engineering services.

These costs were set in the context of the overall cost of rolling out an LTE network, using detailed real-world data collated by WiROI. The modeled network was assumed to have a total of 5.5 million active subscribers at the end of Year 5, and that multiple policies were being applied (including VoLTE), resulting in a total load on the policy server at Year 5 of 42,000 TPS.

> The model found that policy capex represents 18 percent of all core capex for the legacy product, while it was only about 8 percent of core capex for the next-generation product

This modeling showed that legacy policy server costs are significant in the context of core network capex. The model found that policy capex represents 18 percent of all core capex on the same five-year view. By contrast, the next-generation product cost as a proportion of core capex is only about 8 percent.

Figure 7: Policy Server Spend as a Percentage of All Core Capex, LTE Build, Five-Year View



Source: Company or Heavy Reading

In the context of **all** capex – excluding any spending on site development but including spending on base stations, backhaul equipment, core network equipment and maintenance – the next-generation policy platform costs amount to approximately 2 percent.

Even with a fully-loaded policy deployment running a wide range of use cases, including VoLTE, the cost of the next-generation policy deployment remains relatively modest in the context of total capex. It's also worth noting that policy server costs rise linearly with time, in line with the TPS load, and never exceed 2 percent in any one year.

Among other factors that we believe account for the difference in costs between the two modeled products were:

- Extensive use of virtualization techniques, which reduces hardware costs and enables pricing that rises linearly with processing requirements
- Elimination of the need for a third-party database to store state information, and the software and hardware costs associated with using these databases
- A more open and more flexible policy creation environment, reducing engineering/professional service costs associated with introducing new policies or amending existing ones

In summary, this paper has shown that, with a next-generation policy product, it is possible to massively scale policy deployment without adversely impacting cost structures. Given the central strategic role this modeling exercise assumes for policy, controlling a wide range of services and service attributes, this strongly suggests a good ROI should be possible, even for a complex policy deployment.