Bringing Intelligence to the Cloud Edge

Bringing compute capabilities closer to the real value

Executive Summary

The telecommunications industry is moving to cloud-based technologies at the network edge to help tackle the explosion of mobile video consumption, service mobility, virtualization of Customer-Premise Equipment (CPE), the Internet of Things (IoT), and other latency-sensitive applications. But unfortunately today, much of the compute still happens at the center of the network or on the client device. This situation results in more traffic, higher latency, and less flexibility.

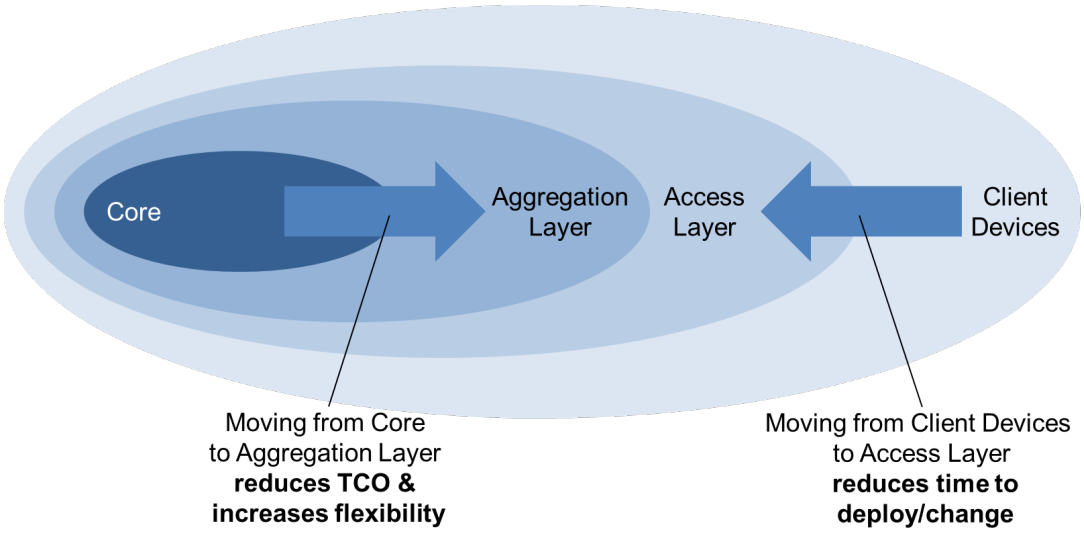
Moving processing closer together—away from the two furthest points where it happens today, the network core and the client, toward the network edge and aggregation layers—can speed deployment and reduce TCO for carriers. In these environments, flexible general-purpose compute will drive the most value. If "data is the new currency", then positioning compute should be like choosing an oil well site: as close to the resource as possible, yet comprehending the geographic transport issues.

The Changing Carrier Landscape

Carriers have transitioned rapidly to packet switching networks as the world has embraced internet and cloud technologies. Services like VoIP (Voice over IP) have moved from a scourge to a monetizable service that carriers like Telefonica and Comcast are now promoting. **Data** is now the key differentiator of access plans, and calling/texting is practically valueless when provisioning new customers.

While much of the recent IT investment has been for building cloud infrastructure at the edge/access area, much of the real compute is still happening at the client and the network core where it is both expensive and harder to deploy/manage. The network core is more proprietary and expensive with less innovation; money spent there is focused mainly on operations and rarely creates a competitive differentiation. As carriers scope out their investments, **the concept of a more intelligent flexible cloud can help guide carriers towards better business outcomes**—driving investment towards more flexible and higher ROI compute innovation that is closer to where the network's real value resides.

Figure 1: Moving Compute Brings Significant Benefits to Carriers



The key tenets that underlie this intelligent flexible cloud strategy are:

- Moving compute, storage, and applications out from the network core, **closer to the data**
- Migrating virtualized client compute **to the network edge** where it can be deployed/re-provisioned faster
- Focusing investment on driving the **highest flexibility and ROI**
- Building **agnostic, standardized infrastructures** to exploit open platforms and frameworks on which to develop new applications and business models
- Enabling a **heterogeneous infrastructure** for seamless application support

Innovation drives the competitive advantage in a market that is moving quickly. Carriers are working to make their infrastructures more flexible by building out service clouds instead of relying on purpose-built platforms. Carriers that don't take aggressive steps today to stay ahead may find their business targeted not only by traditional rivals, but also possibly by new challengers (companies and technologies), as the race to deliver compelling new services to the end user intensifies.

Optimizing for More Flexible and Cost-effective Compute

Today's innovation is happening at the edge of the network at the device and access level. With customers clamoring to take advantage of greater mobility and the emerging Internet of Things (IoT), the stakes are even higher. IDC estimates that 90% of all IoT data will be stored on service provider platforms, as half of the IT networks move from excess capacity to constrained capacity (with almost 10% being totally overwhelmed).

For service providers, shuttling all of this data being collected at the edge back to the center of the network for execution, analysis, and contextualization will create needless amounts of network traffic and congestion. Moving processing from the center of the cloud—to either the aggregation or edge layer—reduces both traffic and latency times.

Processing data at the gateway will rise in prominence as the amount of data (especially IoT data) grows exponentially. IDC supports this concept and believes that 40% of the data will be stored, processed, analyzed, and acted upon closer to the edge.

Network value will align with data, and carriers will build out new cloud infrastructure or business models to monetize data closer to the edge. These compute platforms can help extract the meaning from the massive amounts of data, then pass that meaning to the core, instead of simply carrying/forwarding huge streams of unstructured data.

For instance, facial recognition and vehicle traffic telemetry (both of which rely on timeliness and low latency) are creating and moving massive amounts of data. But only a small portion of information needs to move up to the core. By processing data at the edge, congestion is greatly reduced. The key information (exceptions and flagged conditions) can move more quickly to the core, having been stripped of the "noise" collected from control and signaling. This process brings more intelligence and value to the application by moving the work closer to data, at the gateway. But to process more at the gateway, networks require more intelligent capabilities.

To take advantage of the huge amounts of data from IT strategies like IoT, carriers need flexible platforms at the edge that not only handle traditional networking but also employ general-purpose compute and storage to execute a diverse range of applications, all within environmental constraints. At the edge, base station equipment may have to fit into Advanced Telecommunications Computing Architecture (ATCA) standards. ATCA is an area where platforms built on a System-on-a-Chip (SoC) like those based on ARM designs have an extreme power and flexibility advantage. In some cases, we may even see very low power compute devices that rely only on power over Ethernet.

Focusing Investment

A carrier's investment chain begins at its network core, extends to an aggregation layer, then out to an access layer/edge, and then finally to the customer device (on-premises IT, mobile device or set-top box). Historically, most of the processing work was being done either at the core or on the customer device; the access and aggregation layers were simply collecting and moving data. Moving the processing both out from the core and in from the customer brings some significant financial advantages.

At the center of the network, investments like core chassis routers (Cisco) and large-scale vertical systems (IBM and Oracle) focused on operations, not differentiation. These are very expensive and proprietary investments with a high cost for compute cycles. Moving processing from the core to the aggregation layer brings better economics and flexibility. By reducing the amount of traffic that needs to be processed at the core, carriers are less likely to "outgrow" these more expensive, less flexible, and lower ROI capital investments, leading to longer utilization and fewer replacement costs.

Capital invested in the edge of the network generally focuses on building out new services and helps bring more competitive differentiation. Moving processing off of customer devices and into the network edge lowers TCO and accelerates deployment. Providers are moving away from customer-based processing, moving the work to the

network edge—where thousands of efficient, low-cost heterogeneous platforms do the virtualized processing and low-cost storage handles the content distribution.

Building Agnostic, Standardized Infrastructures

With investment and processing moving to the edge, how can carriers design the most flexible, innovative solutions? The key is to **move away from the purpose-built platforms of the past and toward the more flexible, yet standard, agnostic solutions of the future.** Because the cloud world is rapidly unfolding with new innovations every day, it is critical that carriers have the ability to leverage these innovations quickly to monetize them. While some believe the key is simply virtualization, in essence, the more critical element is the abstraction of the hardware layer. Moving the point of abstraction further up the stack unlocks more benefits to the carrier. Breaking the connection between the hardware and software stack allows for more flexibility. Software infrastructures like OpenStack and OpenDataPlane further distance the application from the underlying hardware. These open source API frameworks become the equalizer in the middle, bringing consistency to a wide range of applications and operating environments above, while enabling deployment on a wide range of hardware below.

Underneath this stack, hardware should be able to handle a wide range of emerging workloads and also expose the power/performance-optimized discrete compute components and accelerators back up the stack to the software. Thus, an application can take advantage of the technologies underneath, without having to be functionally architected to them. Requiring a specific instruction set or platform below the application or cloud stack reduces the flexibility and innovation. A homogeneous software stack (like OpenStack or OpenDataPlane) residing on heterogeneous hardware throughout the cloud provides carriers with capability and flexibility allowing them to choose more power efficient and cost-effective platforms like those based on ARM SoCs.

Silicon Platform Requirements

Because operating environments won't always be conducive to standard full-size systems, carriers should look to scale out with smaller, more efficient building blocks. This approach puts a greater emphasis on some of the newer form factors and platforms.

Cloud servers can (and should) be heterogeneous; they should use the best tool for the job. Data centers were traditionally homogenous (x86), but carriers come from a world of heterogeneous systems, so they will need a server strategy that supports both x86 and ARM ecosystems. Through open cloud standards, carriers have the ability to select the appropriate platform (ARM or x86) based on specific need. New ARM-based platforms from companies like Altera, AMD, Applied Micro, Avago, Broadcom, Cavium, Freescale, Hi-Silicon, Marvell, Netronome, PMC Sierra, TI, Qualcomm and Xilinx will bring choice to the market, including form factors and designs that are not possible on traditional X86 platforms. Carriers already have extensive experience with ARM in client and mobile devices, so ARM will be seen as a viable server option. Ultimately, vendor diversity will also benefit carriers; it will prevent proprietary lock-in and give them more

leverage to move quickly as market conditions change, so the addition of ARM greatly increases a carrier's leverage in the market.

Network architecture needs to evolve like servers did years ago, from the traditional vertically integrated approach towards workload optimized with a mix of compute, IO and storage capabilities. The ability to select either x86 or ARM, depending on the workload, will be important. Support for OpenDataPlane, OpnNFV, and OpenFlow will be critical, as well as the flexibility to support other open protocols that emerge. Support for networking standards must cover not only the world inside the data center, but it also must reach outside of the data center as well.

A critical platform element will be the ability to pass compute capability details up through the software stack, so that individual applications can use and exploit the capabilities of the hardware below. From a compute standpoint, the SoC should be able to accommodate the correct blend of the following compute elements:

- **CPU** to host a diverse range of applications
- **GPU** for analytics, HPC, and video workloads
- **NPU** for high performance software driven networking
- **VPU** for high performance video encode/transcode applications
- **DSP** for analytic and wireless and video workloads
- **FPGA** for flexible implementations of high performance packet and DSP pipelines, plus CPU offload and acceleration

A truly flexible cloud platform needs to support flexible and open enabling software layers like KVM, OpenDataPlane, OpenDayLight, OpenStack, and Open vSwitch.

Optimizing the Network

Carriers are well into their transition towards a fully virtualized, fully flexible, and fully packet-driven world, where moving compute closer to the users and data is the best technology play. For true flexibility, the underlying platforms must support multiple instruction sets and be adapted for low power. Plus, they must have performance-optimized heterogeneous architectures that can expose their compute elements up the stack.

Uncertainty about new innovations in the market should lead carriers (and their OEMs) towards more flexible solutions that can be both quickly deployed and easily repurposed as needs change. The fluid nature of technology demands both standardization and specialization at the same time, so robust platforms are essential.

Moor Insights & Strategy sees that the best positioned carriers will be moving investment to the edge and aggregation layers as they bring compute out closer to where the data lives, optimizing the network along the way. In this world, support for underlying technology needs to include optimized heterogeneous SoC platforms with enabling software that targets both ARM and x86 in order to have the most truly intelligent flexible cloud.

Important Information About This Brief

Inquiries
Please contact us here if you would like to discuss this report and Moor Insights & Strategy will promptly respond.

Citations
This note or paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

Licensing

Disclosures
Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness or adequacy of such information and shall have no liability for errors, omissions or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.