AT&T Edge Cloud (AEC) - White Paper

AT&T Labs & AT&T Foundry

Table of Contents

Executive Summary

In recent years, there has been a concerted effort among all companies to move their infrastructure to a centralized cloud, enabled by virtualization. This push started with the vision of reducing time to market for new services and achieving lower total cost of ownership (TCO). This surge in the demand for cloud computing led providers like Amazon and Google to build massive centralized clouds (think data centers) designed for efficiency.

With the emergence of new technologies such as augmented and virtual reality, autonomous cars, drones and IOT with smart cities, data is increasingly being produced at the user end of the network. These use cases demand real-time processing and communication between distributed endpoints, creating the need for efficient processing at the network edge.

"Edge computing" is the placement of processing and storage capabilities near the perimeter (i.e., "edge") of a provider's network. Edge computing can be contrasted with the highly-centralized computing resources of cloud service providers and web companies.

Edge computing brings multiple benefits to telecommunications companies:

- reducing backhaul traffic by keeping right content at the edge,
- maintaining Quality of Experience (QoE) to subscribers with edge processing,
- reducing TCO by decomposing and dis-aggregating access functions,
- reducing cost by optimizing the current infrastructure hosted in central offices with low cost edge solutions,
- improving the reliability of the network by distributing content between edge and centralized datacenters,
- creating an opportunity for 3rd party cloud providers to host their edge clouds on the telco real estate.

The computational resources can be distributed geographically in a variety of location types (e.g., central offices, public buildings, customer premises, etc.,) depending on the use case requirements. This variety requires flexibility in the hardware and software design to accommodate constraints around power, space, security, and other elements. Therefore, edge computing needs to support any type of compute, storage and network (e.g., peripherals, uCPE, COTS, white box, etc.)

AT&T's Domain 2.0 (D2) initiative has always followed a hybrid cloud deployment model to keep latency sensitive network and service functions closer to the edges of the network and move non-"real-time" applications to centralized data centers. This type of planning and deployment method is well-known, and has been adopted by most global telco operators.

This model can also be adapted to edge computing and edge-friendly workloads. We will more than double our D2 AIC compute capacity in 2017. These new edge applications need to be geographically close to users, on a typical scale of hundreds to multiple thousands of edge cloud deployments. At that scale, cost-effectiveness is key, along with native support for acceleration and peripherals.

There are multiple edge open source and standard initiatives (e.g., ONAP [i], Open Stack [ii], ONF [vii], CNCF [iii], ETSI MEC [iv], OPNFV [viii], Open Compute Project [ix], xRAN [x], 3gpp[xi], etc.,) that are converging to create an ecosystem that will support edge computing and edge services. The purpose of this whitepaper is to articulate the business and technical benefits of edge computing, describe edge technologies and solutions for suppliers, and identify emerging potential business opportunities with 3rd party cloud providers. Further, we establish a consortium proposal for global commonality such that operators, open source communities and standards bodies can realize next generation applications such as augmented and virtual reality (AR/VR), self-driving vehicles, IOT, and more.

Background – Cloud Computing Evolution and the Rise of the Edge Cloud

Cloud resources have democratized access to computing and dramatically increased the pace of disruptive innovation

Clayton Christensen's concept of the 'innovator's dilemma' has been an axiomatic principle throughout the rise of enterprise and consumer software markets [i]. In his 1997 book, Christensen describes the process by which a novel product or service can initially take root in an underserved fringe market, but then eventually grow to out-perform and disrupt the incumbent players in established markets as well (Figure 1). The incumbents – especially larger, less agile enterprises – are thus faced with the 'dilemma' of allocating sufficient resources to maintain their competitive edge against potential disruptors while simultaneously increasing performance along the dimensions that their mainstream customers have historically valued.[ii] Though applicable to almost any industry, this concept has been incredibly relevant in the rapidly-evolving software and technology space.[i]

Over the last decade, public cloud providers have dramatically increased both the pace and the impact of disruptive innovation by allowing third party services to reap the benefits of geographic presence and economies of scale without deploying and maintaining their own costly infrastructure. Public clouds have contributed to the success of new business models – facilitating the disruption of the television and movie industry, the disruption of the hotel industry, etc. [iii] – and are rapidly attracting established companies as well.[iv]

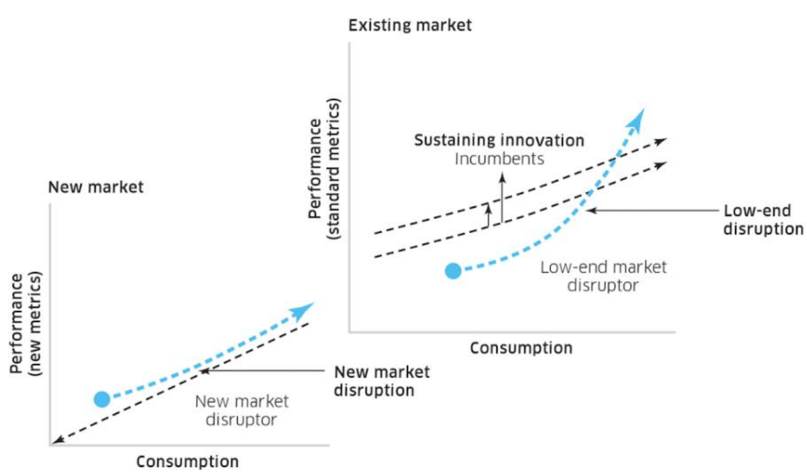

Figure 1: Impact of disruptive innovation upon new and existing markets [iii]

However, the emergence of the public cloud has been revolutionary beyond just the startup and software communities; it has also introduced a new paradigm by which large-scale enterprises can thrive in an increasingly fast-paced, software-driven economy. Public cloud companies have leveraged the sizeable investments made toward their own computing infrastructure to create pay-as-you-go service environments for third-party players.[v] By powering some of the most innovative companies in the world with their Infrastructure as a Service (IaaS) model, these cloud providers are becoming important participants in any new technology trend and effectively sharing in the financial growth and success of each of their customers.[i] And while incumbents in the software and service space are increasingly disrupted, cloud infrastructure players have maintained a steady presence.[v] Thus, a large enterprise can do more than merely avoid being disrupted by new technologies. By leveraging its infrastructure capabilities as a service to enable the growth of emerging ecosystems, it can in fact share in their success.

The public cloud was itself a disruptive initiative. When the first major cloud provider initially launched, their service significantly underperformed the private data centers that were the gold standard for established enterprises. [iv] However, rather than competing for these mainstream customers, they gained traction by designing toward underserved markets, identifying new customers and use cases that were ill-suited for existing infrastructure paradigms. Over time, the offering adapted and grew dramatically, eventually leading even large IT companies relinquish their private data centers. [iv] We believe AT&T must also design toward new, underserved use cases as we evolve our business models and infrastructure into the 5G era.

The IT-Networking Convergence Has Unlocked the Power of the Cloud

As technology has evolved and converged, the flow of information has become streamlined, driving compute and storage off personal devices while simultaneously enabling them to become more powerful and intuitive with each subsequent generation. However, computing revolutions are not a byproduct of the device technology alone – the entire ecosystem and infrastructure must be developed to support these seismic shifts (Figure 2).

In the 1990s, telecommunications companies (telcos), which previously offered primarily dedicated point-to-point data circuits, began to offer Virtual Private Network (VPN) services. Rather than building out physical infrastructure to allow for more users to have their own connections, telcos were now able to provide users with shared access to the same material resources. Because these operators could optimize resource utilization toward overall bandwidth usage efficiency, they could offer the same quality of service at a fraction of the cost. In these early stages, the term "cloud" was used to represent the computing space between the provider and the end user. In 1997, Professor Ramnath Chellapa of Emory University defined cloud computing as the new "paradigm where the boundaries of computing are determined by economic rationale rather than technical limits alone." This has become the basis of what we refer to today when we discuss the concept of cloud resources.[vi]

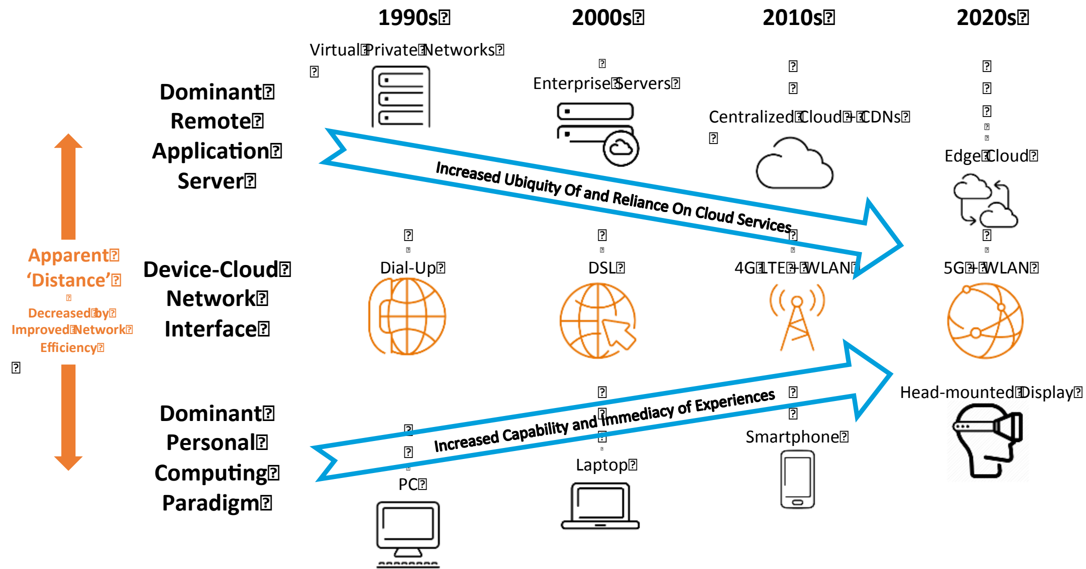

Figure 2: History and future projections of the IT-Networking convergence

As computers became ubiquitous, engineers and technologists explored ways to make large-scale cloud computing power available to more users through time-sharing. They experimented with algorithms to optimize the infrastructure, platforms, and applications to prioritize computing resources and increase efficiency for end users. In the 2000s, major movers in the cloud arena pioneered the concept of delivering enterprise-level applications to end users via the Internet.[vii] Over time, the IT-Networking ecosystem has become increasingly adept and dynamic. Cloud data centers have become more powerful and geographically distributed, and networks have become dramatically more efficient in order to support them. As a result, remote application servers have become easier to access, bringing more information, media content, utility applications, and realistic forms of communication directly to the consumer.

Content Delivery Networks Brought Static Content to the Edge

Content Delivery Networks (CDNs) play a critical role in the seemingly immediate content access that users experience today by establishing Points of Presence (POPs) that store localized caches of content geographically nearer to end users. By hosting these large content files at the edge of the network, CDNs are also able to alleviate performance variance, core network congestion, and high operating expenses for massive data transmission.

Due to the increased streaming demands of the past few years, large content providers have explored models to further improve performance and reduce delivery cost. By establishing commercial agreements directly with local and regional network operators, they can place their own CDNs within the operator network itself, eliminating the delays induced by transferring to an external InP. These large CDN providers are continuously expanding their geographic presence.[iii]

The Emergence of the Compute-Driven Edge Cloud

While CDNs significantly improve user access of static, predefined content such as videos and web pages, they are not designed to be application servers for dynamic content such as AR/VR, real-time analytics, etc. CDNs essentially operate as caches for content files that they pull from their central origin server[vii] and are not equipped with the requisite computing infrastructure to dynamically generate content streams.[viii]

Pseudo-dynamic (sometimes referred to as 'event-driven') content can be distributed with CDN architecture by breaking streams into small segments and caching multiple possible versions of each segment in event-driven lookup tables.[ix] For example, adaptive bitrate streaming is accomplished by encoding several versions of the same video file encoded at a variety of data rates. The CDN can then adjust to network fluctuations by delivering appropriate stream segments based on the inputs such as available bandwidth.

Unfortunately, the complexity of *truly* dynamic applications like interactive content can be exponentially greater due of the degree of variability in the experiences – making the event-driven CDN implementation option costly and impractical. Anticipating and caching content that can be rendered based on every single combination of actions a user could possibly perform would be extremely difficult, if not impossible.

Thus, to address the low-latency requirements of highly dynamic and compute-intensive applications, major cloud providers have been expanding their geographical footprint and positioning their data centers closer to the edges of the network. By investing in increasingly localized resources, they are able to meet the needs of their customers more efficiently, securely and economically.[iii]

However, though cloud provider nodes are becoming increasingly distributed, they are not truly local. Factoring in over-the-air radio communication for mobile networks, and delays induced by queuing and network transfers, even these edge clouds are unlikely to deliver the ultra-low-latency and performance reliability that many classes of emerging applications will demand. Furthermore, the requirements for dynamic content require an even stricter adherence than those for static content given that delivery cannot be easily cushioned with mechanisms such as buffering and adaptive bitrate streaming. A logical step would be for infrastructure architects to mimic the model of CDN evolution and place these edge-computing nodes within the operator network itself.

AT&T Edge Cloud Architecture

Since Domain 2.0's original inception, AT&T continues to shape our network architecture, construction, acquisition, and deployment based on three simple imperatives:

- **Open our network** – Characterized by a modular architecture, highly programmable via robust network API's and policy management.
- **Simplify and scale our network** – Characterized by a common shared infrastructure platform, high degrees of automation achieved using ECOMP[v], and the ability to adapt and perform efficiently as our traffic volume and characteristics change.
- **Increase the value of our network** – Characterized by cost performance leadership, improving time to revenue, reducing cycle times, industry leading security/performance/reliability, all while delivering the best customer experience possible.

In addition to these three original imperatives, AT&T has embarked on an effort to "data-power" our network through an initiative called "Indigo", which layers on sophisticated, timely, and high value analytics for a range of business insights.

By design, access architectures were excluded from the original Domain 2.0 plan given the unique needs and requirements in that domain. With the emergence of 5G, densification of LTE, and the early successes with disaggregated vOLT[1], the time is right to include access (Wireless and Wireline, software and hardware) into Domain 2.0 and is hereto in referred to as Domain 2.0 Virtual Access (D2vA).

AEC is agnostic to any access type (e.g., Wireless, Wireline) and is built to support many services at the edge including D2vA.

Edge-clouds in AT&T's context will largely be modeled as an extension of the principals of network virtualization and software defined networking as applied to data centers through AT&T's Domain 2.0 and Network 3.0 (Indigo) initiatives. These principals allow for elastic

---

[1]) virtual OLT http://opencord.org/wp-content/uploads/2016/03/Virtual-OLT.pdf

infrastructure that can be reconfigured/repurposed on the fly and in accordance with workload demands.

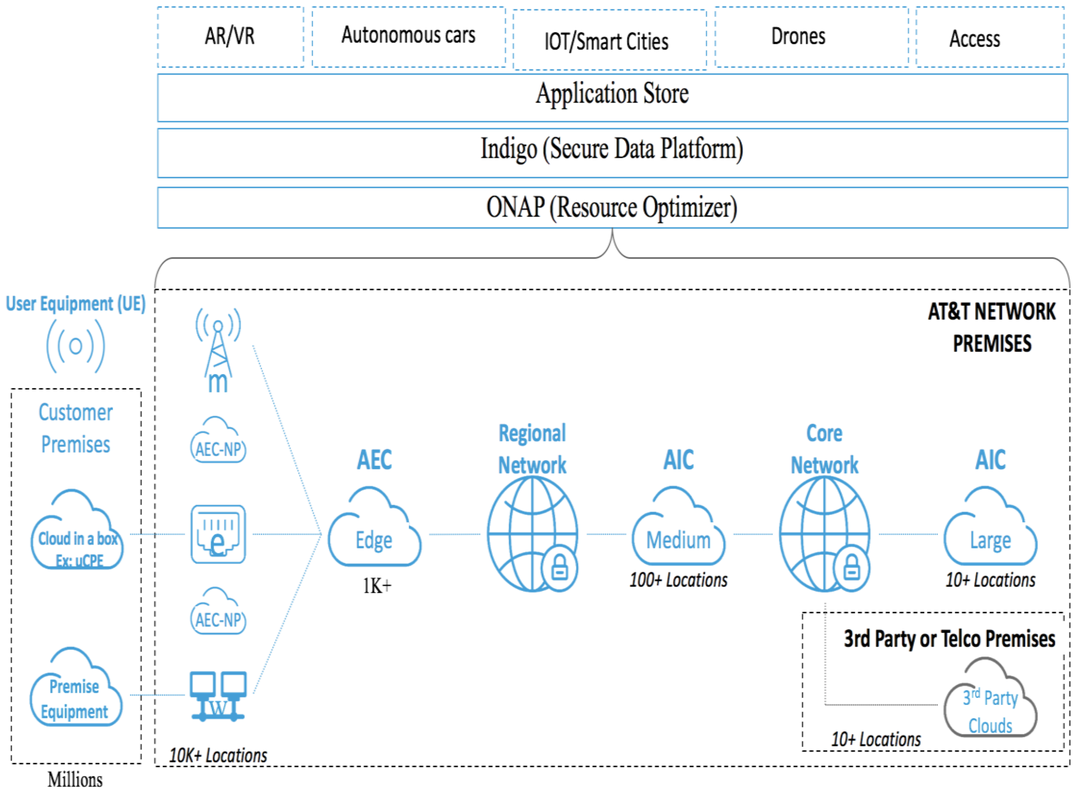

Figure 3 – AT&T Edge Cloud (AEC) Architecture

Figure 3 illustrates the AT&T Edge cloud architecture. The Edge Cloud initiative (AEC) is largely the natural progression of AT&T Integrated Cloud (AIC). AT&T currently has physical network centers located across the United States. Central offices and other physical locations are each equipped with multiple broadband network communication capacities as well as the physical assets of space, power, and HVAC required to place and operate datacenters at the network's edge. Edge computing capabilities could also be deployed in millions of customer premises or access locations such as cell towers.

ONAP is the orchestration mechanism to allow seamless orchestration and workload management at any edge location and datacenter cloud deployment. ONAP is designed to support multiple Virtual Infrastructure Managers (VIMs), orchestrating 3rd party clouds either at the 3rd party location or telco premises.

Indigo enables the possibility of data mining across all of these cloud environments. The intent of this architecture is to enable a common framework for the subscribers and developers in to request resources from the AT&T infrastructure.

Edge Computing Drivers

Edge computing requires more effort and investment than adding a cloud at the selected edge location. Many moving parts must come together. These are the key drivers from a telecommunication perspective that power edge computing:

- **Right content at the edge could reduce backhaul traffic –** Data from the edge is processed at the edge.
- **Maintain quality of experience (QoE)** – The reduction in latency and more efficient utilization of network capacity.
- **Decompose and dis-aggregate access function** – Flexible with modularity and loose coupling of both hardware and software.
- **Better network resiliency** – Ability to deploy clusters between edge and data center allowing for shared restoration of capacity and the potential to improve the reliability due to less transport between the customer and the cloud.

Edge Location Tradeoffs

Optimization of edge cloud site placement can help minimize the total cost, while maintaining applications and edge compute dependent quality of experience (QoE). Latency is not the only factor that drives the choice of edge location. Location evaluation should also consider other constraining factors such as space, cost, compute and storage, complexity and reliability.

Various optimization methods can be used to solve these types of capacity placement problems, including both mathematical and heuristic methods. AEC uses such scientific methods to determine the optimal location for the edge placement.

Edge computing capabilities deployed across multiple providers could connect with each other to via APIs to provide a common set of edge services. Figure 4, illustrates the edge placement at various locations and a common orchestration using ONAP [1].

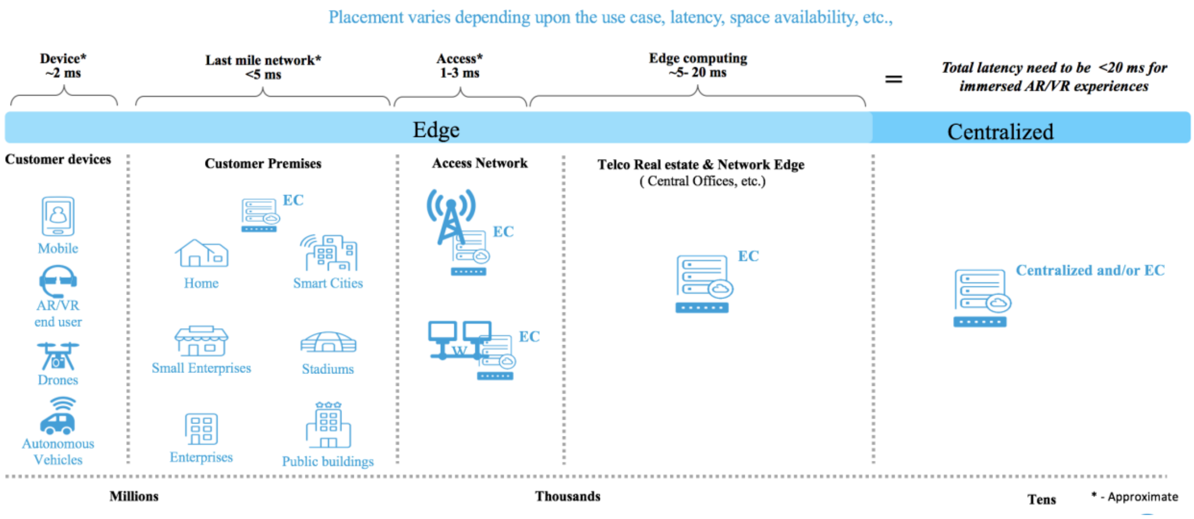

Figure 4 – EC Placement example

Edge Computing Use Cases

Edge computing services are not dependent on radio access technology. The goal is that edge compute should be able to support all access types. While other companies may be focusing on a limited set of services, AT&T is looking to support as broad an array as possible.

Edge computing introduces challenges (e.g., small footprint, large scale, etc.,) as well as new opportunities such as more flexibility for system designers to dis-aggregate computing functionality across platforms and across geographies. It enables developers to build systems that can flexibly distribute computational load across dispersed servers such that compute-intensive functions can occur at physically disjoint locations. This enables new system design choices and new computing paradigms that were not feasible otherwise.

The use cases can be broadly grouped into two categories: Infrastructure and Services.

1. **Infrastructure** – Virtual network functions from multiple access types such as

wireline and wireless

2. **Services –** Services enabled by the virtual network functions hosted in a combined edge and datacenter cloud.

Additionally, by moving compute to the edge, complex computing loads that require real-time machine learning and artificial intelligence capabilities can be enabled on a per-user session basis. Figure 5 is one example of immersed customer experience.

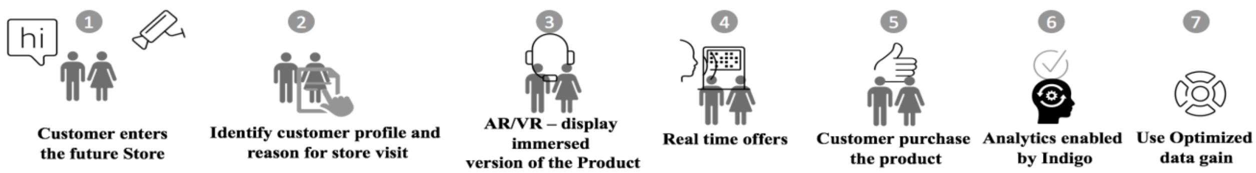

Figure 5 - Example Immersed Customer experience use cases

Co-existence of Centralized Cloud and Edge compute

The need for centralized clouds will remain, but complementary edge computing is required to enable next-generation edge technologies.

Even though edge compute will enable data processing power at the edge of a network instead of the processing at the centralized cloud, not all data needs to be processed at the edge.

As the chosen edge location gets closer to the user, the processing power and storage capacity becomes limited due to space and other environmental constraints. Considering these limitations, it is optimal to store and process necessary data at the edge and remaining data at the Centralized Cloud. Similarly, the data processed at the edge may be required for long-term record keeping and data mining, hence it needs to be transferred and stored in large data centers at the centralized cloud and moved back to edge as needed.

In addition, as the subscriber moves away from one location to another location, access to the application and data has to be managed to ensure a smooth transition and maintain the quality of user experience. If the edge locations do not have the same edge capabilities and a common set of APIs, the subscriber may experience a discontinuity in the service when moving from one location to another.

In order to achieve desired performance and financial benefits of edge computing, a balancing act is required between edge compute and centralized cloud.

The right balance can be achieved by,

1. Supporting the right content at the edge,
2. The right network connectivity between edge to edge and centralized cloud,
3. An open source based developer ecosystem to develop and maintain common set of edge APIs
4. Having all edge providers adopt common sets of APIs and offering seamless edge computing experience to any subscribers
5. Optimizing edge applications to use both edge computing and centralized cloud.

Edge Computing Key Requirements

NFV best practices adopted in the data center cloud need to be continued in the edge cloud to have seamless integration. Edge computing requires substantial effort and investment. Industry wide collaboration across suppliers, other service providers, application/content providers and open source development communities are key to maximizing the benefits of edge computing. AT&T wants to share the set of principles required to speed-up the edge computing implementations.

1. Finite set of configurations with low cost, plug-and-play modular infrastructure to reduce complexity in connectivity, network, compute and peripherals, to be more scalable and lowers costs. Software abstraction based homogeneity that will hide any hardware differences via software
2. Cloud native applications – Design VNF's to be optimized from the beginning to include lighter control planes with simplified security. Cloud control planes need to

be very thin, centralized and secure.

3. Fully automated to utilize near seamless orchestration across edge and centralized cloud including common platform and service orchestration (such as ONAP) that includes automated maturity measurement for operations, designs and services.

4. Zero touch provisioning with autonomous and turn-key solutions for service enablement that enable rapid introduction of services and reduce OPEX

Virtualization Infrastructure Manager (VIM)

Operators cannot afford the luxury of multiple cloud and orchestration stacks between data center deployments versus edge deployments. Therefore, consistent architecture and tool sets are needed between datacenter and edge deployment to reduce design, certification and operational costs. The edge stack needs to be thinner and cost effective compared to traditional data center deployment to support the enormous scale.

AT&T uses open source software throughout the AIC and contributed its ECOMP innovation as open source to the ONAP community. AT&T prefers AIC/AEC architecture that can support plugging any hardware and software, open source, standard or solution. This approach is critical to succeeding in an ever-changing technology landscape. The intent of this section is to outline the functionalities needed in cloud orchestration platform (e.g., Virtualization Infrastructure Manager (VIM)) to open source communities, white box manufacturers and suppliers, so that they may use these requirements to enhance edge capabilities for everyone's benefit.

In addition to datacenter NFV best practices, edge cloud needs to support:

- Both virtual machines and containers
- Micro services
- Network I/O improvements with no overlay such as SR-IOV or simple overlay
- CPU processing offload using GPU and NPU
- Peripheral management (life-cycle)

Edge node could consist of network, compute (including GPU, FPGA, etc.,), storage and peripherals. The size and the configuration of edge nodes vary depending upon capacity demand, use cases served out a location and TCO. Edge node may have to be installed outdoors or in locations with space, security or environmental constraints that differ from those of a data center or central office. The edge architecture needs to have the flexibility to use equipment from different vendors or white box or grey box solutions.

The network is not the only factor that drives the latency. Compute processing at the edge and edge application efficiency also contribute to the overall latency which influences the customer QoE. Optimization of compute processing at the edge is critical to meet the latency demands. Hardware acceleration processors such as GPU, FPGA, NPU, and others could assist in the faster processing, and applications should be in a position to use them. Much innovation is required in this space, particularly to use hardware acceleration cards in the multi-tenant cloud environment.

The edge network needs to support several types of networking connectivity to the network core and non-complex software defined networking [SDN] within the site. Edge nodes needs to be highly secure to avoid hackers tampering with the control plane and core network.

Peripherals are attached to and used with AEC, although they are not an integral part of it (e.g., OLT MAC/PHY for vOLT), and are orchestrated by ECOMP/ONAP. Peripherals are expected to transition from proprietary "black box[2]" to open, commoditized, off-the-shelf "white box[3]" or complete virtualization over time (e.g., VNF). Creating an industry-common functional hardware/software unit as peripheral, with well-established and industry-accepted APIs is the objective of any new peripheral or peripheral type.

Edge nodes need to be flexible to support multiple flavors of servers such as network-optimized, storage-optimized, computing-optimized or general-purpose use, virtual machines, containers, micro services and serverless services. To avoid multiple base server configurations, additional optimizations need to be a plugin to the base configuration.

---

2 A black box is a proprietary device whose workings are not understood by or accessible to its user.

http://searchsoftwarequality.techtarget.com/definition/black-box

3 A white box is an open device made up of commoditized, off the shelf hardware.

https://en.wikipedia.org/wiki/White_box_(computer_hardware

Network optimized servers are suitable for high-performance network functions and could leverage technologies like SR-IOV, Pass-through, DPDK, High-speed NIC (50/100G), NUMA alignment, etc.,

Storage optimized servers could have different disk types to support software-defined storage such as flash, NVME, SSD, etc.

Computing optimized servers could have GPU, NPU, and other acceleration cards to support encryption and decryption.

The platform needs to support both Virtual machines and Containers. All applications should adapt to use Micro services deployed in Virtual Machines and Containers.

Scaling Edge Functions Using Cloud Native Computing

This section provides a quick overview of the influence and potential impact of cloud native computing on container-based network functions at scale. Open source communities and VNF suppliers should adopt cloud native methods to improve the agility of virtual network functions.

The cloud native architecture reference model is comprised of different foundations, namely host, distribution and platform.

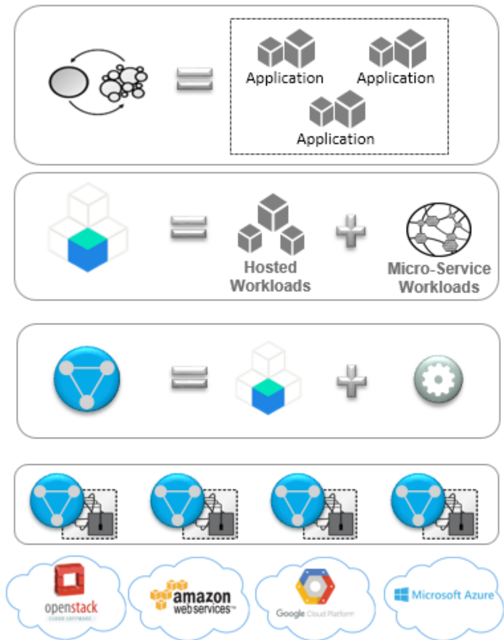

Figure 6 - Cloud Native Architecture

There are some basic characteristics of cloud native applications specified by the CNCF. These characteristics represent the convergence of containerization, cloud infrastructure and a micro service-oriented service architecture:

- Container-packaged, dynamically scheduled and micro services-oriented
- Pivot from scaling infrastructure to scaling function
- Movement away from layers of high-level abstraction to exposing low-level primitives to solve problems
- Shift from persistent physical or virtual infrastructure allocations to short-lifespan atomic functions
- Continued support for applications comprised of tightly coupled software elements deployed as static sets and new support for dynamic services that maintain independence throughout the full software lifecycle

A container image is a lightweight, stand-alone, executable unit of software. It includes everything needed to run it (code, runtime, system tools and libraries). The cloud native ecosystem enables containerized software to run anywhere be it bare metal, OpenStack or in a public cloud. This ecosystem should ensure that the container image is transportable and executed across a diverse set of operating cloud environments dynamically. In addition, it should also provide granular security and service level assurances. Cloud native computing aims to reduce time, cost and complexity without technology, vendor or stack lock-in.

Orchestration & Management

Enhanced Control, Orchestration, Management & Policy (ECOMP) delivers product/service independent capabilities for the design, creation and lifecycle management of the D2 environment for carrier-scale, real-time workloads. Open Network Automation Platform (ONAP) is the open source version of ECOMP and the broader community is working to enhance it further to meet the needs of providers and enterprises.

ECOMP/ONAP capabilities are the natural fit for the effective scaling, orchestration & automated management of the service chains needed to realize the edge compute and next generation services in such a distributed environment. The open nature of the ONAP evolution provides new opportunities for extending the traditionally vendor-specific environments in today's edge networks with dedicated and often under-utilized resources to a multi-vendor, scalable edge environment that allows for flexible scaling of resources where needed.

Open Source Eco-system – Edge Computing enablers

There are multiple, complementary open source initiatives that are converging to create an ecosystem that will support edge computing and edge services. AT&T believes in the convergence of open source communities to achieve the goals together without duplication of effort as time, resources and money are limited within the industry. Below is the example list of open source communities playing a role in the edge computing space that AT&T is actively engaged with. AT&T would like to see these communities come together and develop broader edge computing features to fully harness the benefits of shifting functionality to the edge.

Open source and standards communities currently working on edge spaces such as ONAP, Open Stack, ONF, CNCF, ETSI MEC, OPNFV, Open Compute Project, xRAN, 3gpp should work together in enabling the edge functionalities.

Conclusion

AT&T seeks to engage in a transformation of network technology, operations, infrastructure, software, and APIs in a way that provides greater value to customers and application development partners. The result will be an edge cloud-centric networked capability that can support varied edge workloads along with flexible business models. Comments and engagement are welcomed.

Glossary

Acronym	Description
AIC	AT&T Integrated Cloud – AT&T's cloud enabled using Open Source
AEC	AT&T Edge Cloud – AT&T's new initiative part of Domain 2.0 program and an extension of AIC
AEC-NP	AEC with Network Peripherals
AR/VR	Augmented Reality and Virtual Reality
CO	Central Office
COTS	Commercial Off the Shelf
CDN	Content Delivery Network
D2.0	Domain 2.0 – AT&T's initiative
D2vA	Domain 2.0 Virtual Access – AT&T's Initiative to deploy access in to the AEC/AIC

DPDK	Data Plane Development Kit
ECOMP	Enhanced Control, Orchestration, Management & Policy
EVPN	Ethernet VPN
FPGA	Field Programmable Gate Array
GPON	Gigabit Passive Optical Networks
GPU	Graphics Processing Unit
HMD	Head Mounted Display
IOT	Internet of Things
LTE	Long-Term Evolution
MAC	Media Access Control
MEC	Mobile Edge Computing
MPLS	Multi-Protocol Label Switching
NFV	Network Function Virtualization
NIC	Network Interface Controller
NPU	Network Processing Unit
NVME	Non-Volatile Memory Express
NUMA	Non-uniform Memory Access
OLT	Optical Link Terminal
ONAP	Open Network Automation Platform
PHY	Physical Layer
Peripherals	Specialized hardware devices (e.g., OLT MAC/PHY for vOLT)
PNF	Physical Network Functions
QoE	Quality of Experience
SSD	Solid State Drive
SR-IOV	Single-Root Input/output Virtualization
TCO	Total Cost of Ownership
uCPE	Universal Customer Premises Equipment
VIM	Virtual Infrastructure Manager
VNF	Virtual Network Functions
vOLT	Virtual OLT
VxLAN	Virtual Extensible LAN

References

[i] https://www.onap.org/

[ii] https://www.openstack.org/

[iii] https://www.cncf.io/

[iv] http://www.etsi.org/technologies-clusters/technologies/multi-access-edge-computing

[v] https://about.att.com/content/dam/snrdocs/ecomp.pdf

[vi] https://www.gsma.com/

[vii] https://www.opennetworking.org/

[viii] https://www.opnfv.org/

[ix] http://www.opencompute.org/

[x] http://www.xran.org/

[xi] http://www.3gpp.org/

References for Sections Background Cloud Computing Evolution and the Rise of the Edge Cloud and AT&T Edge Cloud Architecture:

[i] Jesus Rodriguez. “The AWS Tax and the Innovator’s Dilemma”. 2 Jun 2016. Medium.com.

[ii] Christensen, Clayton M. “The Innovator’s Dilemma”. 1997.

[iii] Weldon, Marcus K. “The Future X Network: A Bell Labs Perspective”. 2016.

[iv] Julie Bort. “Netflix, Juniper, and Intuit explain how Amazon is eating the $3.5 trillion IT Industry”. Business Insider. 13 Jan 2016.

[v] Matt Weinberger. “The Cloud Wars explained: why nobody can catch up to Amazon”. Business Insider. 7 November 2015

[vi] “The History of Cloud computing”. http://www.eci.com/cloudforum/cloud-computing-history.html. Accessed 30 Nov 2016.

[vii] Akamai Developer. "{OPEN}ing Akamai – an introduction". https://developer.akamai.com/learn/index.html. Accessed on 15 December 2016.

[viii] Interview with Mike Hodges, CEO of Cinova. 29 November 2016.

[ix] Beheshti, Hooman. "Leveraging your CDN to cache "uncacheable" content". Fastly. 5 May 2015.