STRATEGY ANALYTICS
Research, Experts, and Analytics

Service Providers

AI + Edge Computing Essential in the 5G Era

Service Provider Strategies (SPS)

While 5G could be deployed without Artificial Intelligence (AI) and edge compute, intelligent edge will be essential to new revenues and lower costs for 5G. Strategy Analytics predicts that these two technologies combined will be 'table stakes' to deliver on the promise of 5G.

Combining the benefits of AI with compute power at the network edge offers compelling value propositions, with benefits for both network operations and new service opportunities, improving the return on investment (ROI)

Service Providers

Sponsored by intel

June 2018

Susan Welsh de Grimaldo
Tel: +1 617 614 0724
email: *swelshdegrimaldo@strategyanalytics.com*

Introduction: Pushing Artificial Intelligence (AI) to the Edge Key for 5G

Intelligent edge will be critically important for 5G. While 5G could be deployed without artificial intelligence and edge compute, Strategy Analytics predicts that these two technologies combined will be 'table stakes' to deliver on the promise of 5G.

- **Operators see a role for AI** to both improve operations and create new service opportunities. As networks get more complex and serve many more endpoints with 5G, AI becomes more essential.
- **Compute power at the network edge can simultaneously** impact network efficiency and improve service experience for both consumer and enterprise/industrial use cases. One of the biggest benefits of edge compute is lower latency, which is critical for numerous 5G use cases. However, other values exist such as 'geo fencing' the data, higher security, control of backhaul, and new innovations due to the values edge compute brings.

Service Providers who combine the benefits of AI with compute power at the network edge will be able to offer compelling value propositions that enhance the return on investment (ROI) of 5G.

Defining AI and Edge for the New Network Paradigm

Both AI and edge compute are terms that have been used in multiple confusing ways. AI has been used as an umbrella term to include multiple topics from machine learning (ML) and data analytics to true automation, and covers both network centric and user/service centric use cases. In the context of the new network paradigm based on virtualization enabled by Software Defined Networking (SDN) and Network Function Virtualization (NFV), there is an opportunity to push various functions—including AI processing—out to the edge of the network in a distributed architecture. But when we say 'pushing AI to the edge'— what edge are we referring to? Intel defines edge computing as applications, data, and services located at the edge of a network rather than in a centralized datacenter. With edge computing, it is feasible to push flavors of AI and functions that formerly required big data centers into micro data centers (µDC) and AI processors closer to the user—-at the originating device, at the enterprise or in an operator's local access network and even in the RAN. **Importantly, who owns the edge and who will offer AI at the edge?**

Since there are multiple edges, AI processing can be done wherever it is most efficient for any specific application, service or network operation. AI in one edge may even support an AI process at another:

- **Device as edge –** one edge is an end user or IoT device that generates and consumes data and will likely have some form of embedded AI. Strategy Analytics predicts that in 5 years nearly every smartphone chip will be an AI chip and that use cases will expand beyond today's focus on 3D face detection, image recognition and Animoji (see "Smartphone Apps Processor Forecast (2008-23): On-Device Artificial Intelligence (AI) Set to Take 25 Percent Share in 2018"). These smart devices may send and receive data to and from additional AI systems that have more compute power than the device itself or that can aggregate data from many such devices for analysis and action.
- **Enterprise premise network edge** – this edge can support AI processing running at any location in an enterprise (or government) facility, or even on a piece of hardware in a vehicle, drone or machinery, and will collect and process data from smart devices as they consume or generate data e.g. video surveillance cameras or industrial robots or other devices with sensors. Local AI may need connectivity to update ML models from the

cloud, but may not rely on constant connectivity if the local AI processor can support the work load and services needed.

- **Operator network edge**– this edge, often referred to as multi-access edge compute (MEC,) may be located at a micro data center sitting at a radio tower, edge router, base station, Central Office or Internet gateway with compute power and an AI stack/platform to host applications and services for both the operator itself and its partners who utilize the AI capabilities.

Intel clearly sees that "one of the challenges this new networking paradigm creates stems from the fact that the edge of the network are consistently shifting and moving," explains Caroline Chan, VP and GM, 5G Infrastructure Division, Network Platform Group, Intel.

Vision: Why AI + MEC will Deliver on the Promise of 5G

Both AI and edge compute are essential for 5G to architect and manage the network in ways that enable cost efficiencies, enhance performance and deliver next generation services.

Ubiquitous connectivity, big data and AI forms a virtuous cycle for next generation connected computing: ubiquitous connectivity enables connected devices of all form factors to generate more and more data. The huge volume of data can help improve the performance of AI, in particular the Deep Learning model. Improved AI drives deeper user engagement, and will in turn generate even more valuable data.

However, in the conventional cloud/client model, the link between the centralized cloud and the client has increasingly become a bottleneck, as the virtuous cycle continues to drive ever more data and increasingly lower latency services and applications.

Some of the main benefits that AI + MEC together can deliver for 5G are: low latency for real-time services, enhanced security and backhaul cost savings (by not sending all the data up to AI in the cloud). Most importantly operators can provide a platform for innovation and open edge services to partners and developers to create applications that support consumer, enterprise and multiple verticals while adding significant value to their business.

The vision for MEC with AI can be divided into two general categories: benefits for network operations and enablement of new services.

Network Operations: Enable automation needed for 5G to scale efficiently, handle traffic, etc.

5G will be expensive to scale without AI for automation across operations, including traffic optimization, network resource management, predictive analytics for improved customer experience and cost savings. Intel explains:

> "In network automation/optimization, AI will… (be) critical at the edge and inside devices. Because many times, the decisions have to be made real-time, for example choosing the best data rate, choosing the best AP/network. Also edge/device will have rich measurements/other data for optimal local decision making instead of sending the data to a remote cloud. Distributed intelligence at the edge/device can be a real differentiation for those scenarios."

Strategy Analytics notes that some of the internal network operations of CSPs that will particularly benefit from AI at the network edge include:

- Characterizing traffic in real-time for Wi-Fi offload or 'Local Breakout' to nearest fixed broadband connection
- Intelligent traffic redirection across the RAN or Metro Area Network to the nearest processor, storage or database resources including:
    - Video Cache/Media Servers
    - IoT Application Service Platforms
    - Emergency /Public Safety Service Access
    - mHealth Services
- RAN specific Network Slice Management e.g. for Peer-to-Peer (P2P) or Device to Device (D2D) traffic
- Zero-touch service management in the standalone (SA) phase of 5G with service-based architecture (SBA)

5G will need to dynamically configure and scale many types of devices and applications as needed. Intel explains:

> "For any given application and device type, we need to provide an end-to-end network service that meets the quality of experience expectation. The scale and dynamics of such system makes it very challenging to fulfill the E2E service expectation with handcrafted algorithms. AI will be the key to automate/optimize E2E service provisioning in such a complex network by taking inputs from end-to-end and across protocol layers. The SDN/NFV transformation of the network gives the flexibility to better configure/provision the network. But how to configure/provision/optimize the network will rely on AI powered solutions."

The efficiency of doing as much intelligent processing as possible at the edge with AI can dramatically reduce the cost of 5G deployment and service delivery on an ongoing basis.

New Services: Enable revenue streams and new business models

The capabilities and value proposition of AI at the edge with 5G connectivity will enable innovative approaches to meet low latency/real-time service needs—something not previously available with mobility—thus creating a new playground for innovation in services. If operators bring compute and microservice functionalities that are now 'cloud-native' closer to the user and open up edge processing and AI capabilities to third party partners and developers they will create opportunities for innovation and new business models. Intel guides that:

> "the powerful data processing and analytics capabilities that have traditionally lived in the heart of the data center must be **strategically placed closer-and-closer to the data generating and consuming endpoints**, at the "edge." By **expanding the powerful capabilities of the data center outward**, service and network providers can deliver more powerful services, reduce application latency by processing more data closer to the edge, and optimize TCO."

Two examples of services requiring high bandwidth, low latency communications (5G) and real time intelligence (MEC):

1) Intelligent Manufacturing: high precision manufacturing and robotics will require AI located on premises to ensure real-time responsiveness. This may also support increased security and privacy.

2) AR/VR: future immersive and interactive services such as AR and VR may demand both high bandwidth and low latency. In order to satisfy this highly demanding user experience, both content and artificial intelligence will need to be pushed to the edge of the network.

AI at the network edge can significantly benefit consumer, enterprise and partner services, through improved user experience, enhanced security, and cloud functionality without the overhead cost and added latency of moving the traffic to a central location for processing. For example, AI at edge can avoid several of the security and privacy issues in a public cloud. Specific use cases can rely on AI at the edge, in particular the ability for Communications Service Providers (CSPs) or their partners to offer "as-a-service" models, with AI-as-a-Service sitting at network edge for enterprise and industrial use cases. By placing compute with AI the network edge, those services that require AI and security become more accessible to a broader range of enterprises to help digitally transform their businesses and enhance their offerings. Similarly, services that need to have real-time decision-making based on a range of sensor, video, or other data inputs can leverage AI at the edge with deep learning capabilities (see Intel blog, https://ai.intel.com/artificial-intelligence-at-the-edge/ for additional service examples).

Dynamic Network Slicing Will Enhance Efficient 5G

As mentioned above, AI and MEC combine to support network operations and enable services. As operators move toward 5G and adopt network slicing to deliver differentiated classes of service for enterprise and vertical markets, AI at the edge can play an important role to ensure efficient use of resources and offer AI services for partners and small and medium enterprises (SMEs).

In particular, AI at the edge will be essential for managing Dynamic Network Slicing so that slices become a mechanism for cost effective use of network resources that scale up and down to meet the needs of enhanced 5G service offerings across many different types of services and connected devices. For more on network slicing, see Strategy Analytics' report: "Network Slicing the Key that Unlocks 5G Revenue Potential – Where 5G Meets SDN/NFV".

How to go from Vision to Reality: Addressing Challenges

The vision of the value propositions of AI and MEC make it clear that they may be highly beneficial to 5G success. However, there are a number of challenges to be addressed to get from that vision to the reality of AI + MEC deployments at large scale as commercial 5G networks are built out over the next decade.

Technical challenges, including architecture and standards

AI Architecture: CSPs that are developing AI capabilities are debating whether do a single centralized 'brain' or optimized point solutions for different business functions (see TM Forum report, "AI: The Time Is Now," December 2017). With AI distributed to the network edge, there is an additional set of options that raises a new challenge of how to federate and update multiple AI/ML models at diverse edge points. Intel suggests "we need to look at new AI architecture such as Federated Learning and optimized hardware implementation to enable intelligence at edge/device."

Additional operational challenges also arise including:

- how to get good quality data sets to train AI/ML/deep learning
- how to develop the right expertise and hire AI and analytics skills/employees with CSP domain knowledge as other industries, digital giants and digital native start-ups compete for the same talents
- how to avoid 'AI Hype', prioritize and keep up with growth of AI explosion

Open Source: As operators and solutions providers seek to address technical challenges, the role for open source may offer a way to spur faster, more cost effective development and innovation. For example, AT&T and Tech Mahindra collaborated to create an open source AI platform, Acumos, hosted by the Linux Foundation. Intel has also announced open sourcing components of the Wind River Titanium Cloud for edge computing:

> "Titanium Cloud brings mature technology to the open source community which is optimized for edge computing. Wind River Titanium Cloud technology is built on collaboration with open source communities and by opening these Titanium Cloud components, we are now able to upstream our enhancements and make these technologies available for the whole ecosystem for the first time. The Intel NEV SDK contribution is a set of reference libraries and APIs to enable edge computing solutions for different network deployment scenarios. The SDK is designed to remove the need for application developers to understand the complexity and implementation attributes of the underlying network protocols. The Intel NEV SDK libraries are validated and deployed with Wind River Titanium Cloud technology." (Caroline Chan, Wireless Symposium Keynote, May 2018)

Standards: Multiple standards bodies and working groups are addressing challenges and standards development for AI and MEC to insure interoperability, including ETSI's Experiential Networked Intelligence Industry Specification Group (ENI ISG), which "is defining a Cognitive Network Management architecture, using Artificial Intelligence (AI) techniques and context-aware policies to adjust offered services based on changes in user needs, environmental conditions and business goals." Telecom Infra Project (TIP) has both the TIP Artificial Intelligence and Machine Learning Project Group and the TIP Edge Computing Working Group. Leading providers of AI and MEC solutions are participating to drive standards development, along with CSPs and other ecosystem players:

> "Intel believes that open standards organizations foster industry-wide collaboration and innovation on standardizing technologies that propel the industry forward. Intel is a member of the Partnership on AI, which is focused on rallying the industry around a set of standards for AI that promise to ultimately bring down costs and make AI more accessible to more people. Intel is a leading contributor to a number of open standard organizations working in the NFV and SDN area. Intel's involvement in these open standards organizations includes ETSI Mobile Edge Computing (MEC), ETSI Network Functions Virtualization (NFV), Internet Engineering Task Force (IETF) and the Open Networking Foundation (ONF)." (See: https://newsroom.intel.com/news/intel-joins-partnership-on-ai/ and https://networkbuilders.intel.com/network-technologies/openstandards)

Financial and business model, including cost structures/ROI and how to work with partners and developers

For each of the edges, there are questions of who owns the edge, and who owns the AI/compute at the edge—real estate owner/tower company/operator/partner/enterprise etc. —or has access to it as a partner, as well as who benefits from value created at each edge, and what are the value propositions for that edge. Multiple edge processes with AI may work together and also work with AI in the cloud or centralized data center and serve multiple use cases with different business models. CSPs will need to evaluate the cost and investment ownership

in parallel with the revenue and business models and who earns what share of the value to determine ROI—and it is likely to be a complex analysis given the range of use cases that may be supported over time. For example:

- An intelligent edge at an enterprise location has the ability to do AI without the need for constant connectivity to the cloud. In this case it will be hard for CSPs to extract value unless they offer additional capabilities such as cloud based analytics and on-demand bandwidth to the enterprise.
- Network edge with AI at a base station can serve operational functions in the network for the CSP, while also supporting services for end users and 'AI-as-a-Service' offerings that can be simultaneously opened up to third party partners and developers to enable services that access the intelligence of the network edge AI platform. In this case the operator may either own and operate the intelligent edge or utilize it in an outsourced revenue share model e.g., with a tower company or a building owner who finances the compute power and operates the facility or even with an AI datacenter provider who is moving to offer micro data centers at the intelligent edge.

CSPs need to think through their edge role depending where the edge sits and to promote the benefits of an operator-network based edge with 5G as they work with end user customers and service partners to develop business models and solutions.

Public policy

To avoid regulatory problems and ethical issues, AI services need clear government policies that stimulate positive uses of AI while protecting citizens from potential malevolent uses. In addition, these AI policies may need to address the economic and societal transition to increased use of AI based automation i.e. skills development, job transition, private/public partnerships, etc. Industry leaders like Intel are actively discussing AI policy frameworks in multiple markets with government leaders and agencies:

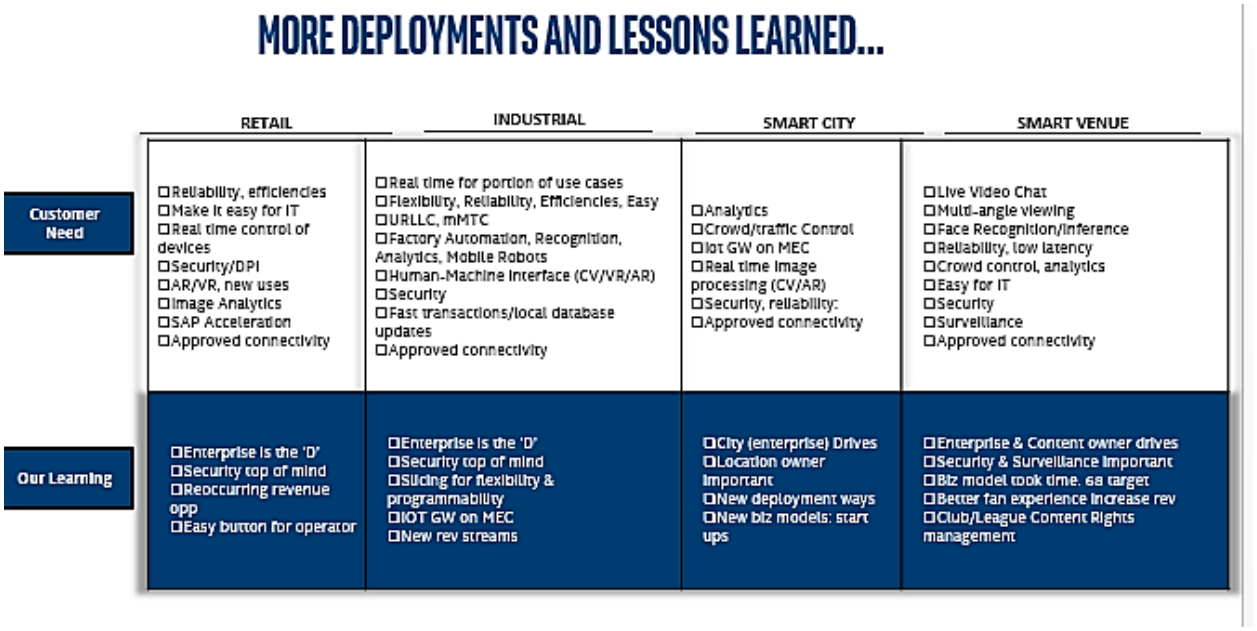
> "With all this projected good coming from AI, it's important that we also address unintended consequences. We believe the government should help society transition to a safe and innovative yet ethical AI future… by preparing an AI workforce, dealing with workers displaced by AI, ensuring transparency around the use of AI, protecting privacy, and at the same time encouraging AI innovation to drive positive social benefits. Intel is engaging with governments now to make sure the public policy environment's rules ... fit the changed game of AI innovation." (See Intel blog: https://blogs.intel.com/policy/2017/10/18/naveen-rao-announces-intel-ai-public-policy/)

Early Enterprise Edge Learnings—Things to Consider

Based on numerous trials and deployments that it has conducted with enterprise verticals, Intel is sharing key learnings that it found to be common across customers, and guides that these findings are important to ensure adoption of edge based vertical market solutions. Specifically:

1) **Enterprise must drive deployment:** The technical solution must solve a specific business problem or add to output.
2) **Ease of adoption is a must:** New technology needs to accelerate time to market for both enterprise and operators.
3) **Location is important:** For some industry/enterprise use cases, having local, on premise solutions is critical to success.
4) **Security and reliability:** Both are important to end customers as they integrate edge services with AI.

Intel has seen a lot of early traction across multiple verticals, such as Retail, Industrial, Smart City and Smart Venues – the table below from Intel is indicative of the customer needs and Intel's learnings:



Source: Intel

Implications: Call to action

AI with MEC may be one of the most essential architectural evolutions of 5G; without AI combined with MEC, 5G may not live up to its full promise and the costs to deliver on evolving customer needs could be prohibitive.

CSPs and their ecosystem partners need to push forward to deliver AI at the network edge in order to:

- Scale 5G and deliver on the breadth of use cases and large number of devices in 5G, by supporting automation in 5G networks from end-to-end
- Drive innovation and enable new services that need real-time intelligence capabilities with low latency
- Enable new business models with partners and developers
- Reduce OPEX of 5G networks and services
- Support the demands for security and analytics with AI

CSPs, enterprises/verticals and their solution partners need to collaborate to create the broader wireless ecosystem and move forward both AI standards and the development of new business models. The architecture for MEC and AI needs to evolve with a dual focus that will reduce 5G deployment and operational costs and enable a flood of high value new services and partnerships.

How Can Strategy Analytics Help You?

Strategy Analytics provides strategic and tactical support to global clients through a range of customized solutions:

- Multi-Country primary research assignments using leading edge tools and techniques
- User Experience Design and Innovation engagements
- Real Time Mobile Consumer On-Device Tracking projects
- B2B Consulting Projects and Whitepapers

Please contact us at Custom@strategyanalytics.com with any questions and for further details and solutions on how we can work with you to create a custom solution that addresses your specific needs.