Solution Brief

CxO, Engineering VP, Network Operation VP
Telecommunications

Capgemini Pushes AI Learning to the Edge with Advanced Privacy Preservation

With federated learning, network edge clients learn a shared prediction model while keeping all training data at the edge. Capgemini and Intel deliver federated learning-enabled "Edge AI" for communications service providers

The transformational promise of artificial intelligence (AI) and machine learning (ML) for communications service providers (CoSPs) presents opportunities to advance network performance, conduct predictive analytics, reduce operational costs, generate new revenue, and improve customer experience.

The adoption of AI in smart manufacturing, Internet of Things (IoT), augmented reality/virtual reality (AR/VR) and smart city connectivity applications means billions of devices are generating data at the network edge. Traditional AI training involves importing massive amounts of data to a centralized location where machines are taught to properly interpret data and continually learn based on changing data inputs.

But massive data movement from the edge comes with high cost, bandwidth constraints, and legal and privacy concerns with regulations like the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) laws in the US and European Union.

Capgemini*, an Intel® Network Builders ecosystem member, and Intel are working together to introduce CoSPs to federated learning (FL), a collaborative AI training methodology that conducts training at the network edge without moving or exposing sensitive data reserved for centralized AI training. The net result of FL includes increased data diversity, preserved privacy, and reduced latency while achieving learning effectiveness.

Security and Privacy AI Issues

User growth, spectrum diversity and increased complexity are driving increased AI use for CoSPs to scale operations beyond human workforce capacity, and for networks to self-sustain wherever possible.

Network AI promises to self-configure, monitor, manage, and correct system and network issues without human intervention. Network AI drives reduced costs, improved network quality and customer experiences, and creates new revenue opportunities in verticals using IoT, AR/VR, and robotics.

For network AI to be effective, it requires significantly large and diverse volumes of data from which to learn. Traditionally, this has meant capturing and centralizing data from various sources and transporting them to a centralized node, either at a data center or in the cloud (Figure 1).

While centralized AI learning produces good model accuracy, it has significant issues that limit its scalability and effectiveness due to lack of compliance with privacy regulations. When massive data collection and transfers back to a central data center or cloud are required to create robust learning, cost and network bandwidth availability limit the diversity of available data.

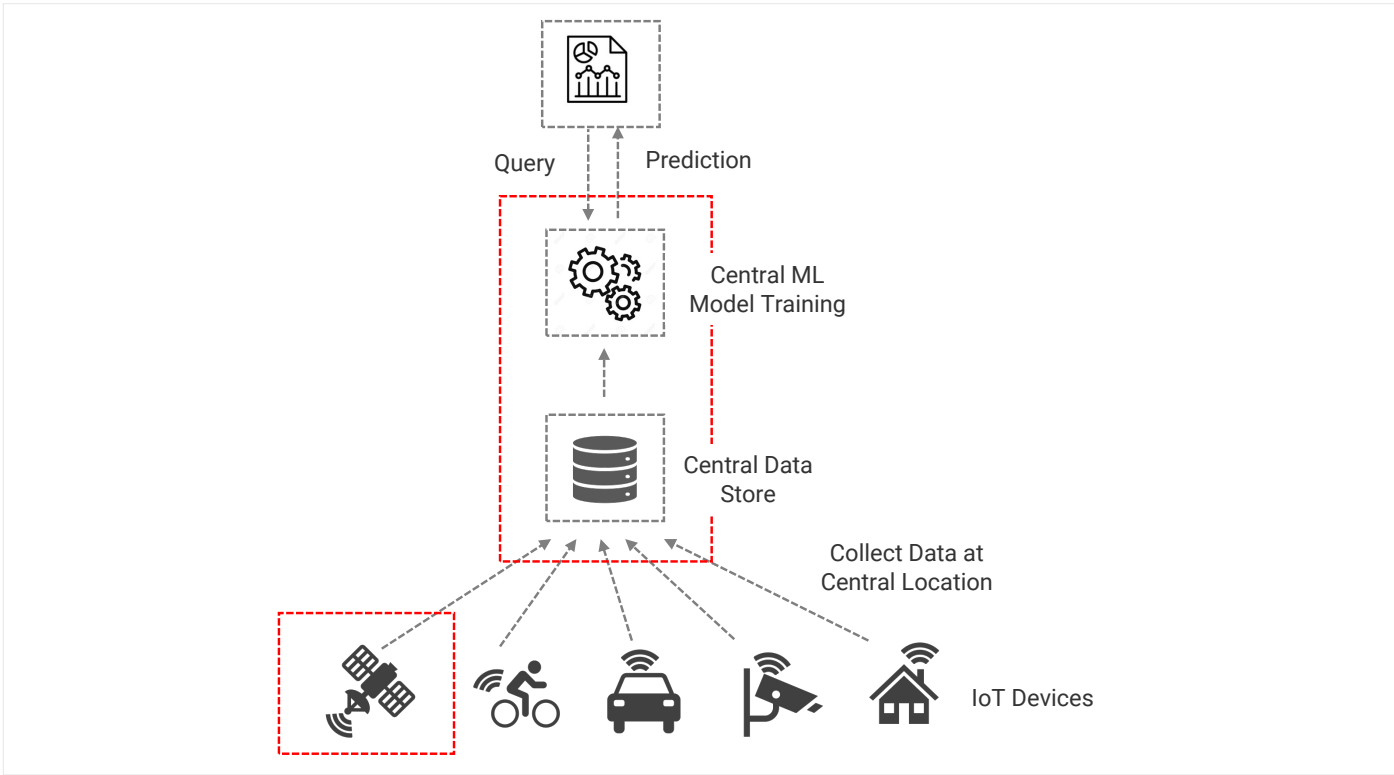

Figure 1. Centralized Learning Model for AI (Source: Capgemini)

Some countries have laws that restrict using personally identifiable information (PII) or healthcare-related data for AI training. Scarcity of sensitive data in these cases throttles the effectiveness and importance of AI's impact. This results in AI training models that do not match the needs and opportunities of an entire market, which is critical for CoSPs.

In many cases, this lack of data volume and diversity can cause biased AI systems and underrepresent key audiences. For example, doctors starting to use AI diagnostically are finding it difficult to represent all communities in the AI learning process because only a limited number of medical devices are able to collect the needed training data quantity and quality, as well as afford the costs associated with data capture, storage, and transfer to AI learning engines. This lack of needed data has resulted in training models having incomplete data across the market[1].

Federated Learning: A Revolutionary AI Paradigm

To overcome the challenges with centralized AI training, the concept of Federated Learning (FL) was developed. FL enables multiple edge nodes or devices to collaboratively learn from a shared prediction model without having to share sensitive data or classified/private information. Minimal data movement across the federated learning environment includes only the model training parameters and iterative updates. Specific individualized data collected by edge devices is not transmitted or shared. The use case which launched FL generated keyboard predictions using mobile phone texting to train models across users without needing to collect user data.

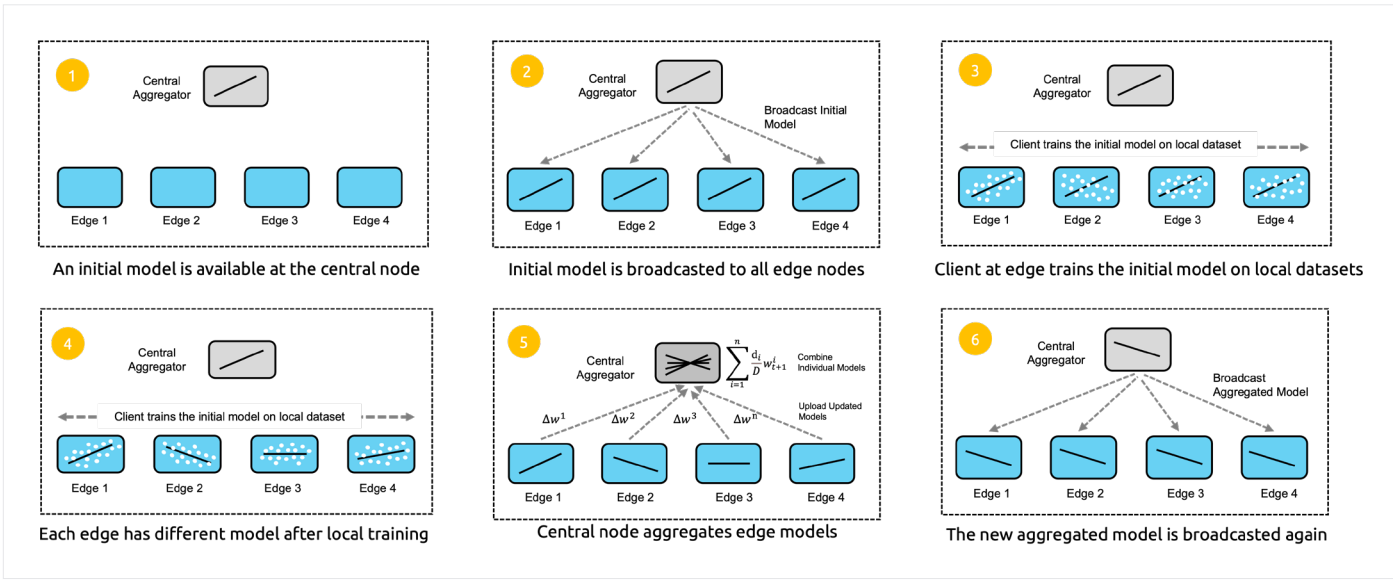

Figure 2. Federated Learning AI Model Training Goes Through These Six Steps

An FL framework used in model training includes a central aggregator and several edge clients. The central aggregator broadcasts one initial model to each edge client where data is generated. The model trains using local data at the edge whereby the edge clients train the first model iteration using the model's local datasets. Trained edge models then go back to the central aggregator and an updated newly trained model is created via a weighted learning algorithm. The new, more robust model is rebroadcast, iteratively updating edge clients with new learning (Figure 2). As the FL model is locally trained at edge, so data is not exposed to additional security threats. FL also increases diversity of data inputs, and minimally impacts network bandwidth.

Federated Network AI Benefits and Top Use Cases

CoSP Benefits

- Enhanced data privacy and security
- Increased scalability and adaptability
- Faster and continuous model training
- Increased data diversity and model generality
- Reduced OpEx (operating expenditure)
- Enhanced customer experience
- Lower latency
- Less power consumption

Top CoSP Use Cases

- Predictive maintenance
- Root cause failure analysis
- Intelligent service rollout
- Intelligent service orchestration
- Network slice management
- Subscriber churn prediction
- Network intrusion detection
- Intelligent capacity planning

Compared to traditional machine learning (ML), federated learning poses several challenges when used for telecommunication and IoT networks[2].

1. Unpredictable training participants. Mobile and industrial devices may experience intermittent connectivity. There may be thousands of devices, and the communication with these devices may be asynchronous. The training coordinator needs a way to discover and communicate with these devices.
2. Asynchronous communication by design. As previously noted, devices like IoT sensors may use communication protocols like message queuing telemetry transport (MQTT) and rely on IoT frameworks with asynchronous pub/sub messaging. For example, the training coordinator cannot simply make a synchronous call to a device to send updated weights.
3. Passing large messages. Model weights may be several MB in size, making them too large for typical MQTT protocol payloads.This can be problematic when only low-bandwidth network connections are available.
4. Emerging frameworks. Deep learning frameworks like TensorFlow and PyTorch do not yet fully support FL. Each framework includes emerging options for edge-oriented ML, but these new options are not yet production-ready and are adequate only for simulation work.

Capgemini's NetAnticipate FL Framework

Capgemini's NetAnticipate framework (Figure 3) is a highly scalable, cloud-native, self-learning AI platform for realizing autonomous network operation. NetAnticipate analyzes a substantial number of hidden and hierarchical influencers to predict potential network anomalies, builds autonomous decisions and takes preventive actions. An autonomous feedback loop ensures the network self-learns over time to improves the actions it takes over time. NetAnticipate uses best in breed AI deep-learning algorithms for creating self-learning closed loop automation. It orchestrates various deep learning algorithms for identifying network anomalies in real-time and forecasting future anomalies using multi-variate timeseries analytics.

NetAnticipate then prescribes preventive actions to fix the anomaly before it can start affecting the network. Actions taken are improved over time through deep reinforcement learning techniques. Reinforcement learning uses paradigms for sequential decision making under uncertainty, solely through the rewards and penalties generated from previous actions it has performed in the network.

The NetAnticipate AI platform is designed to deliver critical value to CoSPs through:

- Enhanced end-user experience and reduced customer churn through proactive network issue detection and meeting dynamic customer needs
- Higher operational efficiency by simplifying network operations and facilitating a consistent, error-free network
- Guaranteed service level agreements (SLAs) to avoid the financial impact of SLA penalties and improve brand reputation
- Improved employee productivity by freeing up expert resources from day-to-day tasks like debugging and root-cause-analysis
- Reliance on trusted and proven telecom expertise, to minimize the risk of inadequate solution capabilities and gain a competitive advantage

5G and edge computing are leading to the generation of large amounts of data at the network edge. As per Gartner's prediction, by 2025, 75% of enterprise generated data will be created and processed at the edge[3]. Moreover, ABI Research forecasts that 43% of AI tasks will take place on edge devices by 2023[4].

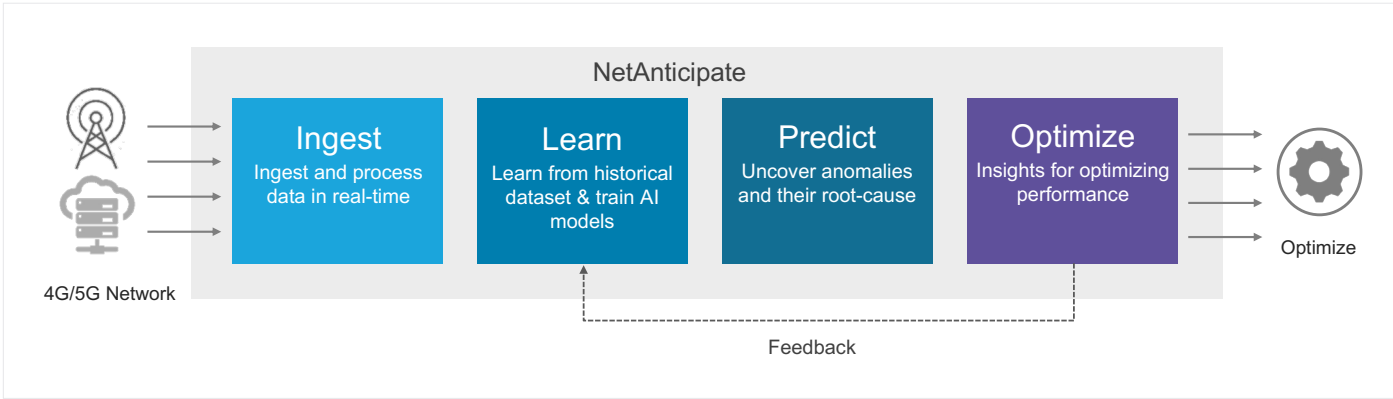

Figure 3. The Capgemini NetAnticipate Framework Overview

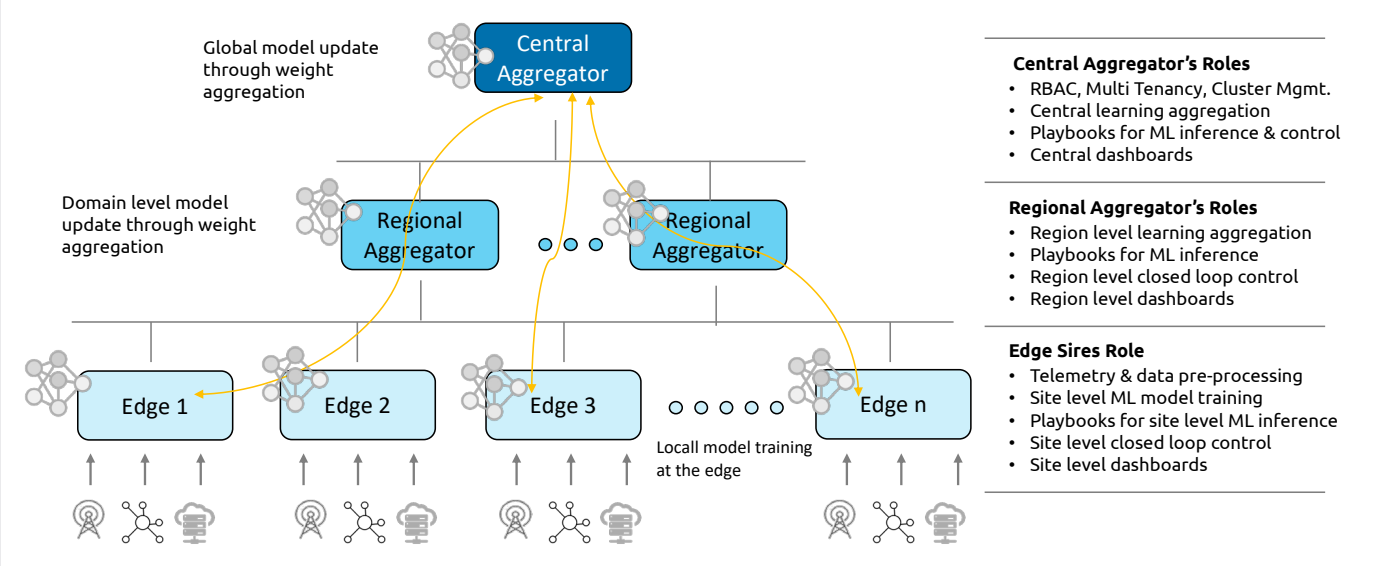

Figure 4. NetAnticipate Enables Disaggregated and Scalable Edge AI Processing Using Federated Learning

NetAnticipate has adapted FL for telecommunication and IoT networks to enable edge devices to collaboratively learn a shared prediction model while keeping all training data at the edge, thus addressing concerns about data privacy and time lag that model updates in traditional machine learning techniques generate. With NetAnticipate, FL solutions can be deployed in a highly disaggregated architecture that scales both horizontally and vertically (Figure 4) and also addresses the challenges posed by conventional FL described in the previous section.

The NetAnticipate edge AI solution is specifically focused on the telecom industry. Following are some of the relevant use cases. Privacy preserving refers to technologies that allow personal data to be protected while it is used:

1. Privacy preserving machine learning in multi-vendor and multi-geography radio access networks (RANs) for advanced use cases like dynamic spectrum management.
2. Privacy preserving federated learning for high quality V2X communication for autonomous vehicles through learning at the network edge.
3. Privacy preserving machine learning for high quality communication among robots and automated guided vehicles (AGVs) on a factory floor in a manufacturing plant, through learning at the network edge.
4. Privacy preserving machine learning for high quality communication among drone swarms for precision agriculture at the network edge.
5. Privacy preserving machine learning for high-quality communication and coordination for distributed cloud gaming at the network edge.

To demonstrate NetAnticipate's edge AI capability, Capgemini has developed a 'smart mobility' proof of concept, using a dataset consisting of user experience (UE) key performance indicators (KPIs) collected from two major Irish mobile operators, across different mobility patterns: static, pedestrian, car, tram, and train. Each edge client shares only a subset of mobility patterns in the AI dataset to better predict UE quality of experience (QoE) for each route (see Figure 5):

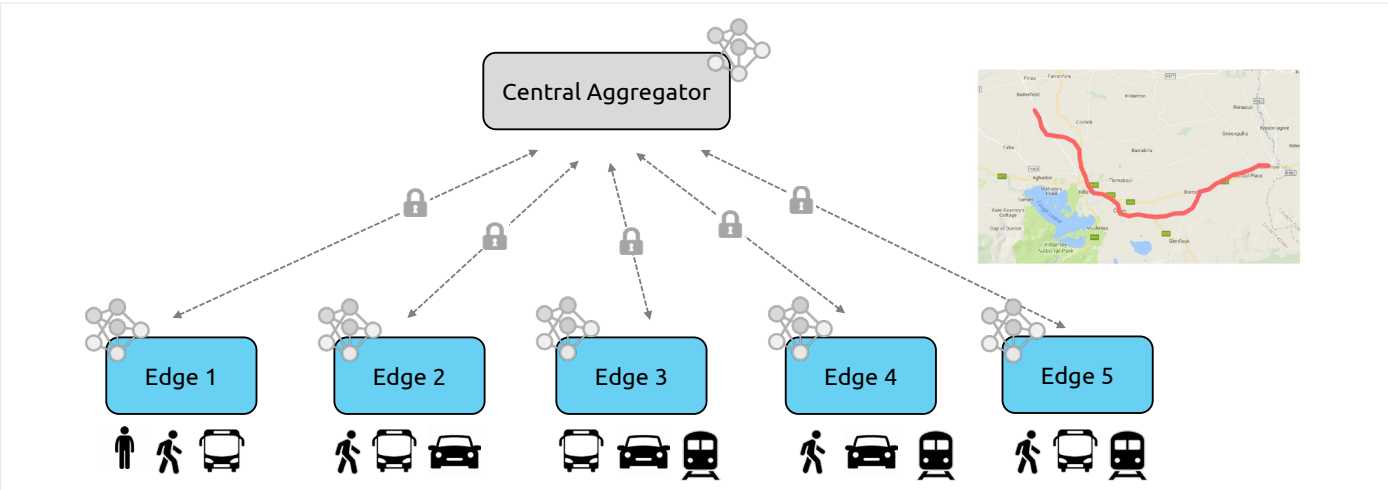

Figure 5. FL Architecture for Smart Mobility Proof of Concept

This experiment compared three FL frameworks: a. OpenFL, b. TensorFlow Federated and c. Flower. OpenFL and Flower FL frameworks performed closer to production grade product maturity, and OpenFL performed much more reliably and robustly in the overall architecture[5].

Future work planned includes a point-to-point decentralized FL using a serverless architecture.

Accelerating Federated Learning for Carrier Grade Edge AI

Servers based on 3rd Gen Intel® Xeon® Scalable processors are optimized for AI, serving many workloads, anywhere, at any time. 3rd Gen Intel Xeon Scalable processors incorporate AI acceleration, end-to-end data science tools, and an ecosystem of smart solutions[6].

Based on a balanced, efficient architecture that brings AI everywhere, from edge to cloud, 3rd generation Intel Xeon Scalable processors increase core performance, memory, and I/O bandwidth to accelerate moving diverse workloads from the data center to the edge[6]. CPUs are available with up to 40 powerful cores and built-in workload acceleration features including Intel® Deep Learning Boost (Intel® DL Boost).

Intel® Deep Learning Boost

Intel® DL Boost acceleration is specifically incorporated for the flexibility to run complex AI workloads on the same hardware as existing workloads. Based on Intel® Advanced Vector Extensions 512 (Intel® AVX-512), and Vector Neural Network Instructions (VNNI), Intel DL Boost delivers a significant performance improvement by combining three instructions into one[7]. This aggregation maximizes the use of compute resources, improves cache utilization, and avoids potential bandwidth bottlenecks.

Intel and OpenFL

OpenFL is a community-supported, deep learning, open-source framework for FL originally developed by Intel Labs and Intel Network and Edge Group (NEX), in collaboration with the University of Pennsylvania. OpenFL is designed to resolve 'cross-silo' FL problems when data is split between organizations, clients, or remote data centers. OpenFL enables data scientists to create federated learning experiments more easily by minimizing user entry points, simplifying processes to establish federations, and registering and running FL experiments.

Conclusion

The federated learning paradigm is revolutionary in that it helps to preserve security and privacy, is minimally impacted by latency and network bandwidth, and is scalable as network edge clients continue to grow exponentially and diversely so that over time FL solutions become more accurate and effective. For CoSPs, FL models make AI solutions more viable, cost-effective and secure, and help deliver more benefits and enabling more use case possibilities across the network. Intel processors, especially 3rd Gen Intel Xeon Scalable processors, are designed to deliver performance, flexibility, and efficiency in hosting such AI workloads.

Interested in speaking with our experts? Contact us at Intelsolutions.global@Capgemini.com.

Learn More

Capgemini Engineering NetAnticipate

Capgemini Engineering

3rd Gen Intel® Xeon® Scalable Processors

Intel® Network Builders

Reference

[1] https://www.scientificamerican.com/article/health-care-ai-systems-are-biased/
[2] https://aws.amazon.com/blogs/architecture/applying-federated-learning-for-ml-at-the-edge/
[3] https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders
[4] https://www.abiresearch.com/press/hardware-vendors-will-win-big-meeting-demand-edge-ai-hardware/
[5] https://youtu.be/BkMapLvFELw — see Appendix for technical specifications of servers used in YouTube video
[6] https://www.intel.com/content/www/us/en/content-details/630321/3rd-generation-intel-xeon-scalable-processors.html
[7] https://www.intel.com/content/www/us/en/content-details/630321/3rd-generation-intel-xeon-scalable-processors.html?wapkw=3rd%20generation%20intel%20xeon%20scalable%20processors

Appendix: Technical Specifications of Servers Used in YouTube Video:

SuperServer SYS-620P-TR	
https://www.supermicro.com/en/products/system/Mainstream/2U/SYS-620P-TR	
Processor	2 x Intel® Xeon® Scalable 6338N 2P 32C/64T 2.2G 48M 11.2GT 185W
Memory	128 GB Memory
SSD	1 x 960 SATA SSD with Tray
Exp Slots	4 PCI-E 4.0 x16 Low Profile (LP) 2 PCI-E 4.0 x8 Low Profile (LP)
LAN	On Board Dual Gigabit LAN with Intel® i350 Ethernet Controller
Addon NIC	1 x Intel® Ethernet Network Adapter E810-XXVDA4
	1 x Intel® Ethernet Network Adapter E810-XXVDA2
P. Supply	2U Rackmount Server with Dual Power Supply

Author: Subhankar Pal, Global Software & Digital Innovation Leader, Intelligent Networks, Capgemini Engineering