White Paper

Communications Service Providers
Network Slicing

TCS Cognitive Framework Adds Efficiency to 5G Network Slicing

Tata Consultancy Services improves 5G network throughput and compute resource efficiency by leveraging artificial intelligence and machine learning. Solution runs on edge servers powered by Intel® Xeon® processors.

Network slicing is a unique feature of 5G networks that provisions multiple customized logical networks on a common physical infrastructure and assigns necessary logical resources customized to the data flow of each network slice. Network slicing ensures that different data flows get the right combination of bandwidth and quality of service (QoS) features for that application. But the combination of dynamic data flows with static network slices provides the potential for congestion on some slices even if the overall network is not at capacity.

Intel® Network Builders ecosystem member Tata Consultancy Services (TCS) has created an artificial intelligence/machine learning (AI/ML)-based framework for cognitive network slicing that utilizes edge servers based on Intel® architecture CPUs. This framework delivers more efficient deployment of network slices to maximize both server resources and network bandwidth. This paper talks about the key benefits of AI/ML in a mobile network ecosystem and how TCS's solution can help optimize resources.

Network Slicing

5G was created to support a wide range of services from cellular data for mobile users to internet of things (IoT) to real-time communications for vehicle-to-everything (V2X) applications. Each of these services requires different bandwidth, latency, and quality of service (QoS) metrics. IoT, for example, does not need a lot of bandwidth, but needs connectivity for a large number of wireless sensors. Mobile broadband services, on the other hand, are designed to support video streaming and need much more bandwidth. V2X services require added reliability and very low latency.

Traditionally, operators offer differentiated services by configuring physical networks with specific business and use case requirements. These dedicated physical networks or physical slices are configured manually using networking devices like routers, switches, and physical network elements (PNFs) over multiple networks, which is quite challenging as it requires skilled network engineers to configure the network functions, underlying infrastructure, and connectivity across these networks.

Moreover, if there are network configuration changes or upgrades required, there is a need to manually reconfigure the network devices along with re-connections of cables, in many cases.

Network slicing technology for 5G, however, leverages software-defined networking (SDN) and network functions virtualization (NFV) to deliver multiple logical slices of network performance that enable 5G networks to meet this service variety.

Table of Contents

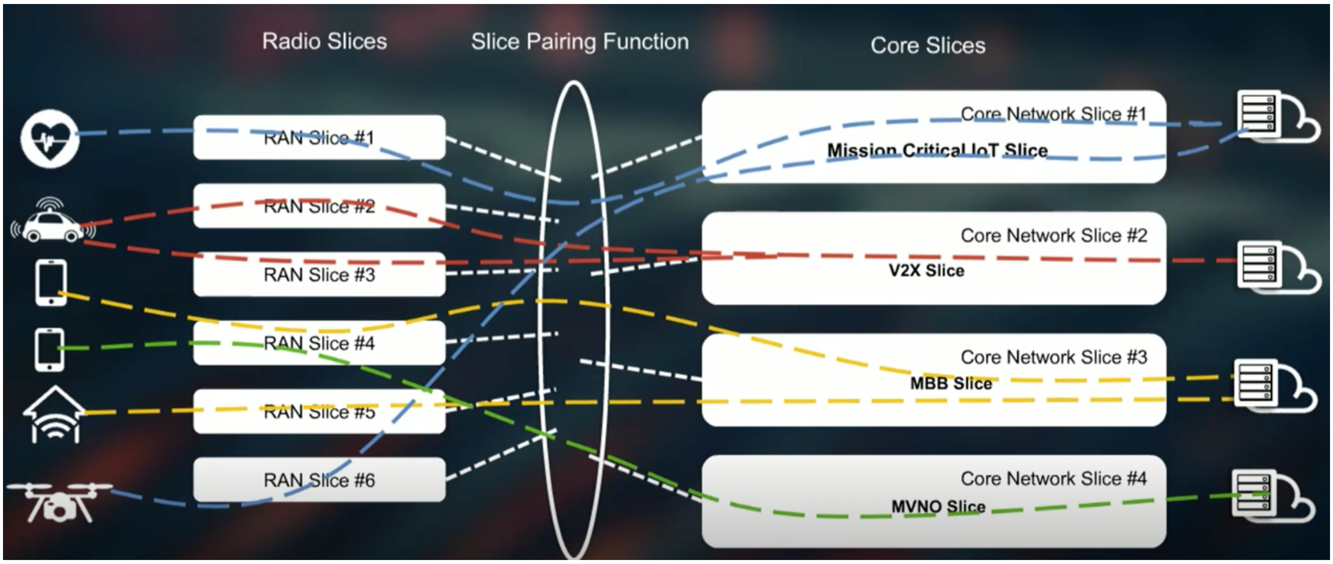

Figure 1. Network slice logical view

Figure 1 is a logical view of how network slices are developed. Users connect to the radio access network (RAN), which detects the type of traffic and creates a RAN slice that corresponds to a predefined core slice. A slice pairing function, provided through a network orchestrator, matches the RAN slices to the core slices to create the proper connection QoS parameters.

TCS Cognitive Framework

The TCS Cognitive Framework network slicing solution provides an end-to-end next-generation mobile network platform with an orchestration layer that enables on-demand service deployment with an automated and shorter deployment cycle, whereas the management layer helps manage the service lifecycle. It leverages open source cloud technologies like OpenStack, Docker, or Kubernetes for platform development and utilizes SDN principles for network isolation.

This solution is built on the TCS Microservices Development and Deployment Platform (MDDP), leveraging a set of strategically chosen services. The platform also provides a powerful zero-touch orchestration engine to ease complex multi-region deployments and bring lifecycle management features to the solution.

Key functional components of the TCS Cognitive Framework include the following:

- **Unique Framework:** TCS's comprehensive MDDP framework offers strategic services that could be readily used to ease the development of 5G network services. These services range from simple machine-first logging to complex cross-region context and cache replication. Furthermore, the framework also offers state-of-the-art orchestration services to efficiently deploy and manage 5G network services in a hybrid cloud environment.
- **Radio Access Network Slicing:** TCS has developed an enhanced MAC scheduling algorithm for dynamic resource allocation to support multiple network slices in order to meet the specific quality of service (QoS) requirements of user equipment (UE) and the service types they require.
- **Cloud-Native Mobile Core:** The solution's programmable mobile core data plane adds flexibility and agility to the mobile network. This solution exploits SDN and NFV principles to design a mobile core with a cloud-native, service-based architecture where the QoS on each data plane is programmed according to the specific network requirements of each business service.
- **Programmable Data Plane:** Dynamically provisions the data plane in the 5G mobile core by leveraging Open vSwitch to meet QoS needs.
- **TCS Cognitive Framework:** A closed-loop modular framework that allows a user to implement AI/ML algorithms.

Improving Bandwidth, Compute Efficiency

The TCS Cognitive Framework brings intelligence to the deployment of the 5G network in order to improve the efficiency of bandwidth deployment and of compute resources. The framework is cognitive in that it learns about the applications in use on the network from the data flows and can adjust network parameters based on this information.

The TCS Cognitive Framework leverages AI and ML algorithms to add intelligence to the network slice framework that can maximize both slice bandwidth and compute resources. Figure 2 shows how the framework gets the real-time data from the radio network that is analyzed according to business rules and policies to understand the intent of the data flows. The data is then fed into a machine learning algorithm for ongoing training and feedback. Based on the input data and its learnings, the algorithm generates predictions that can be then fed back to the network or sent to human network managers for intervention, which, in turn, improves the efficiency of the decisions. The predicted parameters are also sent back to the algorithm for ongoing training and feedback.

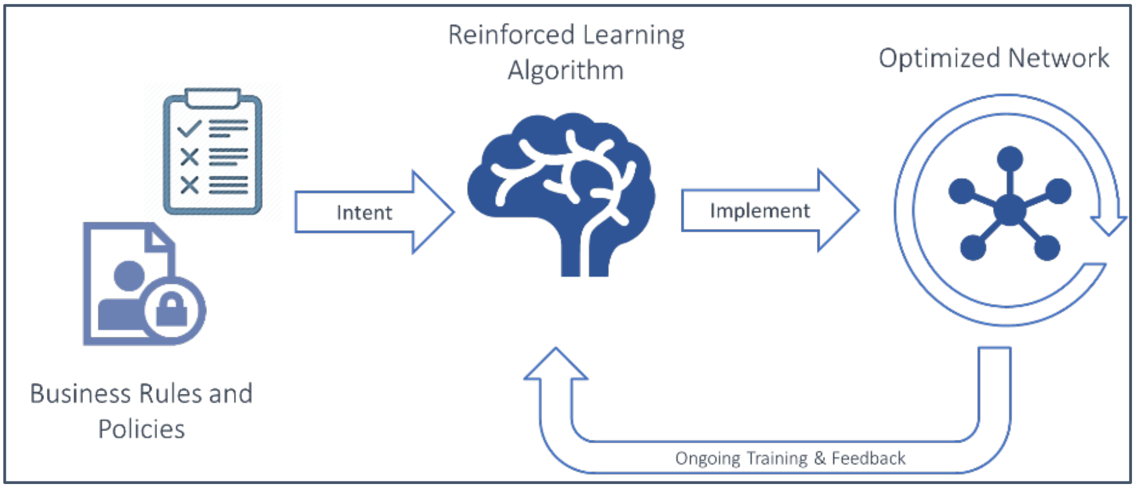
Reinforced Learning Algorithm
Optimized Network
Intent
Implement
Business Rules and Policies
Ongoing Training & Feedback

Figure 2. Logical view of TCS Cognitive Framework

As seen in Figure 3, the framework consists of four major modules with a dedicated functionality:

- The **Sense** component exposes an interface to collect real-time data from the access node. It uses a defined standardized application programming interface (API) to collect radio parameters from mobile access nodes, infrastructure, and stores in its database. This database includes UE and radio parameters, telemetry data from NFV infrastructure (NFVI) and uses defined policies.
- The **Analyze** module takes inputs data from the Sense module and extracts meaningful statistics and other characteristics of the data and forwards it to the Decide module.
- The **Decide** component uses a reinforced learning model to predict future values based on previously observed radio and infrastructure parameters over a time period.
- The **Respond** module, based on the predictions from the previous layer, sends back the response to either management layer or to the node or to the orchestrator to reconfigure and optimize the network and its parameters. This module also uses the dataset to train the model and learn from experience for future.

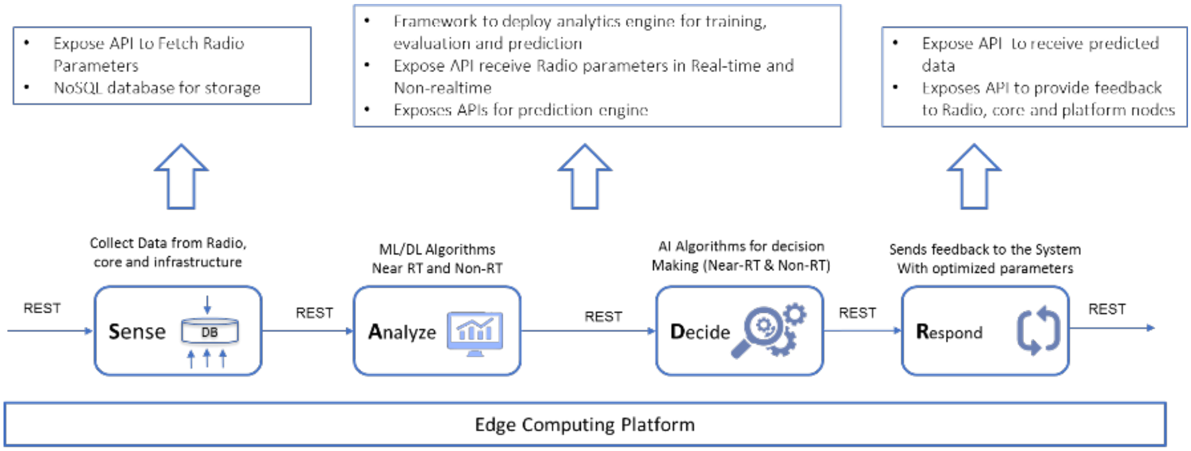
Expose API to Fetch Radio Parameters
NoSQL database for storage
Framework to deploy analytics engine for training, evaluation and prediction
Expose API receive Radio parameters in Real-time and Non-realtime
Exposes APIs for prediction engine
Expose API to receive predicted data
Exposes API to provide feedback to Radio, core and platform nodes
Collect Data from Radio, core and infrastructure
ML/DL Algorithms Near RT and Non-RT
AI Algorithms for decision Making (Near-RT & Non-RT)
Sends feedback to the System With optimized parameters
REST
Sense
DB
REST
Analyze
REST
Decide
REST
Respond
REST
Edge Computing Platform

Figure 3. Functional view of TCS Cognitive Framework

The TCS Cognitive Framework needs high-performance edge servers for its AI/ML functionality and real-time response and can benefit from the performance of servers based on Intel® Xeon® Scalable processors. These processors set a state-of-the-art level of platform convergence and capability across compute, storage, memory, network, and security. Edge servers built on the Intel Xeon Scalable CPU platform can deliver agile services with enhanced performance and groundbreaking capabilities compared to prior CPU generations. The Intel Xeon Scalable platform is designed for high-performance edge compute applications and can drive operational efficiencies that help lead to improved total cost of ownership (TCO) and high productivity for users.[1]

Optimizing Network Resources

Network slicing, with the help of SDN and NFV, decomposes a single mobile network connection into multiple logical networks that serve customized business needs to meet required QoS levels. Infrastructure resources for each service/slice are configured by an orchestrator. But if the business needs change, the resources need to be realigned by the orchestrator or virtual infrastructure manager (VIM). This requires manual intervention by a technician with some historical understanding of each slice and its usage pattern. This method is not an efficient way of optimizing the network resources.

The TCS Cognitive Framework can automate this realignment. To demonstrate this, below is an example that addresses the challenge of dynamically provisioning and optimizing the network resources such as bandwidth, CPU, memory, and others for different network slices based on the multiple parameters and UE usage pattern and their slice type.

Using the TCS Cognitive Framework and an open source slice framework, two slices are deployed with varied QoS requirements for each simulated UE. These simulated devices have specific data traffic patterns and are admitted into the slice that has the best aligned with their use case. At regular intervals, the sense process of TCS Cognitive Framework captures a range of parameters, including the number of active UEs, individual bandwidth utilization, slice bandwidth utilization, RAM, CPU utilization, and usage patterns across all the slices. This data is fed into the TCS Cognitive Framework as a training dataset.

This process trains the model over a period of time until it can predict the resource requirements of each slice before they become overloaded. Based on the model, this information is then fed back into the orchestrator to redistribute the resources based on the requirements and optimize the slice performance.

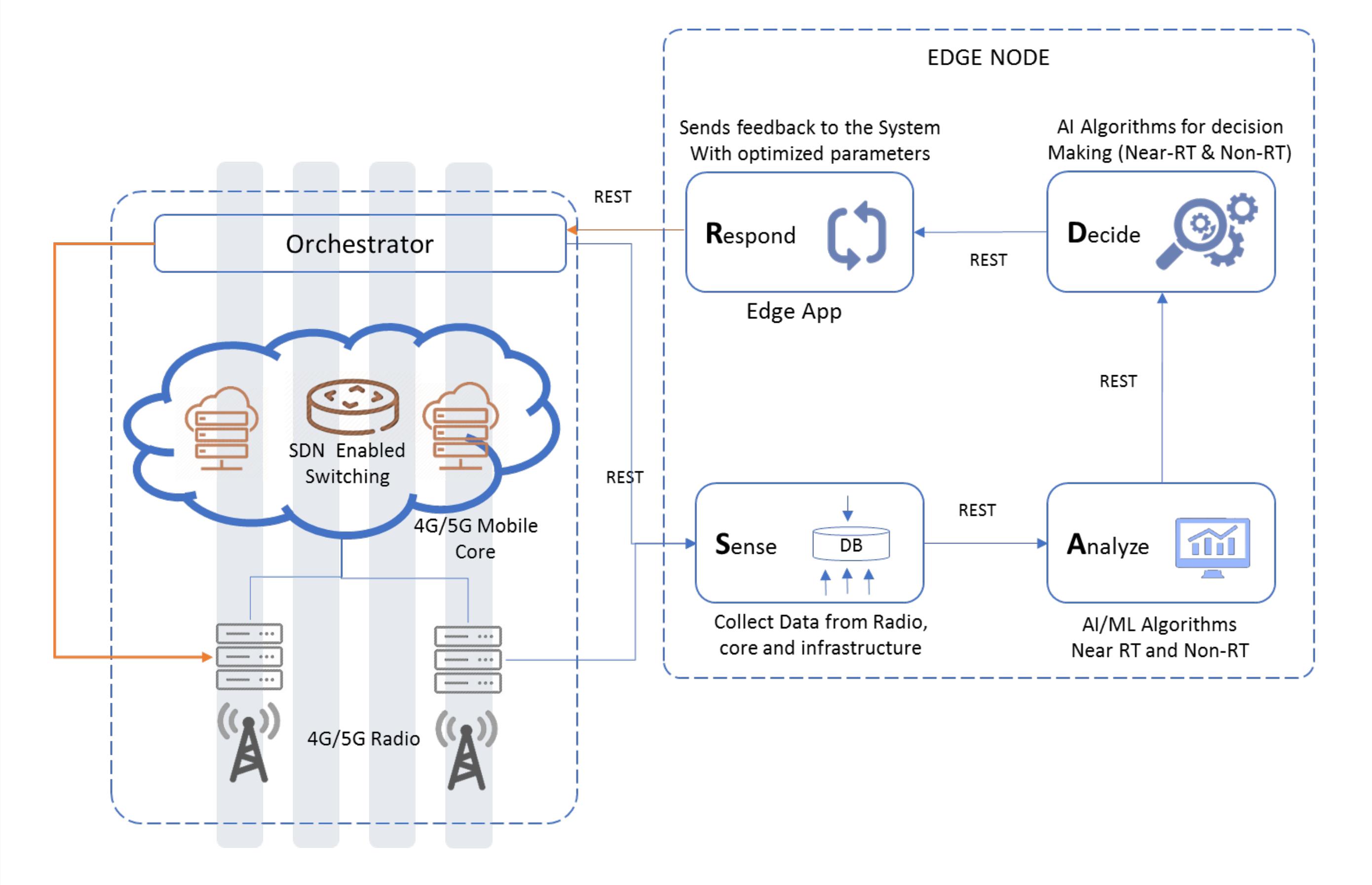
EDGE NODE
Sends feedback to the System With optimized parameters
AI Algorithms for decision Making (Near-RT & Non-RT)
REST
Orchestrator
Respond
Decide
REST
Edge App
REST
SDN Enabled Switching
REST
4G/5G Mobile Core
Sense
DB
REST
Analyze
Collect Data from Radio, core and infrastructure
AI/ML Algorithms Near RT and Non-RT
4G/5G Radio

Figure 4. Network slice deployment

In this case, the TCS Cognitive framework uses the Intel® Optimization for TensorFlow framework, on a server powered by a 10-core Intel Xeon processor with 64 GB RAM, to implement the supervised learning algorithm to create a linear regression model that sits as an edge applications to analyze and decide (see Figure 4). This edge application predicts the pattern and creates a relationship between network resource parameters and data traffic patterns. The goal of the algorithm is to predict an accurate network resource value for each slice at any point of time and notify the orchestrator with that value. Based on the training dataset captured, TCS has been able to train the model to yield accuracy to approximately 93% over a period.[2]

Simulation Results & Optimization

The performance of AI model algorithm is evaluated using system parameters according to 5G standards.[3] For evaluation purposes, a 5G test bed was created with two slices, one for a smart factory with more than 50 UEs and second slice for a college campus, with 25 UEs in each slice and with varied resource requirements.

In the college campus slice, the maximum data rate for the UEs in the generic slice is set to 1 Mbps and min is set to 100 Kbps. The data rate for UEs in the smart factory slice is set to a maximum of 500 Kbps and minimum of 100 Kbps, assuming a mix of IoT and enhanced mobile broadband (eMBB) resource requirements.

Below are the few key compute KPI formulas used by the algorithm to calculate the resource usage per slice.

1. Slice bandwidth usage: Summation of bandwidth usage by all the UEs in the slice per transmission time interval (TTI).
2. Active protocol data unit (PDU) sessions/slice: Number of active PDU sessions in a single network slice instance.
3. Max bandwidth/slice: Allocated bandwidth for each slice.

The graph below represents the reallocation of under-utilized bandwidth from the college campus slice during the time of the day when the utilization is low, to the smart factory slice where there is need for additional bandwidth. It is worth noting that the algorithm does not allocate the unused bandwidth from the smart campus slice during the daytime hours (9 am to 6 pm), when there is the potential for that bandwidth to be used by students and faculty. The algorithm develops this intelligence based on historical and current usage patterns.

This can be further extended to hardware resources such as CPU and RAM. If the utilization is low, then based on the closed loop information, the orchestrator can re-allocate the CPUs and RAMs to other processes that have a need for those resources.

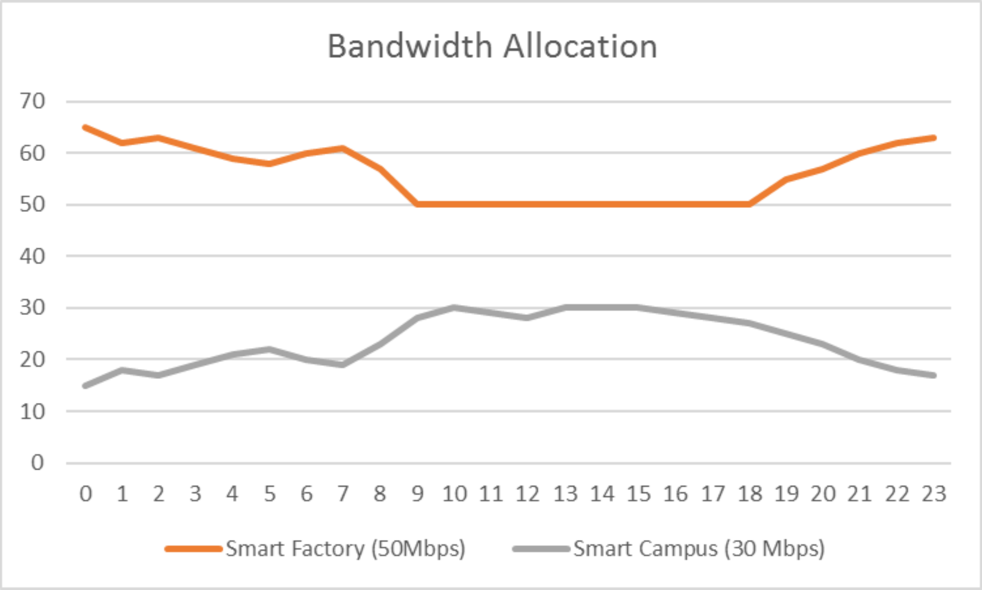

Figure 5. Dynamic bandwidth allocation

Conclusion

Network slicing is a key feature of 5G networks because it allows for a wide range of services to be supported across a common wireless infrastructure. Mobile network operators (MNOs) will be able to develop services for merging IoT applications involving drones, utilities, and manufacturing with unique slices created and optimized for those data flows. With the TCS Cognitive Framework, MNOs or enterprises can more efficiently use network and compute resources to maximize the utility and value of these network slices.

Learn More

Tata Consultancy Services (TCS)

TCS is a member of the Intel® Network Builders ecosystem

Notices & Disclaimers

[1] https://www.intel.com/content/www/us/en/products/docs/processors/xeon/2nd-gen-xeon-scalable-processors-brief.html

[2] Data provided by TCS, October 2020.

[3] Tests were conducted by TCS in October 2020. TCS recommends using Intel® Xeon® Scalable processors for optimal performance. However, due to hardware availability issues caused by the COVID-19 pandemic, the simulation described in this paper used a server powered by a 10-core, 2.2GHz Intel® Xeon® processor E5-2630 v4 running 20 threads per socket (microcode 0x1). The CPU had Intel® Hyper-Threading Technology turned on, and Intel® Turbo Boost Technology also turned on (up to 3.1 GHz). The system BIOS was SeaBIOS 1.21.1. System memory totaled 64 GB. Server operating system was Ubuntu 18.04.4 LTS with Kernel version 4.15.0-112-generic. Other software included GCC Compiler v7.5.0, Scikit-Learn v0.23.2, and TensorFlow v2.30.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

0421/DO/H09/PDF ♻ Please Recycle 345549-002US