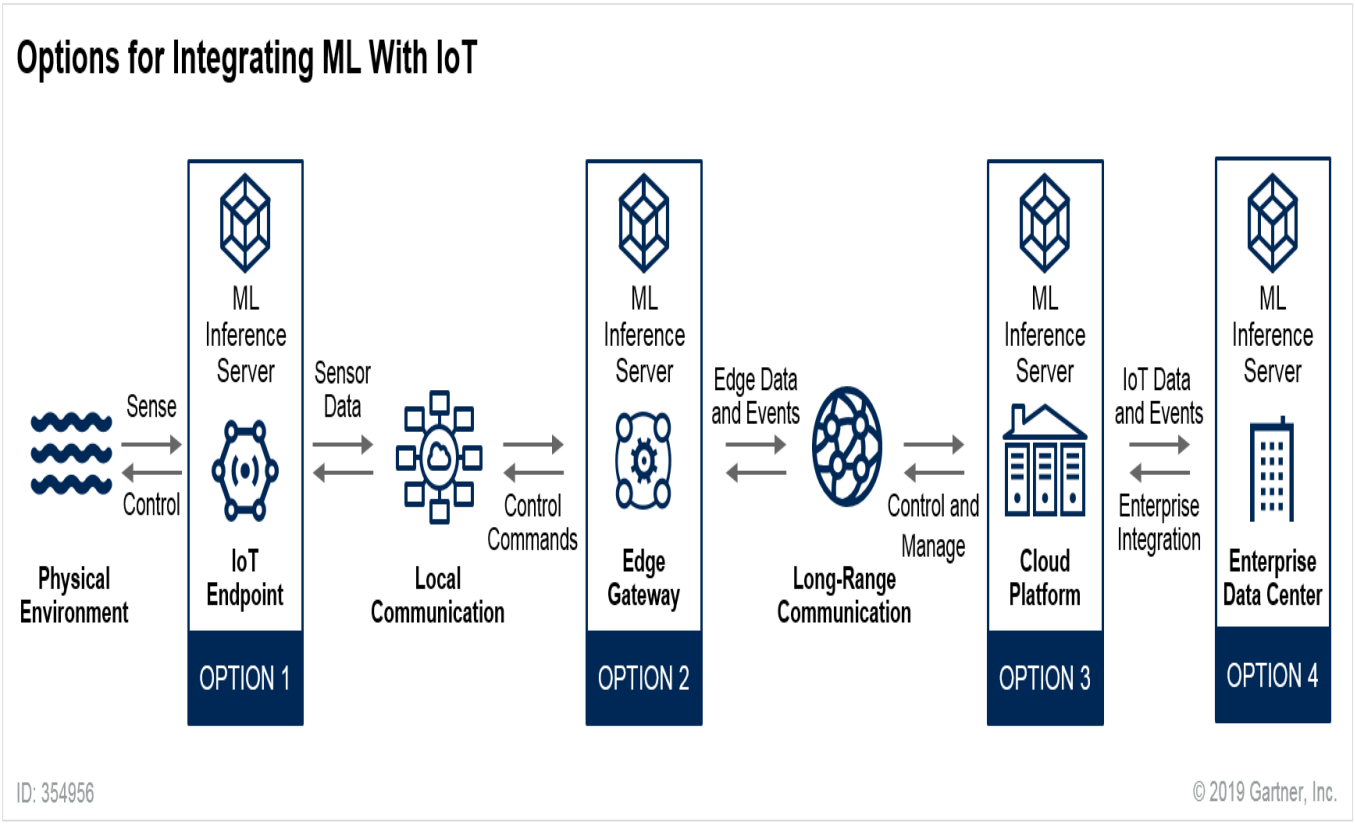
Four Options for Integrating Machine Learning with IoT - Paul DeBeasi

Machine learning projects are inherently different from traditional IT projects in that they are significantly more heuristic and experimental, requiring skills spanning multiple domains, including statistical analysis, data analysis and application development. Most organizations have defined the process to build, train and test machine learning models. The challenge has been figuring out what to do once the model is built. Integration, deployment and monitoring are essential aspects to provide for continuous feedback once the models are in production.

IoT is one of the most disruptive forces organizations must contend with today. IoT solutions integrate multiple technology and business operations and impact mission-critical processes and products. IoT technologies continue to evolve and morph quickly with few true standards.

As the number of IoT endpoints proliferate, the need for organizations to understand how to design systems that integrate machine learning inference with IoT will grow rapidly. Given the fact that IoT solutions are distributed systems, a key design question is "Where should my organization deploy the machine learning inference server in the distributed IoT system?". (For an overview of machine learning inference servers.

The figure below illustrates four options that will form the foundation for creating a system design that integrates machine learning with IoT.



Options for integrating Machine Learning with IoT

- **Option 1 — IoT Endpoint**: In this option, the machine learning inference server is integrated in an IoT endpoint such as a microcontroller-based system (e.g., smart thermostat). It can provide machine learning inference services for a single IoT endpoint.
- **Option 2 — Edge Gateway**: In this option, the machine learning inference server is integrated in an IoT gateway (e.g., an edge server). It can provide Machine Learning inference services for one or more IoT endpoints at a single edge location.
- **Option 3 — Cloud Platform**: In this option, the machine learning inference server is integrated into a cloud-based IoT platform (e.g., AWS IoT and Azure IoT). It can provide Machine Learning inference services for many IoT endpoints across many IoT edge locations.
- **Option 4 — Enterprise Data Center**: In this option, the machine learning inference server is integrated with the on-premises data center. It can provide Machine Learning inference services for many IoT endpoints across many IoT edge locations.

Note that it may be necessary to integrate the machine learning inference server at more than one location to achieve your design goals. For instance, you may need to design an ensemble Machine Learning pipeline that uses an machine learning inference server in each of the IoT endpoints as well as an machine learning inference server in the IoT gateway. In this case, you can create a hybrid design by using design elements from two or more of the reference architectures in the report described below.