A **Micron** White Paper

The Secure Intelligent Edge

Micron's innovations at the edge—enhanced by Tempered Networks and integrated by P2 Solutions Group—provide the most performant and most secure intelligent edge solutions available

We've all seen the forecasts pointing to the tens—and even hundreds—of billions of IoT devices expected to be deployed over the next few years. That ubiquity speaks to the remarkable flexibility of these devices and the simplicity of their deployment. However, there is a vast difference between sensors deployed at the edge and actual edge intelligence—to say nothing of reliability, performance, and equally important, security. Achieving all these vital objectives requires a holistic approach when engineering effective solutions that comprehend the entire ecosystem—and its myriad vulnerabilities.

The Big Picture

Let's begin with an overview of the major components, endpoints, and data flow aspects of the edge environment:

First is the **edge** itself. More than just a location, the edge comprises both the environment and the framework deployed to sense, capture, filter, compute, process, store, analyze, communicate—and sometimes actuate—events, each of which can involve sophisticated technologies that must not only work together seamlessly, but also accomplish their work in small, inexpensive, and power-stingy form factors.

Just this side of the edge are the **hubs**, **gateways**, and/or **communications cells** responsible for bidirectional traffic control and data aggregation. This side is also concerned with questions of when, where, and how preprocessing/filtering is to be performed across the network, as these functions in their various forms are partitioned across the ecosystem.

Finally, we have the **cloud** and the nature of the analytics performed here based on data collected or passed on to it from the edge or its gateways.

Across this matrix—and common to all nodes—are matters such as data management, device provisioning and management, control, monitoring, and diagnostics; connectivity (2G, 3G, 4G, LTE, 5G, RFID, Bluetooth, ZigBee, satellite, Ethernet, LAN/WAN, etc.); authentication and security, and various stacks. Viewed this way, we can see the edge is merely a part of a much larger, and potentially complex, ecosystem.

Then there are the business objectives—strategic, tactical, and logistical. For example, what specific problems are being solved? What is to be the chain of custody of the data across the network? What are the specific compute needs of the various endpoints deployed across the network? Indeed, the intelligent edge comprehends a great many little pictures that ultimately comprise a much larger operational picture.

At Micron, working with our integration partner P2 Solutions Group, we begin the engineering of our secure intelligent edge solutions by examining the value we can bring to address, control, or improve operations at each node of the ecosystem.

Micron

In examining our value, we characterize, for example, the opportunities for the application of deep learning and other novel compute elements at each node (accounting for the various types of neural networks, throughput, and bandwidth profiles), as well as the memory requirements (type, density, bandwidth, etc.), and communications aspects—and all in the context of an environment where data must be exploited to the greatest degree possible in real time, while it is still in motion—and in secure fashion—with the overall system extracting valuable information to support distributed and remote decision making.

Making the edge work, then, is ultimately more of a data challenge than a device deployment and connectivity challenge. Indeed, extracting data from devices is one thing; protecting it, understanding what it all means, and being able to act on it is quite another. So, while the market has been seemingly obsessed with getting smart gadgets connected online, there's been comparatively little attention and innovation in the area of enabling the profitable consumption of the data that these edge devices generate—with the requirement that the data remain secure. Consequently, many edge/IoT "solutions" fail in the real "last mile."

Data, Data, Everywhere...

Where the edge is concerned, it is vital to understand that all sensors—and the data they generate—are not created equal. And that inequality comes in many forms. In the larger context of smart infrastructure, for example, there may be many interrelated components—components that may vary widely in their data types, formats (structured and unstructured), periodicity, quality (sensor data can be notoriously "dirty"), transmission medium, and other attributes. Yet key to making these elements work together across their respective domains is the ability to orchestrate these disparate data sources in order to aggregate and make sense of them at progressively higher levels of abstraction. When you can do this, you can effectively enable new forms of intelligence via the patterns the system is able to detect (and correlate in the time domain) from the various and diverse endpoints.

Combined with our unique deep learning-enabled ability to also tag, label, feature extract, triage, and route data as it streams into the system's aggregation points, the networking of myriad edge sensors can in turn enable the generation of hypergraph analytics that can be leveraged to identify and flag larger trends and generate alerts as appropriate. Thus, enabled by intelligently networked edge, on-premise, and cloud-based processing, raw field data is transformed in real time into valuable information.

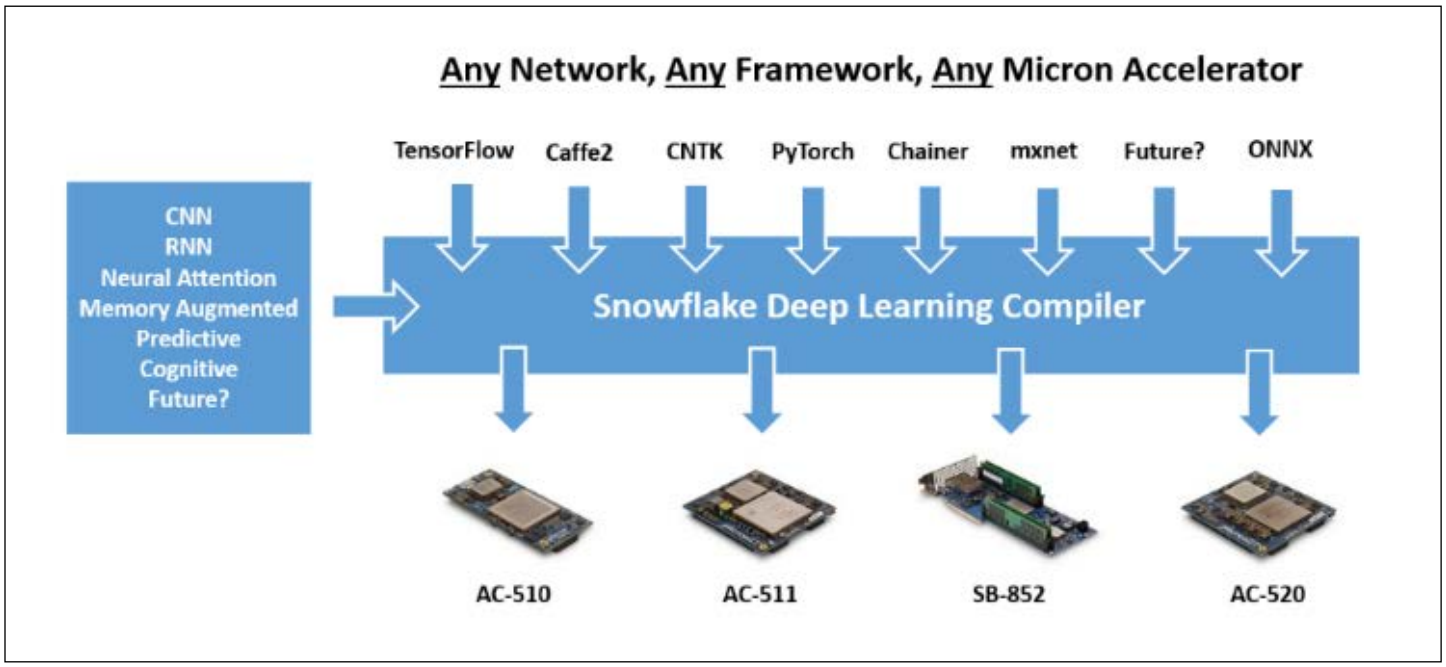
Whether the data collected at the various endpoints involves full motion video, wide area motion imagery, electronic signals, chemistry, communications, or other data/signal types, Micron's secured machine/deep learning capabilities can aid in the automation of processing, exploitation, and dissemination of such signals whose applications can span virtually any tactical, mobile, or other edge platform to provide integrated intelligence support for larger operations centers *anywhere they are needed*.

Developing, Deploying, and Scaling Edge Intelligence

To speed the development of intelligent edge deployments, Micron's integrated deep learning development environment was engineered to be exceptionally easy to use, enabling developers to go directly from any machine learning framework (for example, TensorFlow, PyTorch, Caffe, etc.) directly to hardware implementation, accelerating virtually any neural network with unprecedented ease.

Unlike other machine learning solutions, there's no requirement for developers to have expertise in HDLs like Cuda, OpenCL, or Verilog—Micron's neural network compiler does all the work, automatically instantiating trained networks into the hardware accelerators. Once training has been performed in the preferred framework, it can be saved as ONNX interchange format.

Micron's deep learning environment



Let's take a closer look at the hardware itself.

Micron's edge solutions are built upon a novel modular acceleration platform that enables developers to work completely within their preferred machine/deep learning framework while dramatically improving inference performance over competing technologies.

Our solution leverages the seamless integration of a sophisticated deep learning stack and easy user interface. Designed for either a standard PCIe interface in a conventional gateway or server configuration, or a dedicated carrier board for small form factor-embedded applications, our deep learning accelerator modules aren't much bigger than a business card and feature Micron's advanced memory technologies, such as the high-bandwidth Micron Hybrid Memory Cube (HMC). Shown here is the AC-511 that combines a Kintex® UltraScale FPGA with the 2GB HMC and 16GB of DDR4.

The HMC enables breakthrough levels of bandwidth and throughput needed for real-time processing of, for example, the massive amount of streaming camera and other sensor data, while slashing both power consumption and physical size.

Micron offers a family of various module configurations with either Xilinx or Intel® (Altera) FPGAs, as well as PCIe carrier boards that can accommodate up to six FPGA/HMC modules. We also have a growing family of single-board computers that support both PCIe and QSFP interfaces.

Micron's AC-511 High-Performance Computing module

Securing the Intelligent Edge: The Fundamental Challenge

To protect against intrusions—a challenge exacerbated by the proliferation of IoT devices and their associated attack surfaces—Micron's edge solutions integrate a multilevel approach to system security that is enhanced by a novel combination of neural network security techniques, as well as dataset and neural network framework tools that provide strong security on both data and the outputs of neural network training layers, providing robust anomaly detection, including neural network and dataset contamination.

More importantly, thanks to security innovations by Tempered Networks, we are now able to make edge deployments completely invisible to would-be attackers.

When contemplating the securing of the intelligent edge, one thing becomes instantly and abundantly clear: You cannot look to traditional IT security stacks to secure edge infrastructure. Today's IT solutions were not designed to address the scale, availability, security, and visibility demands of IoT—with billions of devices lacking built-in cybersecurity increasingly connected to aging IT networks. The result is exponentially greater liability with complex and porous networks that are extremely difficult to protect with traditional firewalls and segmentation.

7 Ways FPGA-based Solutions Bring Significant Inference Advantages to the Edge

There's no question the edge poses challenges with respect to power and bandwidth. Memory bandwidth is key to machine learning performance, particularly at the edge—the very source of the data—and even more so as computation is becoming more localized. Prediction is being pushed increasingly to the edge where it encounters tight requirements for limited weight, size, and power, to say nothing of very limited external bandwidth. So how well do FPGAs perform under such conditions? Here are seven reasons why FPGAs just may be your preferred choice:

1. A properly engineered inference engine operates on low-precision data with high (near parity) accuracy. Because we need to do more with less at the edge, it's worth noting that FPGAs are designed to perform concurrent fixed-point operations with a close-to-hardware programming approach. In short, fixed point is far more computationally efficient, and yields superior inference performance (algorithms run faster with simpler math). Moreover, FPGAs directly pipeline one layer into another without having to go to memory, saving time and energy by eliminating costly data movement.

2. FPGA architecture is exceptionally flexible, allowing an architecture that better suits lower precision network topologies versus other accelerator platforms. Also, by reducing bitwidth, FPGAs synthesize a larger number of operations and parallel/pipelined kernels within the fabric's fixed area, easily accommodating more network layers (more layers yield higher accuracy)—again, doing more with less. FPGA architecture is inherently parallel *and* pipeline oriented.

3. FPGAs typically operate with 1/10th the power of an equivalent GPU implementation. This lower power consumption requires less thermal dissipation countermeasures, enabling smaller system dimensions and contributing to greater flexibility in deployment.

4. Because of the disparate requirements of edge applications, FPGAs are preferred because they are adaptable to virtually any interface. Their onboard I/Os may be dynamically reallocated without altering the physical connections.

5. FPGAs exhibit significantly less latency than GPUs, even with deeper networks. Algorithms implemented on FPGAs provide deterministic timing, with latencies an order of magnitude less than GPUs.

6. Smaller kernels enable more of the network to fit on a single FPGA. Moreover, FPGAs are superior to GPUs in many architectures because a network can be partitioned across multiple FPGAs without going through the host CPU. Also, FPGAs incur no redundant instruction fetch and decode overhead, as execution is fully in hardware.

7. FPGAs accommodate the higher memory bandwidth needed to "feed the flops." Mating FPGA with HMC (and other high-performance memory technologies from Micron) yields more efficient use of the available bandwidth, particularly for larger data transfers/streaming. While GPUs may have greater raw memory bandwidth, much of it is consumed with instructions and data movement. Consequently, FPGA's effective memory bandwidth, while lower, is actually more compute-efficient because FPGAs omit "housekeeping" accesses not directly involved in calculations.

The effective deployment and management of the edge requires simplicity; the legacy IT security stack only contributes complexity—leaving the network exposed to massive vulnerabilities. Moreover, more complexity also means adding more staff, more rules, more complicated workflows, and creating more opportunities for something to go wrong. And things always go wrong.

This is in large part because IT security is centered on the human factor—always the weakest link. This is not the case with the IoT—the myriad endpoints that comprise data capture at the edge are the weakest link, and there are far more of them than people. Consider just a few of the key differences:

- Securing IT devices that perform many different functions requires a complex set of security components. IoT devices are generally "single function" with a very limited number of hosts with which they communicate.
- Hacking IoT focuses on network exploitation and controlling the devices themselves, whereas hacking IT focuses on users and gaining access to data.
- With the explosion of IoT, the cost traditionally budgeted to secure each IT device will not scale to support IoT.
- With IoT, support is typically handled by the business units or operational groups, not IT. Consequently, applying an IT support model may not align well with IoT workflow.

When considering the many—and consistently reported—failures that result when IT approaches are applied to the edge, IT architects quickly realize:

- Legacy technologies (NACs, VLANs, firewalls) are not built for IoT. Emerging technologies (such as anomalous behavior detectors) are costly and add to the resources needed to maintain the "security posture."
- The integration of multiple technologies and their many moving parts is both costly and complex.
- The configurations do not scale, fail to lock down the devices fully, and controls are difficult to update.
- The workflows to onboard, decommission, and move/change devices require multiple steps spread across departments and technologies.
- While compliance requirements might be "checked off," important security gaps remain.

What, then, is the solution? In short, a single, fully integrated and automated edge network control architecture that completely and easily cloaks it from bad actors.

When cloaked, an edge device or appliance will not respond to any data packet sent to it, unless that packet is properly encrypted, signed, and authorized. This means the device will not even respond to a ping or port scan, rendering it effectively invisible to an attacker. This in turn makes DDoS, vulnerability scans, VPNs, purpose-specific VLANs, and internal firewalls obsolete.

It is to all these ends that Micron has adopted the Tempered Airwall™, a purpose-built IoT cybersecurity platform to secure our intelligent edge solutions.

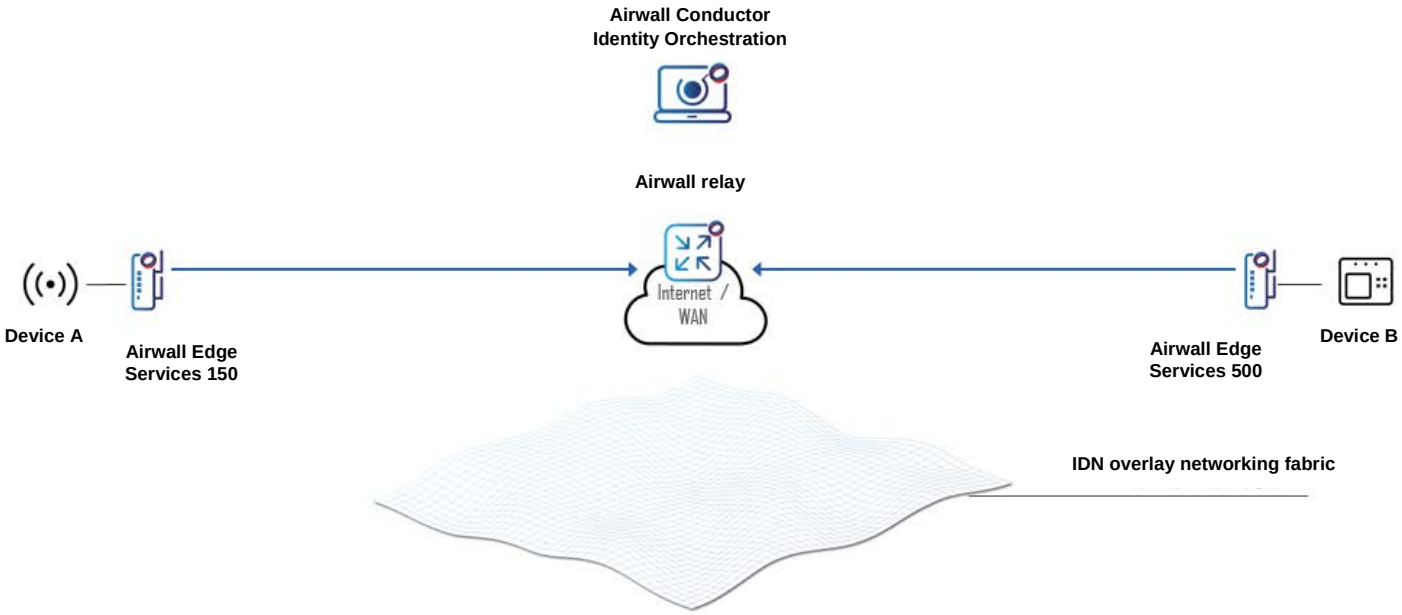
Tempered Networks

The Tempered Airwall addresses the new and emerging demands of converged infrastructure, enabling a massive reduction in the attack surface presented by IoT devices—and doing so without having to upgrade existing IT infrastructure.

Deployed as a trusted layer within an existing network, Tempered's Airwall enhances resilience, scale, and security while eliminating the need for additional firewalls, ACLs, VPNs, or specialized skills. The platform provides 1) north/south and east/west micro-segmentation across all networks; 2) simplified network management at the scale mandated by the edge; 3) support for real-time network analysis with Snort integration; and 4) extended functionality with edge switching capabilities.

The Tempered Airwall is based on the IETF standard Host Identity Protocol (HIP), which both addresses the unique needs of IoT devices and fixes the fatal flaw in TCP/IP networks—that IP addresses are used for both a device's location and identity on a network. Every device with an IP address for an identity can be found. In short, if they can be found, they can be hacked. The Tempered Airwall doesn't use IP addresses for identity. No IP address means devices cannot be found inside an Airwall. If you can't find them, you can't hack them.

Tempered Airwall



Three Key Components of the Tempered Airwall

In securing the entire network from edge to cloud, the Tempered Airwall:

- Eliminates lateral movement within a network.
- Automates security policies at scale for all connected devices with the click of a mouse.
- Provides zero-trust micro-segmentation to authorized personnel and devices.
- Reduces audit scope in PCI, HIPAA, and NIST regulated environments.
- Supports both high availability and device mobility.

In summary, the Tempered Airwall enables breakthrough IoT security, along with much needed flexibility, mobility, and resiliency. And equally important, when integrated with Micron intelligent edge solutions, it yields compelling ROI, rapid time-to-value, unprecedented efficiency and efficacy, and true, effective compliance.

Finally, by virtue of Micron's deep learning stack, the system is able to provide additional and robust functionality for the secure network layer with:

- Packet inspection
- Traffic analysis
- Graph analytics
- Usage patterns
- Optimizations and systems analysis
- Personalized feature options

Taken together, the fully integrated solutions provide not only the most performant, but also the most secure intelligent edge solutions available.

Making Sense of the Edge

Imagine a smart ecosystem as complex as that of a city—a city that encompasses potentially millions of endpoints, each of which is continuously generating its own data exhaust. We can take in a view of that ecosystem by considering the many different sensor types that might be deployed—sensing a wide variety of states including any number of process variables, physical properties, proximity and positioning of mobile objects of interest, chemical properties, electrical properties, and various forms of imaging. They would further encompass specific applications, including industrial, motor vehicles, security, medical, and other information technologies.

Here, then, is a sampling of the myriad sensor types whose data is not only intrinsically interesting even when siloed, but when taken in the larger context of events unfolding across a city, can be transformative—particularly where there are interdependencies and/or correlations among them that can be graphed to provide additional levels of insight:

- Temperature (AC, environmental, industrial, agriculture, health, resistor, thermistor, IR, IC)
- Proximity (retail, vehicle, parking, capacitive, photoelectric, ultrasonic)
- Pressure (water, heating, industrial)
- Water (turbidity, pH, oxygen-reduction potential)
- Chemical/smoke and gas (carbon dioxide, breathalyzer, hydrogen sulfide, etc.)
- IR (heat, healthcare, security)
- Level (fuel gauging, point, continuous)
- Load meters
- Flow
- Occupancy waste
- Parking image (CCD)
- Motion (passive IR, ultrasonic, microwave)
- Accelerometer (Hall effect, piezoelectric, capacitive)
- Gyroscope (car navigation, ADAS, UAV control, rotary, vibrating, optical, MEMS)
- Humidity (industrial, environmental, materials storage)
- Optical (and electro-optical) (ambient light, photodetectors)
- Audio

High-performance neural networks can be created for adding intelligence to these and virtually any other data source. Moreover, the sensors and their processing via Micron edge solutions can be deployed in fixed, mobile, and/or on dynamic platforms, including drones and other devices that are capable of capturing and transmitting video with other forms of data.

About P2 Solutions Group

P2 Solutions Group is Micron ACS's strategic VAR and integrator, developing and deploying secure intelligent edge solutions globally in healthcare, smart building/city infrastructure, energy, and other critical areas.

micron.com