Solution Brief

Intel Programmable Solutions Group
Intel® Stratix® 10 FPGA
Neuromorphic Computing

Neuromorphic Computing at Human Scale on Reconfigurable Hardware

The grand challenge to reverse-engineer the brain enters a new era made possible by Intel® FPGAs.

WESTERN SYDNEY UNIVERSITY

International Centre for Neuromorphic Systems

Summary

This paper describes a project underway in Australia to build a large-scale neuromorphic computer with hardware that can be reconfigured using software. The primary technology that enables such configurable hardware is the FPGA, and other technologies incorporated into Intel® FPGA products enable this project to scale to a size comparable to the numbers of neurons and synapses in the human cortex.

Building a new kind of computer

The International Centre for Neuromorphic Systems (ICNS) at Western Sydney University (WSU) is planning to build a computer unlike any other—a scalable neuromorphic compute system consisting of field-programmable gate arrays (FPGAs) interconnected by a high-performance computing (HPC) network fabric. The goal is to enable research into new frontiers of artificial intelligence (AI) and neuroscience by creating the world's first configurable neuromorphic computer at brain-scale.

What is brain-scale computing? The idea is to create a computing environment comparable to the number of neurons and synapses in the human cerebral cortex. Estimates of the number of neurons in the cerebral cortex range from 10 to 20 billion, and the number of synapses is much higher—from 60 to 240 *trillion*.[1] Achieving a scale similar to this with artificial neurons and synapses is the challenge of brain-scale computing, but it promises to deliver novel ways to study information processing in biological brains, including when things go wrong, and to develop better machine learning (ML) and AI.

Neuromorphic computing is the branch of computer science that seeks to emulate the structures and functions of the biological brain—neurons, synaptic connections, and spikes—as closely as possible in hardware. It is an interdisciplinary field requiring in-depth understanding of biology, computer architecture, and ML.

An FPGA is a special kind of microchip that has the unique ability to be configured and reconfigured at the hardware level using software, making it possible to implement different models for organizing artificial neurons on the same hardware.[2] The feasibility of using FPGAs for neuromorphic computing has been proven in concept by the WSU team on a single FPGA board, but such a project has never before been implemented at brain-scale.

Once this configurable neuromorphic platform is built at WSU, it will provide unprecedented opportunities for research and development in neuromorphic hardware design by enabling new designs to be implemented without new hardware, quickening the pace of efforts to break the neuromorphic code.

Three major developments converged to make this project possible. First is the seminal research and proof-of-concept (PoC) done at WSU, starting with the knowledge that FPGAs can be configured and reconfigured to simulate different kinds of biological structures in the cerebral cortex, and advancing the idea by developing a highly efficient kernel and methods for scaling to a large number of neurons.[3]

Second, the latest generation of Intel® Stratix® 10 FPGAs with high-bandwidth memory integrated on the BittWare 520N-MX board makes available a powerful and affordable off-the-shelf hardware platform.[4]

Finally, Intel's Configurable Network Protocol Accelerator (COPA) technology provides a high-bandwidth and low-latency interface to connect large numbers of FPGAs together in a highly scalable fashion—a capability that was missing in the past, limiting the use of FPGAs at scale.[5] The COPA technology also provides simple programmatical access from the FPGA to the host computer's resources, such as the memory and hard drive, thereby providing the potential to virtually scale up the system to human brain size (though at the cost of real-time performance).

This paper discusses neuromorphic computing as a promising new path toward reaching breakthroughs in AI and advancing our understanding of neural computation in our brains. It explains how FPGAs work and why they are critical to accelerating the cadence of neuromorphic research. And it describes how COPA and other Intel technologies provide the key to unlocking the power of FPGAs in brain-scale neuromorphic computing.

An interdisciplinary endeavor

The neuromorphic accelerator at WSU is designed to be an iterative and multi-disciplinary development platform. HPC architects and experts in ML and neuroscience are working together to deliver a general-purpose, off-the-shelf system that allows the acceleration of discovery in this field.

The opportunity to explore and collaborate across disciplines is a two-way street for computer scientists and neuroscientists. Neuroscientists should find the neuromorphic compute system valuable as a tool for testing theories that prove difficult to test in living brains. For example, it is theorized that the human brain likely uses a form of spatial-temporal encoding, but today's implementations cannot replicate this behavior at scale. A neuromorphic computer at brain-scale allows for network exploration and experimentation to gain more insight, especially into thorny topics like the interaction, at a global level, of multiple local learning rules that allow for local self-organization.

Computer scientists can draw upon the findings of neuroscientists about the biological brain to design similar artificial structures of neurons and synaptic connections, and they can then test those structures' efficacy and iterate for better performance. The PoC for this project developed at WSU did just that. It modeled a hierarchical structure of neurons and synapses based on the current understanding of the structure of the cerebral cortex, and it verified the model by simulating a simplified auditory cortex with 100 million neurons.[3]

Neuromorphic computing as a path forward for AI

In recent years, AI has been focused largely on deep learning (DL) neural networks. The era of DL discovery has waned, and the era of DL industrialization is now underway. Not since the advent of generative adversarial networks (GANs), around five years ago, have there been any real breakthrough discoveries in DL. Instead, advances in the field have been fueled largely by scale: ever more massive datasets used to train bigger neural nets running on supercomputers with increasing numbers of CPUs and graphics processing units (GPUs).

This path has reached a point of diminishing returns. Compute power is not maintaining its earlier rate of acceleration, while the costs in terms of hardware price and power consumption are becoming prohibitive for widespread research.[6] Moreover, the DL approach is inherently limited to creating more accurate predictive models in narrow, specialized subjects. While useful, this provides no progress toward the goal of inventing general AI, or true AI, that functions more like a biological brain. Unlike a DL supercomputer, a two-year-old's brain does not need to be trained on 10,000 images to recognize a cat. The next step in the evolution of AI is to break the neuromorphic code by creating hardware that emulates more closely the structures and functions of a biological brain.

The work at WSU is about simulating large-scale and structurally connected spiking neural networks (SNNs) using simple leaky integrate-and-fire (LIF) neurons. Neither of these concepts is new or unique to this project. LIF neurons are a well-established way to emulate in silicon the behavior of biological neurons that generate sharp electrical potentials, or "spikes," across their cell membranes. Likewise, SNNs have been created that model the way the brain works by emulating natural neural networks that exist in biological brains. Each artificial neuron in an SNN can fire independently of the others, sending pulsed signals to other neurons in the network that directly change the electrical states of those neurons.

What's different about this project is the technology it uses. The most successful SNN projects in the past have been built on application-specific integrated circuits (ASICs). SNNs built on CPUs and GPUs are painfully slow, so specially fabricated ASICs, such as the Intel® Loihi research chip, have been created to provide the performance needed for large-scale neuromorphic computing.

The downside of using ASICs is that they are extremely time consuming and expensive to design and fabricate. Once built, ASIC chips cannot be changed. Whatever hardware design was developed cannot be modified to explore different configurations. This is why the WSU project is revolutionary for the field—because it's built on FPGA technology, which means that the underlying hardware structures can be reconfigured to integrate new findings into the hardware platform as discovery continues.

Harnessing FPGAs for neuromorphic computing

The slow manufacturing cadence and rigid configuration of ASICs inhibits discovery. FPGAs allow hardware to be updated to test new theories and to match the latest discoveries. The goal is to create an open source solution based on off-the-shelf hardware that provides researchers and industry professionals a common platform to perform brain-inspired computation and discovery on a much wider scale.

An FPGA is an integrated circuit of programmable logic gates that, unlike CPU, GPU, or ASIC hardware, can be configured and reconfigured using software. This is important because how neurons in a brain are organized is not fully understood—it's a moving target that requires experimentation to emulate. Theories need to be tested and refined; and testing theories requires the ability to modify the underlying hardware. FPGA-based neuromorphic computing provides a configurable platform for research and discovery.

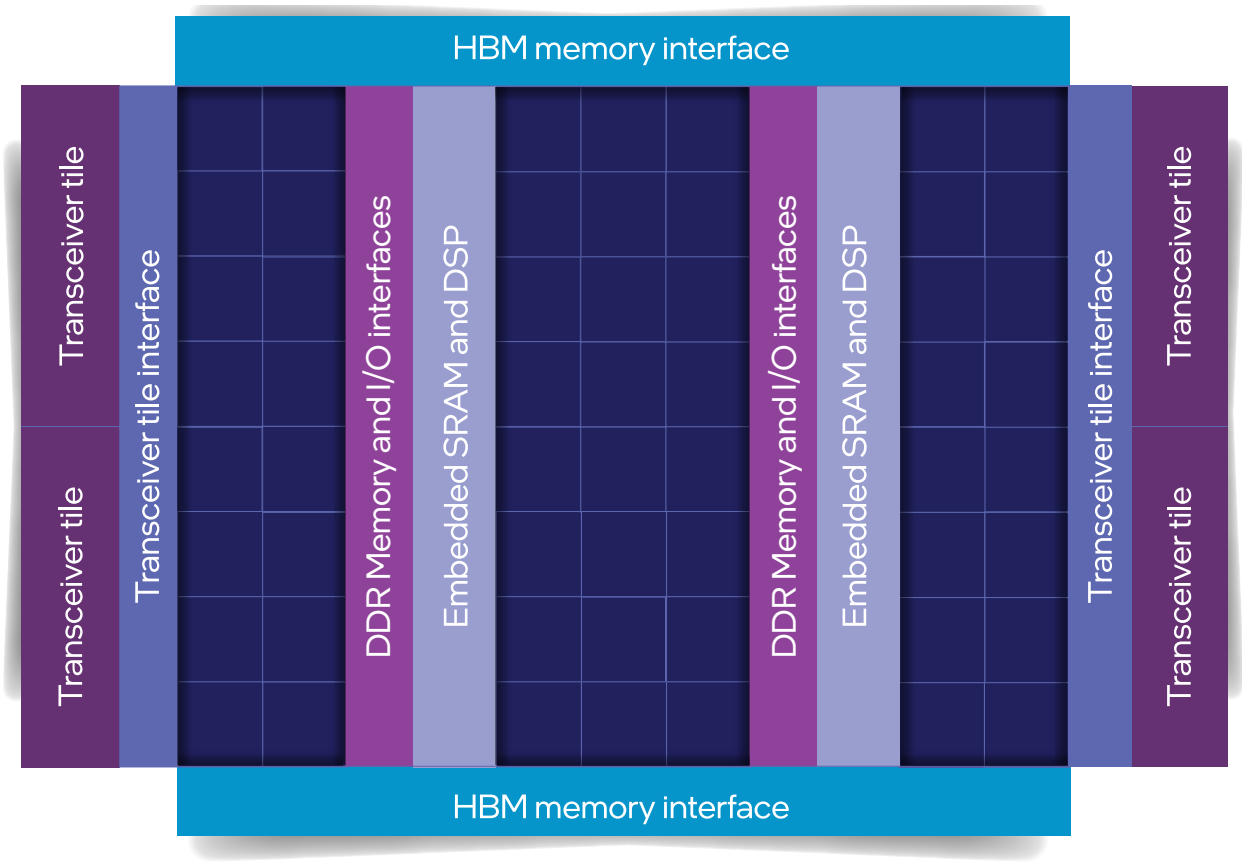

Figure 1. Elements of an FPGA

The PoC completed at WSU demonstrates how the programmability of FPGA hardware can be used to explore highly sophisticated neuromorphic computing models. Led by Mark Wang, the team used FPGA programmability to emulate a hierarchical pattern of synaptic connections found in the human cortex, as shown in Figure 2.

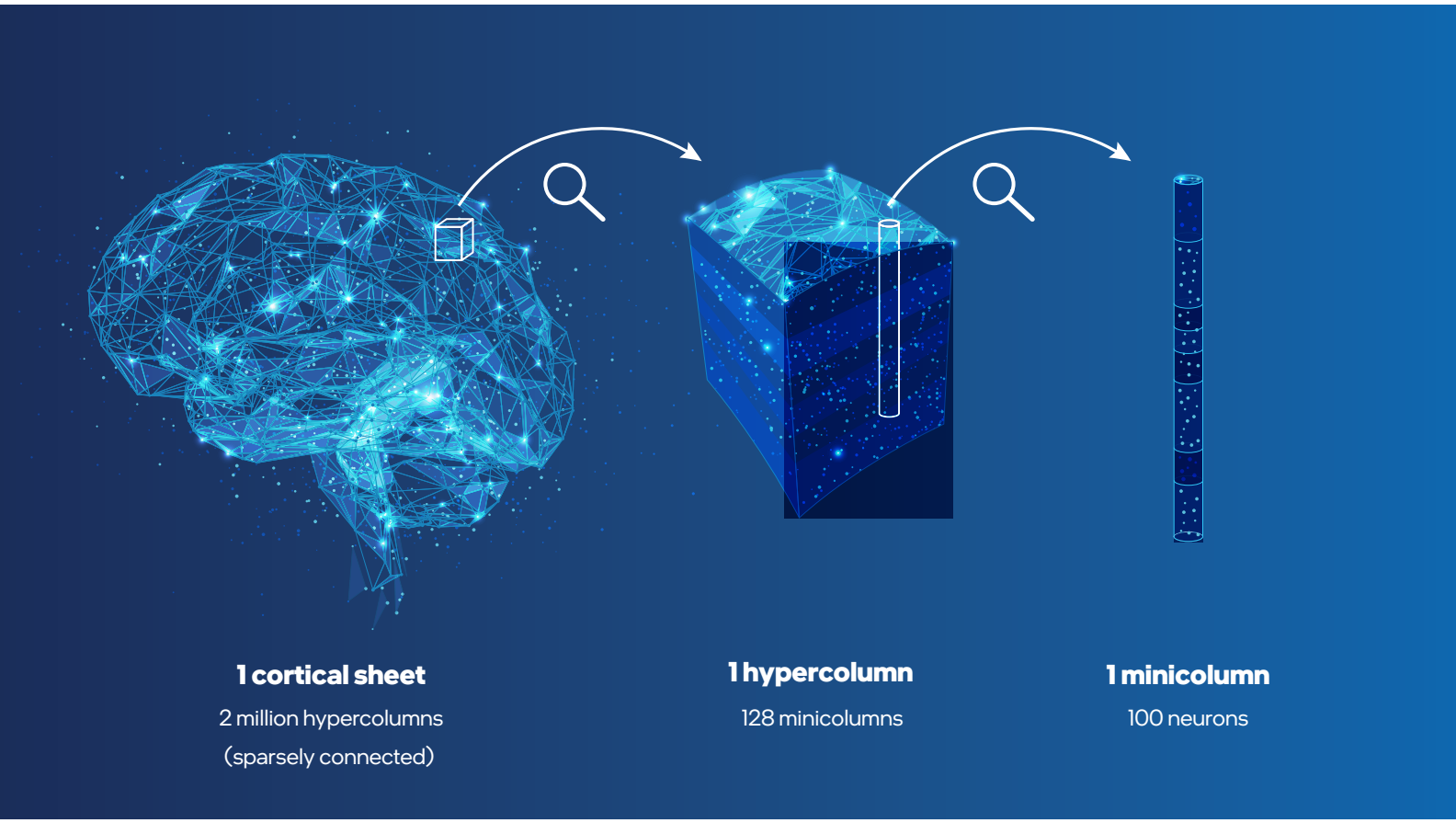

Figure 2. The functional building blocks of the neuromorphic compute system are a network of sparsely connected hypercolumns, each consisting of up to 128 connected minicolumns of 100 tightly connected neurons each[3,7]

The main novelty of Wang's work is the abstraction of a neuromorphic architecture into clusters represented by minicolumns and hypercolumns (as shown in Figure 2), analogous to the fundamental structural units observed in neurobiology. A hierarchical communication scheme allows one neuron to have a fan-out of up to 200,000 neurons—that is, the output from one neuron can be the input for up to 200,000 others.

> "Without this approach, simulating large-scale fully connected networks needs prohibitively large memory to store look-up tables for point-to-point connections. Instead, we use a novel architecture, based on the structural connectivity in the neocortex, such that all the required parameters and connections can be stored in on-chip memory. The cortex simulator can be easily reconfigured for simulating different neural networks without any change in hardware structure by programming the memory."[3]

The PoC was implemented on a single Intel Stratix V FPGA, and it was able to simulate spiking neural networks with up to 2.6 billion LIF neurons in real time. The current project will scale the platform up to 168 Intel Stratix 10 MX FPGAs in three server racks, and it will support simulating SNNs at a theoretical peak performance of 86 trillion synaptic operations per second—which is on par with the human cortex, estimated at 20–118 trillion synaptic operations per second.[8] This kind of scale is made possible by a number of advanced Intel technologies beyond just the FPGAs.

Key Intel technologies

FPGAs are the core technology powering this scalable neuromorphic computer. While configurability is the game changer, other FPGA features are also important to the project. For one thing, FPGA computing is fast—not as fast as an ASIC, but much faster than CPUs or GPUs when it comes to connecting billions of artificial neurons to each other in complex ways. FPGAs also offer greater density than ASICs, as they can use smaller, state-of-the-art technologies. And as off-the-shelf commercial products, FPGAs are substantially more affordable than ASICs.

Two Intel technologies are particularly important to the project: high-bandwidth memory (HBM) on the FPGA chip and the COPA technology.

High-bandwidth memory

HBM plays an important role in this neuromorphic computer. It is an in-package memory solution that represents a happy medium between lower capacity on-chip memory and lower bandwidth external memory. HBM provides enough capacity to implement the 2.1 million look-up tables that define the point-to-point connection patterns among the neurons, minicolumns, and hypercolumns in the brain-scale computer. And it provides enough bandwidth to access those look-up tables quickly—about 10 times more bandwidth from an HBM tile than from a traditional DDR4 DIMM.[9] An Intel Stratix 10 MX FPGA package with two HBM tiles delivers up to 512 GBps of memory bandwidth. It is generally quite difficult to exhaust this memory bandwidth, and the IP kernel developed at WSU is designed to take full advantage of the high bandwidth.

Storing and accessing 2.1 million look-up tables on external memory would be painfully slow. Using much faster HBM located right next to the FPGA chip beside the logic circuits provides the speed necessary to support computing at this scale.

COPA

FPGAs have not typically been used as autonomous nodes in distributed computing scenarios, in large part because there has been no standardized hardware/software infrastructure for such networks. Intel's COPA technology addresses this need by providing a customizable framework for connecting FPGAs into large clusters using a dedicated high-speed Ethernet network. COPA technology is a critical system building block for architecting large-scale FPGA systems by managing complex functionality across FPGA nodes, in addition to between an FPGA and a host computer.

The scalability provided by COPA technology is critical to the brain-machine project because it enables arbitrarily large clusters to handle large-scale neuromorphic workloads. COPA provides more than network connectivity. It tightly integrates acceleration as part of networking, thereby permitting data transformation and filtering operations to be performed directly on the data, both during transmit and receive. The framework is highly customizable and can be tailored for optimal performance. The software framework supports an open standards application programming interface (API), OpenFabrics Interfaces (OFI), with extensions for invoking the accelerator functions as part of network communication.

WSU Neuromorphic Computer Components

- The WSU machine will consist of three full-height racks, each with a 64-port 100 Gb HPC switch to connect to each of the COPA-enabled FPGAs, and a 24-port 10 Gb network switch.
- The three racks will house 42 2U host servers with 3rd Generation Intel® Xeon® Scalable processors, each with a total of 128 GB of main memory and 2 TB of solid state drive (SSD) storage.
- Each of the 42 servers will host 4 BittWare 520N-MX boards featuring Intel® Stratix® 10 MX2100 FPGAs with 16 GB integrated HBM2 memory.
- Projected thermal design power (TDP) will be 38.8 kW at full load.

A new era of scalable, accessible brain-inspired research

The team at WSU is collaborating with Intel on the world's first brain-scale neuromorphic computer using Intel FPGAs. This highly scalable neuromorphic compute system will be unique, but far from exclusive—the whole project is about reproducibility and accessibility.

One such machine is not expected to be enough to meet worldwide demand for access. That's why the WSU machine is designed to be replicated elsewhere. The environment is a regular data center, the hardware is available off the shelf, and the software stack is open source. The WSU machine thereby provides the blueprint and software stack for instantiating the technology at other universities and other data centers.

The WSU machine is the prototype for a common neuromorphic compute platform for the research community. The expectation is that, with the off-the-shelf component approach and an open software stack, similar machines can be replicated anywhere in the world, at any scale, to meet compute needs and to integrate new discoveries. Ideally, in the near future, these machines would live in the cloud. This will open the door to neuromorphic computing as a service, so that neuroscientists and neuromorphic engineers from around the world will be able to perform brain-scale neural network simulation by requesting compute time rather than hardware with their proposals.

The vision is that this project can dramatically accelerate the field of neuromorphic computing. Neuromorphic computers like this one—some smaller and some larger—will proliferate around the world, making neural network simulation ever more widely available and affordable. Discoveries and findings by a community of interdisciplinary researchers will be shared and verified on each other's substantially similar FPGA-based compute systems. And after the initial phase of research into AI and neuroscience, there will likely be commercial opportunities to apply new AI methods to various applications. This approach should just become a tool for ML and data analysis for various industries.

To be clear, FPGA-based machines are by no means the last word in neuromorphic computing. They are, rather, an important early step on a long journey. These machines will be powerful tools for research and discovery, to test theories and refine models, and to make breakthroughs in understanding how the brain works and how it can be simulated. After new models have been discovered, tested, and refined, then they can be used to build specialized and integrated processors with better performance characteristics.

Right now, the need is for a reconfigurable hardware platform that enables rapid iterations to incorporate new findings and make them available to the wider community. The FPGA-based neuromorphic compute system that WSU is building will be the first platform to meet that need, and a prototype for more to follow.

Learn more

Visit the International Centre for Neuromorphic Systems site at:
westernsydney.edu.au/icns

Read more about Intel FPGA products and technologies:
intel.com/content/www/us/en/products/programmable.html

Get details on the BittWare 520N-MX FPGA card:
bittware.com/fpga/520n-mx/

Learn about Configurable Network Protocol Accelerator (COPA) technology:
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9280342

[1] The lower estimates are from: G.M. Shepherd. *The Synaptic Organization of the Brain.* 1998. The higher estimates are from: C. Koch. *Biophysics of Computation: Information Processing in Single Neurons.* 1999.

[2] For a good introduction to FPGA technology, see: Intel. "Architecture All Access: Modern FPGA Architecture | Intel Technology." May 2021. https://youtu.be/EVy4KEj9kZg. Especially helpful is the city-planning metaphor starting at 12:56 in the video: an FPGA starts like a shell of a city with empty buildings and streets, and you get to decide which buildings serve which purpose and how the traffic flows among them.

[3] R.M. Wang, C.S. Thakur, and A. van Schaik. "An FPGA-Based Massively Parallel Neuromorphic Cortex Simulator." *Frontiers in Neuroscience.* April 2018. frontiersin.org/articles/10.3389/ fnins.2018.00213/full.

[4] Intel. "Intel Stratix 10 FPGAs & SoC FPGA." intel.com/content/www/us/en/products/details/fpga/stratix/10.html.

[5] Venkata Krishnan, Olivier Serres, and Michael Blocksome. "COnfigurable Network Protocol Accelerator (COPA): An Integrated Networking/Accelerator Hardware/Software Framework." *IEEE Symposium on High-Performance Interconnects.* August 2020. https://ieeexplore.ieee.org/document/9188286.

[6] Training GPT-3, for example, reportedly cost $12 million for a single training run. Source: Ala Shaabana. "The Future of AI is Decentralized." February 2021. https://towardsdatascience.com/the-future-of-ai-is-decentralized-848d4931a29a.

[7] Matthieu Thiboust. *Insights from the brain: The road towards Machine Intelligence.* April 2020. insightsfromthebrain.com/.

[8] Peter Lennie. "The Cost of Cortical Computation." *Current Biology.* March 2003. sciencedirect.com/science/article/pii/S0960982203001350.

[9] Traditional DDR4 DIMMs provide about 21 GBps bandwidth, while 1 HBM2 tile provides up to 256 GBps. Source: Intel. "Intel Stratix 10 MX FPGA." intel.com/content/www/us/en/products/details/fpga/stratix/10/mx.html.

Printed in USA 0821/SS/PRW/PDF Please Recycle 347936-001US