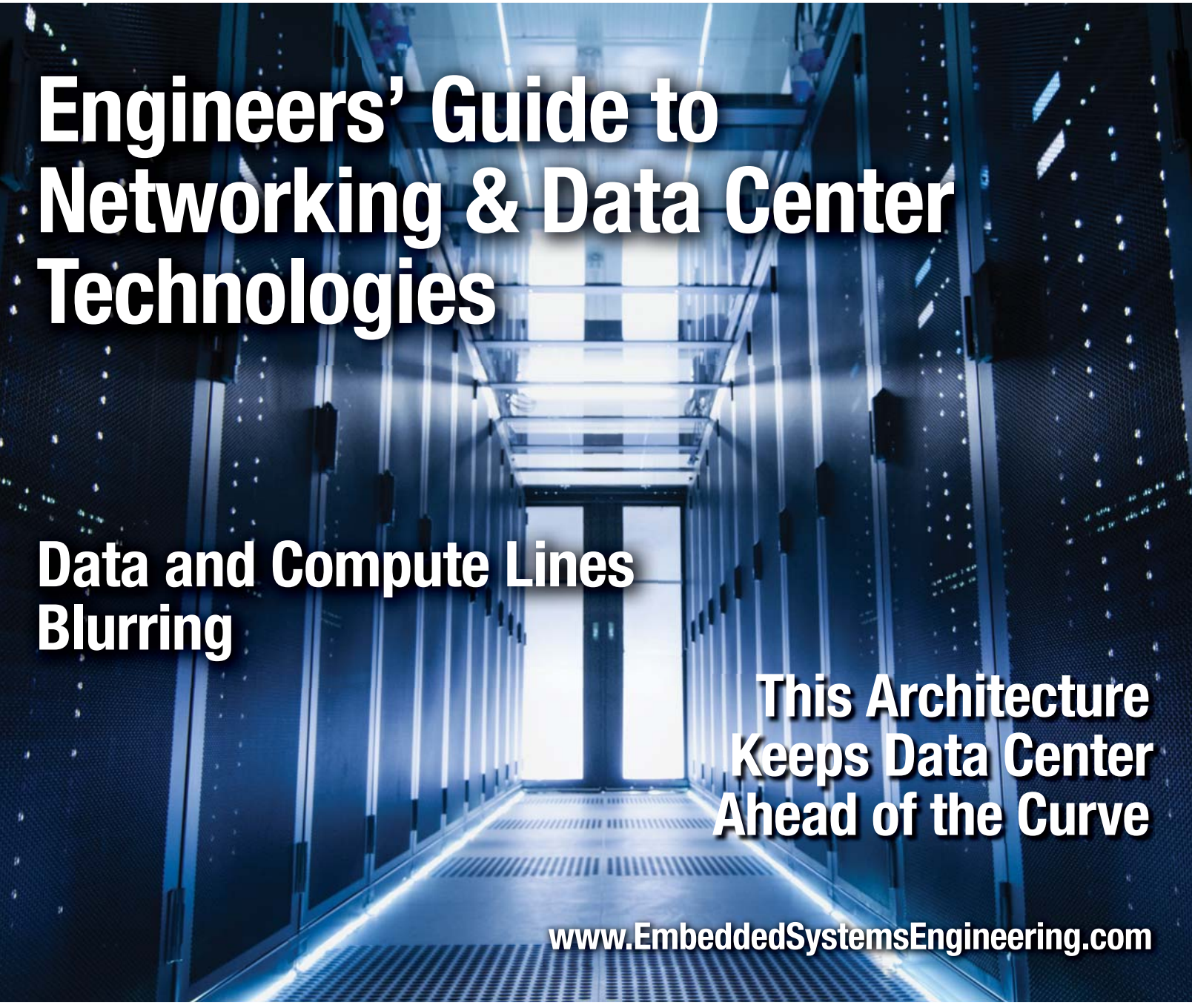
July 2018

EMBEDDED SYSTEMS ENGINEERING

powered by EECatalog

Guiding Embedded Designers on Systems and Technologies



Engineers' Guide to Networking & Data Center Technologies

Data and Compute Lines Blurring

This Architecture Keeps Data Center Ahead of the Curve

www.EmbeddedSystemsEngineering.com

CONTENTS

EMBEDDED SYSTEMS ENGINEERING

Special Features

Product Showcase

ENGINEERS' GUIDE TO NETWORKING & DATA CENTER TECHNOLOGIES 2018
www.embeddedsystemsengineering.com

Vice President & Publisher
Clair Bright

Editorial
Editor-in-Chief
Lynnette Reese | lreese@extensionmedia.com
Managing Editor
Anne Fisher | afisher@extensionmedia.com
Senior Editors
Caroline Hayes | chayes@extensionmedia.com
Dave Bursky | dbursky@extensionmedia.com
Pete Singer | psinger@extensionmedia.com
John Blyler | jblyler@extensionmedia.com

Creative / Production
Production Traffic Coordinator
Marjorie Sharp
Graphic Designer
Nicky Jacobson
Senior Web Developer
Slava Dotsenko

Advertising / Reprint Sales
Vice President, Sales
Embedded Electronics Media Group
Clair Bright
cbright@extensionmedia.com
(415) 255-0390 ext. 15

Marketing/Circulation
Jenna Johnson
jjohnson@extensionmedia.com

To Subscribe
www.eecatalog.com

Extension MEDIA

Extension Media, LLC Corporate Office
President and Publisher
Vince Ridley
vridley@extensionmedia.com
(415) 255-0390 ext. 18
Vice President & Publisher
Clair Bright
cbright@extensionmedia.com
Human Resources / Administration
Darla Rovetti

Special Thanks to Our Sponsor

ADVANTECH
Enabling an Intelligent Planet

Dolphin Interconnect Solutions

Embedded Systems Engineering is published by Extension Media LLC, 1786 18th Street, San Francisco, CA 94107.

Device Lending Enables Composable Architecture

Creating a composable infrastructure by leveraging the latest PCIe standard equates to something like…using pencils in space. Sometimes it makes sense to think up a simple solution that's merely crafty rather than succumb to the hype.

By Lynnette Reese, Editor-in-Chief, Embedded Systems Engineering

Keeping data center infrastructure ahead of rapidly increasing demands can get expensive. Real-time analytics, 5G connectivity, IoT, and Artificial Intelligence (AI) drive growth, but all this innovation is also pushing data centers to improve at a similar pace. For data centers, innovation brings a massive influx of data, shifting requirements, and insatiable business requirements. Data centers must flex while also meeting workload expectations, staying within an operating budget, maintaining efficiency, and leveraging innovation for a competitive edge within the data services market. The rise of hyperscale data centers, driven by big data, IoT, and AI, has massive networking loads supporting a considerable number of external and diverse clients. Hyperscale data centers illustrate the need for more efficient and flexible use of massive amounts of resources.

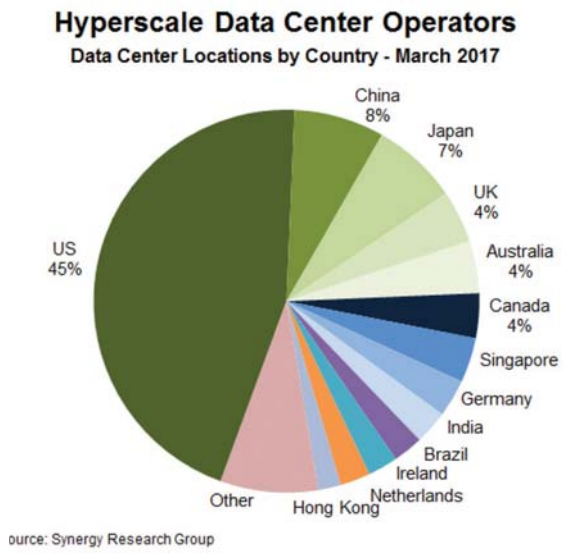


Figure 1: Hyperscale data center networks support many different external clients. The growth of big data with IoT and machine learning/AI are pushing data center infrastructure. As of March 2017, the majority of hyperscale data centers were operated in the U.S.

The volume and spectrum of cloud workloads add pressure that makes inflexibility a non-option. Traditional data architectures made up of servers, storage media, switches, and the like have been available in a large variety of form factors and sizes. The various pieces come together to serve a particular data workload in the data center. However, the workloads of today are changing rapidly, and traditional data center infrastructures cannot flex as fast as needed without adding many hours of labor. The complexity of traditional infrastructures has been mitigated somewhat by converged infrastructures, whereby the compute, storage, and networking fabric converge into a single solution to meet a particular workload. While converged infrastructures relieved hardware-centric challenges, they created another issue, as managing became workload-centric. Having started on the left, then swung far to the right, data center technology has found a sweet spot in composable architecture.

WHAT IS COMPOSABLE INFRASTRUCTURE?

Taking an application-centric approach, composable infrastructure is the answer to data center flexibility. Composable architecture is the next generation data center design—able to support rapidly changing system configurations, facilitate maximum sharing of both real and virtual infrastructure, and support new hardware technology. While similar to a converged infrastructure, composable infrastructure integrates compute, storage, and networking into a single platform by using a software-defined intelligence that maintains a pool of liquid resources. The application-centric composable infrastructure provides a new approach with which to provision and manage assets (both real and virtual). By using disaggregated programmable infrastructure as code, composable infrastructure seamlessly bridges software and hardware while eliminating management silos. The result is lower operating costs through "right-sizing," and a higher level of flexibility.

Several major companies already provide what is also referred to as composable disaggregated infrastructure or composable architecture. For instance, the Intel® Rack Scale Design (RSD) architecture is a com-

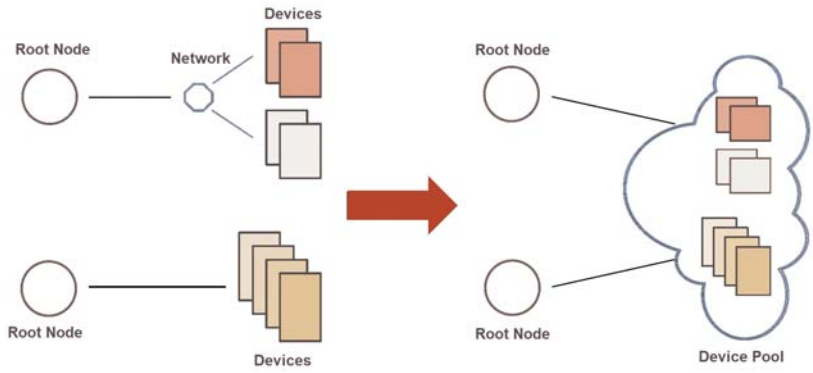


Figure 2: Composable architecture creates a pool of resources that can be deployed on demand in data centers.

posable disaggregated architecture where "hardware resources, such as compute modules, nonvolatile memory modules, hard disk storage modules, FPGA modules, and networking modules, can be installed individually within a rack. These can be packaged as blades, sleds, chassis, drawers or larger physical configurations."[1] Resources include a high-bandwidth data connection through an Ethernet link to a dedicated management network, and at least one Ethernet or other high-speed fabric, such as PCI Express (PCIe). Often two networks, including the management network, are connected to top-of-rack (ToR) switches. Historically, ToR switching has been adopted for rack-at-a-time flexibility, through modular data centers. A rack can connect through ToR to other racks to create a management domain referred to as a pod. Even with composable architecture, pods can be linked together in any network topology that best suits the data center.

A LOWER-COST WAY TO CREATE COMPOSABLE INFRASTRUCTURE

But there's another way to compose resources on the fly to meet changing workloads, and without requiring an Intel RSD compatible rack. Comprehensive rack scale solutions are not always possible due to budget constraints. For machines with access to resources via PCIe, a lower-cost solution can extend composable architecture to existing data centers. A Norwegian company called Dolphin Interconnect Solutions has an elegant solution called device lending. Device lending is a simple software solution that allows one to reconfigure systems and reallocate resources within a PCIe Fabric. Accelerators (GPUs and FPGAs), NVMe drives, network cards or other network fabric "can be added or removed without having to be physically installed in a particular system on the network." Dolphin's eXpressWare SmartIO software enables device lending, which creates seamless management of a pool of devices while maximizing resources. Device lending achieves both extremely low computing overhead and low latency without requiring any application-specific distribution mechanisms. A low-cost composable infrastructure is within reach, as a remote IO resource appears to applications as if local, with device lending software deployed in the PCIe Fabric. Dolphin has been involved with industry standards (including PCI, ASI, and PCIe) since the 1990s.

Device lending works transparently across PCIe connected racks and between servers and modules with no modifications to drivers, operating systems, or software applications. Device lending enables temporary access to a PCIe device located remotely over a PCIe network. Furthermore, performance in accessing a remote device is similar to accessing a local device, since there is no software overhead in the data transfers themselves. Devices are temporarily borrowed by any system within the fabric, and for as long as necessary. When a device is no longer needed, it can be returned to local use or allocated to another system. One can control the Dolphin device lending software using a set of command line tools and options, which can be used directly or integrated into any other higher-level resource management system, such one that might be used with an Intel RSD or different architecture. Dolphin's device lending software does not require any particular boot order or power-on sequence. PCIe devices borrowed from a remote system can be used as if they were local devices until returned. Furthermore, Dolphin's device lending strategy does not require explicit integration into a unified Application Programming Interface (API), since it works by taking advantage of the inherent properties of PCIe to accomplish a composable infrastructure. For more information about how device lending works with hot-adding (or hot-plugging), virtualization, non-transparent bridges (NTBs), IO Memory Management Units (IOMMUs), and DMA remapping, refer to the whitepaper, Device Lending in PCI Express Networks by Lars Kristiansen, et. al. (PDF).

Device lending is an advanced application of the PCIe standard. PCIe is a stable, standard technology that is widely implemented. PCIe is also set to reach 128 GBps in full-duplex mode over 16 lanes with Gen 5. PCIe Gen 5 will be backward-compatible to prior generations

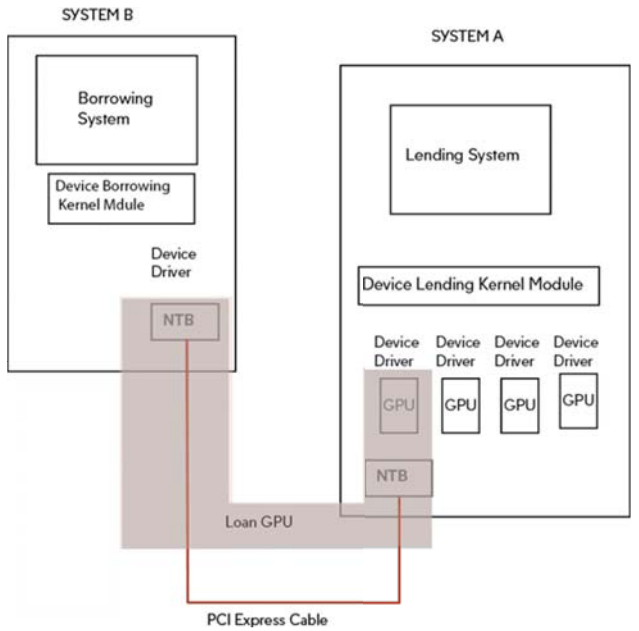


Figure 3: Device lending leverages the PCIe standard at the PCIe level on the stack, so integrating with other APIs, special bootloaders, and power sequencing are not needed. NTB= non-transparent bridge. (Source: dolphinics.com)

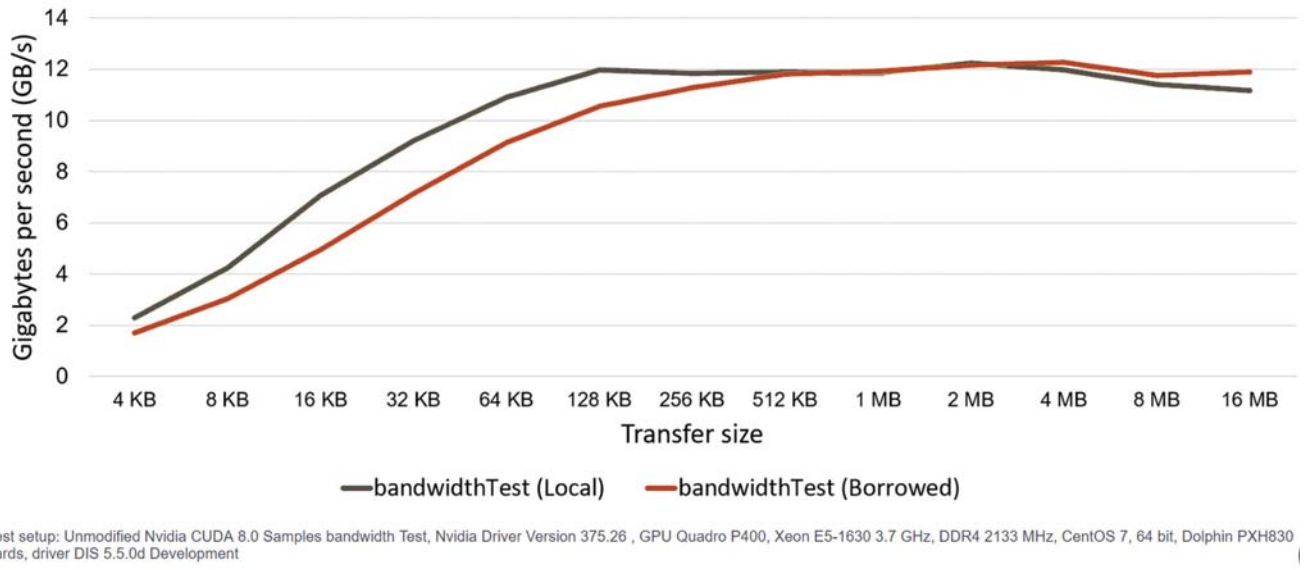


Figure 4: Comparison of bandwidth performance using device lending for a borrowed device (Borrowed) over a physically local device (Local). (Source: Dolphin ¬Interconnect Solutions)

and meet increasing performance needs. PCIe has latencies as low as 300ns end-to-end, dominates I/O bus technology, and has been prolific in the server, storage, mobile, and other markets. PCIe is also a significant player in connecting cloud-based devices that demand the highest performance in interconnects, such as GPU and FPGA accelerators for machine learning and AI. Hyperscale data centers number in the hundreds, with some of the world's most massive run by Google, Facebook, Amazon, and China's Baidu.

PERFORMANCE OF DEVICE LENDING FOR COMPOSABLE ARCHITECTURES

Device lending leverages the PCIe standard to achieve low latency and high bandwidth. Performance accessing a remote device will be very similar to a local device, limited only by the speed of PCIe over longer distances, if any. Dolphin's eXpressWare SmartIO device lending software does not require personnel to make changes to transparent devices or to a Linux kernel. Borrowed devices get inserted into the local device tree and the transparent device driver receives a "hot-plug" event signaling that a new resource is available. According to Dolphin, "If the transparent driver needs to re-map a DMA window, the re-map will be performed locally at the borrowing side, very similar to what happens in a virtualized system. The actual performance is system and device dependent." [2]

The Cisco, Intel, and Hewlett Packard Enterprise (HPE) strategies for accomplishing composable architectures are not that different from device lending software in that they achieve the same goals. HPE promises "a hybrid IT engine for your digital transformation," but is also software-defined.[3] Legend has it that it makes more sense to use a pencil in space rather than design a pen without gravity-fed ink. Urban legends aside, Dolphin's software is definitely a clever way to use the PCIe standard to a low-cost advantage for creating or complementing a composable architecture. The ability to break down fixed compute, storage, and networking fabric into a liquid pool of resources is more than desirable. Composing workloads on demand are making headlines in the IT world, and it doesn't have to have a fancy title to get the job done. Add device lending to the buildup of excitement about composable architecture for meeting the next level of flexibility in data centers.

RESOURCES

1. "Intel® Rack Scale Design (Intel® RSD) Architecture White Paper." Intel, www.intel.com/content/www/us/en/architecture-and-technology/rack-scale-design/rack-scale-design-architecture-white-paper.html.
2. Kristiansen, Lars, et al. "Device lending in PCI Express Networks." 13 May 2016, pp. 1–6., www.dolphinics.com/download/WHITEPAPERS/PCI_Express_device_lending_may_2016.pdf.
3. https://www.hpe.com/us/en/solutions/infrastructure/composable-infrastructure.html, accessed June 4, 2018.

Lynnette Reese is Editor-in-Chief, Embedded Intel Solutions and Embedded Systems Engineering, and has been working in various roles as an electrical engineer for over two decades. She is interested in open source software and hardware, the maker movement, and in increasing the number of women working in STEM so she has a greater chance of talking about something other than football at the water cooler.

"…the capability to store the entire trained model of the neural network…" Q&A with Sylvain Dubois, Crossbar

Autonomous driving is just one of the applications hungry for processing at the edge, giving embedded memory growing strategic importance.

By Anne Fisher, Managing Editor

Editor's Note: "The boundary between data and compute is really blurring now," contends Sylvain Dubois. The vice president of strategic marketing and business development at ReRAM technology company Crossbar also explains why putting data and computing on the same chip is making more and more sense. I spoke with Dubois in May, shortly before Crossbar unveiled its collaboration with Microsemi. Microsemi products manufactured at the 1x nm process node will integrate Crossbar's embedded ReRAM technology.

EECatalog: Across AI, networking, computing, we're seeing an increasing demand for embedded nonvolatile memory [NVM].

Sylvain Dubois, Crossbar: Yes, embedded memory is of strategic importance for CMOS foundries, and if you go to all of the top foundries' Technology Symposiums such as TSMC, Global Foundries, UMC, Samsung, SMIC, they are all looking for ways to have access to embedded NVM (Non Volatile Memory) technologies: Flash all the way to 40 nm and then MRAM and ReRAM for 2x nm and 1x nm.

EECatalog: How is what Crossbar and Microsemi will be doing—integrating embedded ReRAM at 1x nm—going to make a difference for OEMs and developers? Can you describe how a use case would change?

Dubois, Crossbar: A typical use case would involve bringing more computing power to the edge. More processing done locally, this includes wearables and hand-held devices, surveillance cameras and autonomous driving for example. And that brings up the whole topic of AI [Artificial Intelligence] inference at the edge, where you are not necessarily training the AI algorithms in the field but instead using the trained model so that the devices at the edge can recognize a face, a traffic sign. Crossbar's ReRAM technology will make a difference with any pattern recognition task such as object or face detection. It's what we demonstrated at the Embedded Vision Summit, showing how you can bring embedded ReRAM and neural networks together in a one-chip solution to make very low energy computing devices.

Today, what people are doing is trying to store the AI inference model, the weights and features of the neural network in the internal SRAM buffers on the chip. Because SRAM is not a dense memory; it won't be big enough and the models will be partially stored in external DRAM banks that are very expensive and also very power hungry. Both SRAM and DRAM are volatile memory, meaning that they lose their content when powered down. This requires an additional layer of flash memory required to store the model when power is off.

> *"Crossbar ReRAM is enabling a new range of energy-efficient computing architectures compared to legacy SRAM or DRAM-based architectures."*

But now with embedded ReRAM you have the capability to store the entire trained model of the neural network directly on chip. ReRAM retains its content for 10 years even when not powered, this eliminates the need for an external flash memory back-up and enables new use-models where the end-device can be frequently powered down and up to extend the battery life.

What we have designed is a specific memory array—a very wide memory array—with some amount of in-memory computing—pattern recognition, distance computation logic blocks. At the Embedded Vision Summit, we implemented a facial recognition demo showing a classification of a new face across a huge database of other faces in only once iteration.

EECatalog: How did the demonstration turn out?

Dubois, Crossbar: It was very well received as this classification task or comparison of one input across a huge database of objects usually takes a lot of time and power. The value proposition here is that the comparison of one input across a huge database will be extremely deterministic, it always takes the same amount of time whatever the size of the database from very few pictures to 100,000 pictures. The computation is done in only one iteration, few clock cycles.

EECatalog: How does that use case look if it's not being accomplished with ReRAM?

Dubois, Crossbar: Today, if you want to do the same use case with embedded SRAM and external or stacked DRAMs and GPUs, it will be done in a serial manner, where the larger the database is, the longer time it will take, because you have to compare against all these multiple pictures of objects stored in the memory.

We provide a very energy-efficient way—because ReRAM is on-chip and non-volatile —to perform classification of objects, patterns, with fast and deterministic latencies, consuming less energy than SRAM/DRAM memories.

And it's also very secure. Privacy matters when the database includes not only your face but also your biometrics and vocal commands. You don't want the whole conversation in your living room to be processed in the cloud and potentially hackable by malware. Biometrics identification, speech recognition and classification of objects from surveillance cameras are typical use cases for energy-efficient computing and memory on the same chip.

EECatalog: One of the big picture issues here ability to anticipate that next advanced process node and scale to it.

Dubois, Crossbar: Yes, it is important to pick a memory technology that scales because most of these AI chips or advanced SoCs, or microcontrollers, are currently designed at 22nm, 14nm, 12nm or even below 7 nm.

Crossbar ReRAM cells are programmed with a very low voltage across two electrodes causing the metal ions of the top electrode to move and thereby creating an extremely short narrow filament (3 or 4 nm). Growth of this metallic wire forms a conductive path, enabling a very low-resistance state. The ON current that is going through the filament determines the logic 1 state. When you want to have a logic 0 state, we just reverse the electric field so that the metal ions are pulled back to the top electrode, creating a high-resistance state, almost an open circuit.

Based on the metal filament physics that we grow and remove, the difference between ON and OFF current is extremely high, more than 1000X difference, providing great read margins and reliability to the ReRAM technology at the most advanced process nodes. As the filament is just 3 nm, going below 10 nm is something definitely possible with Crossbar ReRAM technology.

The ReRAM cell is so small that it can fit in between the metal routing layers of standard CMOS wafers. This is the reason why we can have a breakthrough architecture with millions of connection points between the logic and the memory compared to maximum thousands of connections with stacked DRAMs. It is a truly monolithic integration of embedded ReRAM and logic in the same silicon.

EECatalog: Anything to add before we wrap up?

Dubois, Crossbar: The boundary between data and compute is really blurring now. Algorithms trained with lots of data and devices are now self-sufficient to perform object identification and pattern recognition with a minimum power budget. Crossbar ReRAM is enabling a new range of energy-efficient computing architectures compared to legacy SRAM or DRAM-based architectures. Crossbar is working with multiple partners to create innovative architectures where data and processing are integrated on same silicon chip.

For edge computing in hand-held mobile devices or home appliances, or cloud computing in data centers, people are starting to realize that they can cut their energy bill quite drastically by putting the data and the computing in the same chip. Most of the system companies are now expanding their strategies towards vertical integration of their business all the way to the chip manufacturing as it makes a lot of sense for a great differentiation.

Advantech Multi-node Servers

SKY-9240

FLEXIBLE HYPER-CONVERGED AND HPC SERVER

- Up to 4 Hot-Swappable CPU Node boards with dual Intel® Xeon® Scalable Processors
- Up to 64x DIMM slots DDR4 2666MHz RDIMM/ LRDIMM 3DS
- Flexible I/O configurations with up to 8x PCIe Gen3 x16 slots (half length, low profile)
- Support for 1 PCIe Gen3 x8 OCP mezzanine per node.
- Optional redundant M.2 storage devices & four 2.5" SSD/HDD slot
- Up to 12x 3.5" SAS/SATA storage on the system
- Hot swappable and redundant AC PSU options
- Carrier Grade BMC (IPMI v2.0 compliant) with fail safe updates, Web Interface, KVM, Redfish

SKY-9240

SKY-9340

HIGH PERFORMANCE BLADE SERVER WITH INTEGRATED DATA PLANE FABRIC

- Converged system with the highest CPU, I/O, Storage density per RU in the market
- Up to 4 hot swappable CPU blades with dual Intel® Xeon® Scalable Processors
- More than 800Gbps of I/O bandwidth
- Integrated data plane fabric for low latency interconnect between nodes
- Flexible I/O configurations with up to 8x PCIe card
- Storage expansion with up to 24x 2.5"/12x 3.5" SAS/ SATA/NVMe front hot swappable drives
- Hot swappable and redundant AC or DC PSU options
- Front-to-Rear Push-Pull Cooling Mode. Two rear pluggable, hot swappable fan modules with fan speed control
- Integrated system management

SKY-9340

CONTACT INFORMATION

ADVANTECH
Enabling an Intelligent Planet

Advantech Co. Ltd
No.33, Lane 365, Yang Guang St.,
Neihu Dist., Taipei, 11491,
Taiwan, R.O.C.
Tel: +886-2-2972-7818
http://www.advantech.com/nc