FREE Download www.aiworld.be

aiworld

Artificial Intelligence UPDATE

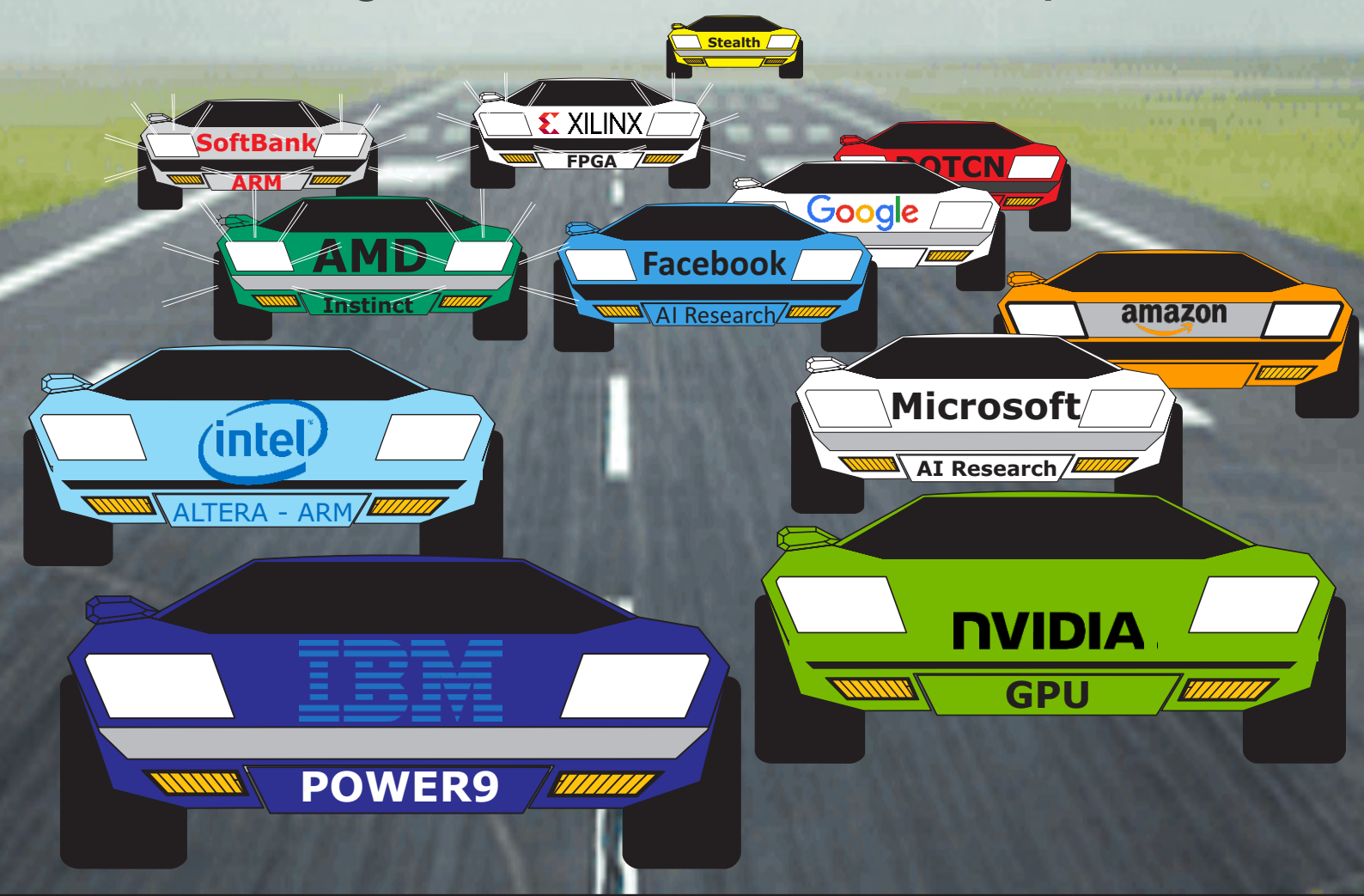
Machine Learning - Deep Learning - Software Research
Products - Technology - Applications - Industry News

AI World is a Global Publication of e2mos

Q1-2017

Artificial Intelligence is Back this Time for Real

The Race is On
Big Calibers in the 1st Lap



aiworld

a NEW e-magazine
but not just another one

Focus and Dedication - Information Selected for You – Time Saving

Artificial Intelligence is back, this time for real and needs full attention.
Unfortunately everyday there is a lot of information of which the majority is of no use and/or duplication.

This NEW e-magazine will focus on key information that you need to know for your business, just like for our other four e-magazines, see on this page bottom right.

Previously A.I. items were published in those other magazines

Dedicated Web Site for aiworld

Same way as for our other magazines

First Edition – 2017 Q1
Summary Key Players

In this first edition of 20 pages (normally 8 or 12) we cover: Nvidia, Intel, MS, Google, GE, ARM, AMD, Cavium, ENEA, Facebook, KU Leuven, Xilinx

1,000+ e-magazines Published

So far we have published over 1,000 magazines for Hi-Tech from Chips to Rack Space Systems, in the mean time we know what our readers are looking for.

Thank you for any suggestion you may have
Enjoy it
Daniel Dierickx

Daniel Dierickx
CEO & co-Founder
at e2mos
Acting Chief Editor

Dear Reader,

Here is your free copy of **AI World**, one of our five e-magazines published by e2mos.

Our aim is to provide you with relevant information directly in relation with your activity.

Those magazines are part of the e2mos « Go-to-Market Platform »

This GLOBAL Platform is a UNIQUE Set of Services for Telecom ICT, Video Broadcast, Embedded Computing, IoT and AI Vendors from Multicore Chips to Application-ready Systems & Rack Space Servers.

Our WORLDWIDE Services include:
- Business Discovery
- Customer Meeting Setup
- Telemarketing
- Call Campaigns
- e-mailings Worldwide
- and our 5 e-magazines, each magazines has its own Website (see below).

It is all based on:
- 30+ Years Customer Relationship and Market & Technology Expertise
- our PREMIER Database started in 1980 and maintained EVERY DAY using many sources and research.

Thank you, Daniel Dierickx

Editor/Publisher:
e2mos www.e2mos.com
Contact mgt@e2mos.com

FREE just Click on the LOGO

aiworld
IoT World
Telecom COTS World
Broadband Broadcast IoT Convergence
Embedded Systems World
ATCA World

Microsoft expands artificial intelligence (AI) efforts with creation of new Microsoft AI and Research Group

Computer vision luminary Harry Shum to lead more than 5,000 people worldwide

REDMOND, Washington — Sept. 29, 2016 — Microsoft Corp. announced on Thursday it has formed the Microsoft AI and Research Group, bringing together Microsoft's world-class research organization with more than 5,000 computer scientists and engineers focused on the company's AI product efforts. **The new group will be led by computer vision luminary Harry Shum**, a 20-year Microsoft veteran whose career has spanned leadership roles across Microsoft Research and Bing engineering.

Microsoft is dedicated to democratizing AI for every person and organization, making it more accessible and valuable to everyone and ultimately enabling new ways to solve some of society's toughest challenges. Today's announcement builds on the company's deep focus on AI and will accelerate the delivery of new capabilities to customers across agents, apps, services and infrastructure.

In addition to Shum's existing leadership team, several of the company's engineering leaders and teams will join the newly formed group including Information Platform, Cortana and Bing, and Ambient Computing and Robotics teams led by David Ku, Derrick Connell and Vijay Mital, respectively. All combined, the Microsoft AI and Research Group will encompass AI product engineering, basic and applied research labs, and New Experiences and Technologies (NExT).

"We live in a time when digital technology is transforming our lives, businesses and the world, but also generating an exponential growth in data and information," said Satya Nadella, CEO, Microsoft. "At Microsoft, we are focused on empowering both people and organizations, by democratizing access to intelligence to help solve our most pressing challenges. To do this, we are infusing AI into everything we deliver across our computing platforms and experiences."

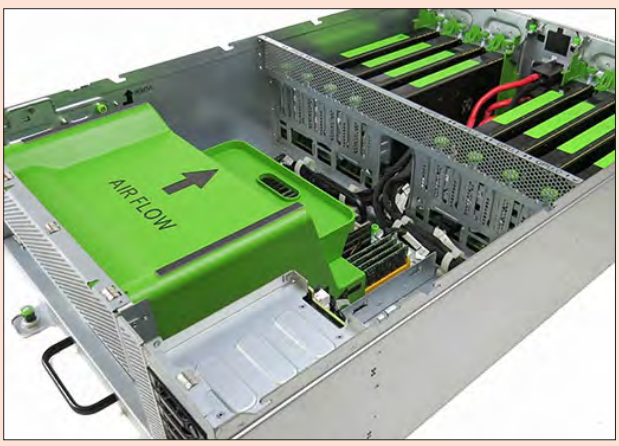
"Microsoft has been working in artificial intelligence since the beginning of Microsoft Research, and yet we've only begun to scratch the surface of what's possible," said Shum, executive vice president of the Microsoft AI and Research Group. "Today's move signifies Microsoft's commitment to deploying intelligent technology and democratizing AI in a way that changes our lives and the world around us for the better. We will significantly expand our efforts to empower people and organizations to achieve more with our tools, our software and services, and our powerful, global-scale cloud computing capabilities."

Microsoft is taking a four-pronged approach to its initiative to democratize AI:
- **Agents.** Harness AI to fundamentally change human and computer interaction through agents such as Microsoft's digital personal assistant Cortana
- **Applications.** Infuse every application, from the photo app on people's phones to Skype and Office 365, with intelligence
- **Services.** Make these same intelligent capabilities that are infused in Microsoft's apps —cognitive capabilities such as vision and speech, and machine analytics — available to every application developer in the world
- **Infrastructure.** Build the world's most powerful AI supercomputer with Azure and make it available to anyone, to enable people and organizations to harness its power. More information about this approach can be found Here.

For 25 years, Microsoft Research has contributed to advancing the state-of-the-art of computing through its groundbreaking basic and applied research that has been shared openly with the industry and academic communities, and with product groups within Microsoft. The organization has contributed innovative technologies to nearly every product and service Microsoft has produced in this timeframe, from Office and Xbox to HoloLens and Windows. More recently, Shum has expanded the organization's mission to include the incubation of disruptive technologies and new businesses. **MORE:** Click Here **Microsoft AI and R&D is hiring** for positions in labs & offices worldwide Click Here

Facebook has offered 22 GPU Supercomputers to European Universities for A.I Research

Two machines have been allocated to KU Leuven in Belgium



ARM extends HPC offering with acquisition of software tools provider Allinea Software

Highlights:

- ARM has acquired Allinea Software Limited ("Allinea"), a leading provider of software tools for HPC
- The acquisition strengthens ARM's HPC offering by extending its product portfolio for development tools to HPC, machine learning and data analytics markets
- ARM will continue to develop, enhance, and invest in Allinea products with support for multiple CPU architectures

Cambridge, UK – December 16, 2016 – ARM has acquired Allinea Software, an industry leader in development and performance analysis tools that maximize the efficiency of software for high performance computing (HPC) systems. Currently, 80 percent of the world's top 25 supercomputers use Allinea's tools, with key customers including the US Department of Energy, NASA, a range of supercomputing national labs and universities, and private companies using HPC systems for their own scientific computation.

"As systems and servers grow in complexity, developers in HPC are facing new challenges that require advanced tools designed to enable them to continue to innovate," said Javier Orensanz, general manager, development solutions group, ARM. "Allinea's ability to debug and analyze many-node systems is unique, and with this acquisition we are ensuring that this capability remains available to the whole ARM ecosystem, and to the other CPU architectures prevalent in HPC, as well as in future applications such as artificial intelligence, machine learning and advanced data analytics."

This acquisition further enhances ARM's long-term growth strategy in HPC and builds on ARM's recent success with Fujitsu's 64-bit ARM®v8-A powered Post K supercomputer, and the launch of the ARMv8-A Scalable Vector Extension. It follows the announcement that ARMv8-A will be the first alternative architecture with OpenHPC support, and the release of ARM Performance Libraries, which provide ease of software development and portability to ARMv8-A server platforms. As this momentum continues, bringing Allinea's expertise into ARM will continue to enable partners with access to a comprehensive software tools suite that address increasingly complex system challenges.

"Writing and deploying software that exploits the ever increasing computing power of clusters and supercomputers is a demanding challenge - it needs to run fast, and run right, and that's exactly what our suite of tools is designed to enable," said David Lecomber, CEO, Allinea. "As part of ARM, we'll continue to work with the HPC community, our customers and our partners to advance the development of our cross-platform technology, and take advantage of product synergies between ARM's compilers, libraries and advisory tools and our existing and future debugging and analysis tools. Our combined expertise and understanding of the challenges this market faces will deliver new solutions to this growing ecosystem."

Allinea's unique tools provide developers with the ability to deal with systems with hundreds, thousands (and hundreds of thousands) of cores. The product suite includes the developer tool suite Allinea Forge, which incorporates an application debugger called Allinea DDT and a performance analyzer called Allinea MAP, and an analysis tool for system owners, users and administrators called Allinea Performance Reports.

Allinea will be integrated into the ARM business with all functions and Allinea's Warwick and Eastleigh locations retained. Allinea's former CEO David Lecomber will join the ARM development solutions group management team.

Source: ARM https://www.arm.com/about/newsroom/arm-extends-hpc-offering-with-acquisition-of-software-tools-provider-allinea-software.php

About ARM

ARM technology is at the heart of a computing and connectivity revolution that is transforming the way people live and businesses operate. From the unmissable to the invisible; our advanced, energy-efficient processor designs are enabling the intelligence in 90 billion silicon chips and securely powering products from the sensor to the smartphone to the supercomputer. With more than 1,000 technology partners including the world's most famous business and consumer brands, we are driving ARM innovation into all areas compute is happening inside the chip, the network and the cloud. www.arm.com

ARM is now a SoftBank company

Further to the Tokyo Stock Exchange announcements listed below, SoftBank Group Corp. ("SBG") is pleased to announce that the Scheme of Arrangement in respect of the recommended acquisition (the "Acquisition") of ARM Holdings plc ("ARM") by SBG came into effect on September 5, 2016 (GMT), and that the entire issued and to be issued share capital of ARM is now owned by SBG and its wholly-owned subsidiaries. Pursuant to the terms of the Acquisition, SBG purchased all of ARM's issued and to be issued shares (excluding any ARM shares already owned by SBG or an SBG subsidiary) for cash, for a total acquisition price amounting to approximately GBP 24.0 billion (USD 31.0 billion or JPY 3.3 trillion). http://www.softbank.jp/en/corp/news/press/sb/2016/20160905_01/

Announcing GPUs for Google Cloud Platform

Tuesday, November 15, 2016
Posted by John Barrus, Product Manager, Google Cloud Platform

CPU-based machines in the cloud are terrific for general purpose computing, but certain tasks such as rendering or large-scale simulations are much faster on specialized processors. Graphics Processing Units (GPUs) contain hundreds of times as many computational cores as CPUs and are great at accelerating risk analysis, studying molecular binding or optimizing the shape of a turbine blade. If your CPU-based instance feels like a Formula One race car but you're in need of a rocket, you're going to love our new cloud GPUs.

Early in 2017, Google Cloud Platform will offer GPUs worldwide for Google Compute Engine and Google Cloud Machine Learning users. Complex medical analysis, financial calculations, seismic/subsurface exploration, machine learning, video rendering, transcoding and scientific simulations are just some of the applications that can benefit from the highly parallel compute power of GPUs. GPUs in Google Cloud give you the freedom to focus on solving challenging computational problems while accessing GPU-equipped machines from anywhere. Whether you need GPUs for a few hours or several weeks, we've got you covered.

Google Cloud will offer AMD FirePro S9300 x2 that supports powerful, GPU-based remote workstations. We'll also offer NVIDIA® Tesla® P100 and K80 GPUs for deep learning, AI and HPC applications that require powerful computation and analysis. GPUs are offered in passthrough mode to provide bare metal performance. Up to 8 GPU dies can be attached per VM instance including custom machine types.

GPUs on Google Cloud

AMD
FirePro S9300 x2

NVIDIA
Tesla P100

NVIDIA
Tesla K80s

Google Cloud GPUs give you the flexibility to mix and match infrastructure. You'll be able to attach up to 8 GPU dies to any non-shared-core machine, whether you're using an n1-highmem-8 instance with 3 TB of super-fast Local SSD or a custom 28 vCPU virtual machine with 180 GB of RAM. Like our VMs, GPUs will be priced per minute and GPU instances can be up and running within minutes from Google Cloud Console or from the gcloud command line. Whether you need one or dozens of instances, you only pay for what you use.

During an early access program, customers have been running machine learning training, seismic analysis, simulations and visualization on GPU instances. **Startup MapD** gets excellent results with a GPU-accelerated database.
"These new instances of GPUs in the Google Cloud offer extraordinary performance advantages over comparable CPU-based systems and underscore the inflection point we are seeing in computing today. Using standard analytical queries on the 1.2 billion row NYC taxi dataset, we found that a single Google n1-highmem-32 instance with 8 attached K80 dies is on average 85 times faster than Impala running on a cluster of 6 nodes each with 32 vCPUs. Further, the innovative SSD storage configuration via NVME further reduced cold load times by a factor of five. This performance offers tremendous flexibility for enterprises interested in millisecond speed at over billions of rows."
- Todd Mostak, MapD Founder and CEO https://www.mapd.com/

The Foundry, a visual effects software provider for the entertainment industry has been experimenting with workstations in the cloud.

"At The Foundry, we're really excited about VFX in the cloud, and with the arrival of GPUs on **Google Cloud Platform**, we'll have access to the cutting edge of visualisation technology, available on-demand and charged by the minute. The potential ramifications for our industry are enormous.."
- Simon Pickles, Lead Engineer, Pipeline-in-the-Cloud

MORE: Click Here

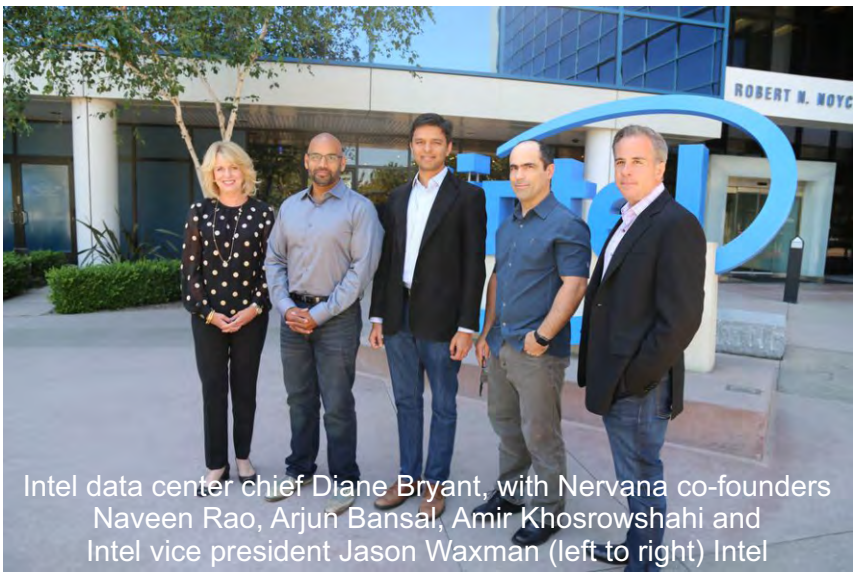
Intel is paying more than $400 million to buy deep-learning startup Nervana Systems

The chip giant is betting that machine learning is going to be a big deal in the data center.

Today, we're excited to announce the planned acquisition of Nervana by Intel*. With this acquisition, Intel is formally committing to pushing the forefront of AI technologies. Nervana intends to continue all existing development efforts including the Nervana Neon deep learning framework, Nervana deep learning platform, and the Nervana Engine deep learning hardware. The combination of Nervana's technology and expertise incorporated into Intel's portfolio will take deep learning/AI solutions to the next level. We will continue to operate out of our San Diego Headquarters and will retain our talent, brand, and start-up mentality.

Nervana started with the idea that we can engineer better solutions for computation by bringing together computer engineering, neuroscience, and machine learning. Amir Khosrowshahi and I brought this idea to Bruno Olshausen (Amir's PhD advisor) before founding Nervana. Bruno said (and I'm paraphrasing here), "Do it. Do it now! You might already be too late." This was the catalyst that pushed us over the edge to pursue the idea fully. At this point we invited Arjun Bansal to join the founding team and Nervana was born. Amir then sought advice from his cousin Ali Partovi, co-founder of Code.org and iLike and a well-connected angel investor. Ali became our first investor and trusted advisor to the company. He made introductions to his network and we pulled together a very high-caliber set of seed investors. Soon after, Carey Kloss and Andrew Yang (close friends and ex coworkers of mine) joined Nervana to lead the hardware development efforts. We're proud to say that in just 2.5 years, we pushed the performance envelope in deep learning and will soon have a revolutionary new architecture to push it even further.

We've always been a mission driven company. Even though we won't be a startup any longer, our mission hasn't changed: we are here to make a dent in the world of computation. With this deal, we can now shatter the old paradigm and move into a new regime of computing. We'll look back in 10 years and see this time as the inflection point of when compute architectures became neural. The semiconductor integrated circuit is one of humanity's crowning achievements and Intel has the best semiconductor technology in the world. Nervana's AI expertise combined with Intel's capabilities and huge market reach will allow us to realize our vision and create something truly special.



Intel data center chief Diane Bryant, with Nervana co-founders Naveen Rao, Arjun Bansal, Amir Khosrowshahi and Intel vice president Jason Waxman (left to right) Intel

Having strong investors who believed in the company was key. We were very fortunate to have board members Steve Jurvetson (DFJ) lead our Series A and Matt Ocko (DCVC) our Series A1. Their inherent interest in innovation and deep technical knowledge was the force that made this company happen. In addition, Playground Global, CME Ventures, Lux Capital, Allen & Co, AME Cloud Ventures, Fuel Capital, Omidyar Technology Ventures, SV Angel and our seed investors were great partners with which to build a company.

Many thanks to Allen & Company LLC who acted as exclusive financial advisor to Nervana in this transaction.

Sincerely,

Naveen Rao, CEO and Co-founder of Nervana

A.I. beating humans? Not Now!

Here are two articles about Google's Diane Greene

A.I. beating humans? Not in my lifetime says Google cloud chief

By Martyn Williams | 23 Nov 2016

The head of Google's cloud business says she doesn't expect machine intelligence to exceed that of humans during her lifetime, despite recent rapid progress that has surprised many.

Diane Greene, who turns 61 this year, said that while researchers are making strides in programming intelligence into computers, there's still a long way to go.

"There is a lot that machine learning doesn't do that humans can do really, really well," she said last Tuesday at the Code Enterprise conference in San Francisco.

Her remarks came hours after Google said Greene's division had hired two leading machine learning and artificial intelligence experts: Fei-Fei Li, who was director of AI at Stanford University, and Jia Li, who headed up research at Snap, the operator of SnapChat.

"Nobody expected some of the advances we are seeing as quickly as we're seeing them," she said, "but, the singularity I don't see it in my sentient lifetime."

Greene had been asked to evaluate, on a scale from one to 10, how close the industry was to "the singularity" -- the moment, forecast by technologist Ray Kurzweil, when machine intelligence would go beyond that of humans.

Greene never got around to putting a number on the current state of research. But she did acknowledge that some people would lose their jobs as machine learning became more useful.

"I think it's really incumbent on us to get the education out there to make sure everyone is digitally literate, because that's where the divide is, and because if you're digitally literate, you're going to have jobs," she said.

At least for now, there aren't enough qualified people in the job market for the number of research jobs available, but that situation's unlikely to last, so Silicon Valley is being asked tough questions about where this push into machine intelligence will take the world.

Greene said the industry was taking the matter seriously but offered only small efforts to solve it, like supplying Chromebooks to schools. **MORE:** Click Here

Video: Full INTERVIEW Click Here

Google's Diane Greene says the singularity won't arrive in her sentient lifetime

Three out of four Google robots disagreed.

by April Glaser@aprilaser Nov 15, 2016

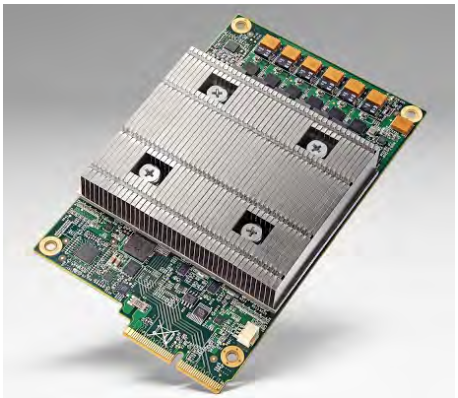
Google's head of engineering, Ray Kurzweil, is known for espousing what's known as the "technological singularity": The idea that artificial intelligence will become so smart that it will take on a form that humans cannot foresee nor comprehend. It's essentially the end of humanity as we know it.

Thankfully, at least one top Google exec doesn't think it's going to happen anytime soon.
Full article Click Here

Google supercharges machine learning tasks with TPU custom chip

By Norm Jouppi, Distinguished Hardware Engineer, Google

Machine learning provides the underlying oomph to many of Google's most-loved applications. In fact, more than 100 teams are currently using machine learning at Google today, from Street View, to Inbox Smart Reply, to voice search.



But one thing we know to be true at Google: great software shines brightest with great hardware underneath. That's why we started a stealthy project at Google several years ago to see what we could accomplish with our own custom accelerators for machine learning applications.

The result is called a **Tensor Processing Unit (TPU), a custom ASIC** we built specifically for machine learning — and tailored for TensorFlow. We've been running TPUs inside our data centers for more than a year, and have found them to deliver an order of magnitude better-optimized performance per watt for machine learning. This is roughly equivalent to fast-forwarding technology about seven years into the future (three generations of Moore's Law).

TPU is tailored to machine learning applications, allowing the chip to be more tolerant of reduced computational precision, which means it requires fewer transistors per operation. Because of this, we can squeeze more operations per second into the silicon, use more sophisticated and powerful machine learning models and apply these models more quickly, so users get more intelligent results more rapidly. A board with a TPU fits into a hard disk drive slot in our data center racks.

TPU is an example of how fast we turn research into practice — from first tested silicon, the team had them up and running applications at speed in our data centers within 22 days.

TPUs already power many applications at Google, including RankBrain, used to improve the relevancy of search results and Street View, to improve the accuracy and quality of our maps and navigation. AlphaGo was powered by TPUs in the matches against Go world champion, Lee Sedol, enabling it to "think" much faster and look farther ahead between moves.

Our goal is to lead the industry on machine learning and make that innovation available to our customers. Building TPUs into our infrastructure stack will allow us to bring the power of Google to developers across software like TensorFlow and Cloud Machine Learning with advanced acceleration capabilities. Machine Learning is transforming how developers build intelligent applications that benefit customers and consumers, and we're excited to see the possibilities come to life.

MORE:
Click Here

MultiCore is beautiful with a MultiCore RTOS

ENEA OSE enables REAL-TIME Acceleration for LINUX on Embedded Multicore Devices

Configuring Linux for hard real-time latency requirements is demanding, but Enea has extensive expertise within real-time operating systems, Inter Process Communication (IPC), Linux, and multicore technologies, and the ability to supply the market with efficient real-time acceleration of Linux based systems.

A heterogeneous multicore system, running both Linux and Enea OSE brings the best of both worlds – the rich ecosystem of Linux combined with the real-time characteristics of OSE. Depending upon the use case and the number of cores available, Linux and Enea OSE are optimized to meet the requirements with an appropriate mix of SMP and/or AMP configurations.

Support for OpenAMP simplifies the configuration and deployment of both operating systems & applications.

Enea OSE supports all major architectures, e.g ARM, x86, MIPS & PowerPC, both in 32-bit and 64-bit mode.

MORE: Click Here

GE acquires Wise.io to deepen its machine learning stack

Posted Nov 15, 2016 by Frederic Lardinois, Writer at TechCrunch (@fredericl)

GE Digital today announced that it has acquired **Wise.io**, a machine-learning powered service that helps businesses find patterns and trends in their vast data stores. At first glance, that may seem like an odd acquisition for a company like GE. It's important to keep in mind, though, that with **Predix**, GE already offers its customers a service that focuses on helping them monitor their equipment, whether that's an industrial tool or an aircraft engine, and predict issues based on the monitoring data.

As GE CIO Jim Fowler told me, the company's developers wrote a few hundred different models for lots of different assets in Predix. The service has some rudimentary AI capabilities, but the addition of Wise.io's technology — which can find patterns on its own — and its team will allow GE to offer a far more flexible model. "As we think about services going forward, you'll see GE enter verticals we haven't been in before," Fowler told me.

In many ways, GE itself is undergoing the kind of digital transformation that many of its own customers are dealing with. It's undergoing what Fowler called a "new industrial revolution," where it has to reposition itself to play in the field of analytics, data, machine learning, etc. To keep pace, the company has been making a number of acquisitions. These include the purchase of **ServiceMax** for almost $1 billion yesterday and **Wise.io** today (as well as the rather quiet acquisition of data integration platform **Bit Stew** earlier in 2016).

GE and Wise.io did not disclose the acquisition price. Wise.io previously raised $3.58 million. The company also participated in the Alchemist Accelerator and Citrix Startup Accelerator programs.

Wise.io will continue to operate as usual and GE will continue to service its existing customers. While the company is mostly buying Wise.io for its technology and talent, this also means it's getting access to quite a few Wise.io customers who probably never considered GE as a vendor. Wise.io's users currently include the likes of Pinterest, Twilio, Thumbtack and Republic Wireless.

GE expects that software will bring in $15 billion in revenue by 2020 (I wonder if it used Predix to set that number…). The vast majority of this ($10 billion) will likely come from existing customers, $1 billion will come from the efficiencies it hopes to gain by driving internal usage of technologies like Wise.io and it expects that $4 billion will come from new customers.

About Predix - A Cloud-based Operating System
The Industrial Internet: Digital Transformation Starts Here
Predix, the platform for the Industrial Internet, is powering digital industrial businesses that drive the global economy. By connecting industrial equipment, analyzing data, and delivering real-time insights, Predix-based apps are unleashing new levels of performance of both GE and non-GE assets.

Video Click Here

GE Digital

Wise.io + GE Digital

servicemax From GE Digital

BITSTEW SYSTEMS From GE Digital

Intel to Acquire Movidius: Accelerating Computer Vision through RealSense for the Next Wave of Computing

By Josh Walden, Josh Walden is SVP & GM Intel New Technology Group (Left on the picture)

Combined with Intel's Existing Assets, Movidius Technology – for New Devices Like Drones, Robots, Virtual Reality Headsets and More – Positions Intel to Lead in Providing Computer Vision and Deep Learning Solutions from the Device to the Cloud

We're entering an era where devices must be smart and connected. When a device is capable of understanding and responding to its environment, entirely new and unprecedented solutions present themselves.

As part of our RealSense™ vision and strategy, we built and acquired critical technologies to ensure our leadership in computer vision and perceptual computing. Simply put, computer vision enables machines to visually process and understand their surroundings. Cameras serve as the "eyes" of the device, the central processing unit is the "brain," and a vision processor is the "visual cortex." Upon integration, computer vision enables navigation and mapping, collision avoidance, tracking, object recognition, inspection analytics and more – capabilities that are extremely compelling in emerging markets.

With the introduction of RealSense™ depth-sensing cameras, we brought groundbreaking technology that allowed devices to "see" the world in three dimensions. To amplify this paradigm shift, we completed several acquisitions in machine learning, deep learning and cognitive computing to build a suite of capabilities that open an entirely new world of possibilities: from recognizing objects, to understanding scenes; from authentication to tracking and navigating. This said, as devices become smarter and more distributed, we recognize that specific System on a Chip (SoC) attributes will be paramount to giving human-like sight to the 50 billion connected devices projected by 2020.

For this reason, I'm excited to announce our pending acquisition of Movidius*. With Movidius, Intel gains low-power, high-performance SoC platforms for accelerating computer vision applications. Additionally, this acquisition brings algorithms tuned for deep learning, depth processing, navigation and mapping, and natural interactions, as well as broad expertise in embedded computer vision and machine intelligence. Movidius' technology optimizes, enhances and brings RealSense™ capabilities to fruition.

We see massive potential for Movidius to accelerate our initiatives in new and emerging technologies. The ability to track, navigate, map and recognize both scenes and objects using Movidius' low-power and high-performance SoCs opens opportunities in areas where heat, battery life and form factors are key. Specifically, we will look to deploy the technology across our efforts in augmented, virtual and merged reality (AR/VR/MR), drones, robotics, digital security cameras and beyond. Movidius' market-leading family of computer vision SoCs complements Intel's RealSense™ offerings in addition to our broader IP and product roadmap.

Computer vision will trigger a Cambrian explosion of compute, with Intel at the forefront of this new wave of computing, enabled by RealSense™ in conjunction with Movidius and our full suite of perceptual computing technologies.

intel + Movidius

Movidius + Intel = Vision for the Future of Autonomous Devices

By Remi El-Ouazzane, CEO of Movidius (Right on the picture)

Today, I'm excited to announce the planned acquisition of Movidius by Intel. Movidius' mission is to give the power of sight to machines. As part of Intel, we'll remain focused on this mission, but with the technology and resources to innovate faster and execute at scale. We will continue to operate with the same eagerness to invent and the same customer-focus attitude that we're known for, and we will retain Movidius talent and the start-up mentality that we have demonstrated over the years.

Our leading VPU (Vision Processing Unit) platform for on-device vision processing combined with Intel's industry leading depth sensing solution (Intel® RealSense™ Technology) is a winning combination for autonomous machines that can see in 3D, understand their surroundings and navigate accordingly. Today, we're working with customers like DJI, FLIR, Google and Lenovo to give sight to smart devices including drones, security cameras, AR/VR headsets and more. But today's smart devices, while compelling, offer just a glimpse of what's to come.

When computers can see, they can become autonomous and that's just the beginning. We're on the cusp of big breakthroughs in artificial intelligence. In the years ahead, we'll see new types of autonomous machines with more advanced capabilities as we make progress on one of the most difficult challenges of AI: getting our devices not just to see, but also to think.

Movidius has been attacking this challenge at the device level - combining advanced algorithms with dedicated low-power hardware. At Intel, we'll be part of a team that is attacking this challenge from the cloud, through the network and on the device. This is very exciting.

Intel Unveils Strategy for State-of-the-Art Artificial Intelligence

Intel Offers Broad Portfolio Spanning Data Center to IoT Devices and Software to Make AI Foundational to Business and Society

NEWS HIGHLIGHTS
Intel announces AI strategy to drive breakthrough performance, democratize access and maximize societal benefits.
Intel introduces industry's most comprehensive data center compute portfolio for AI: the new Intel® Nervana™ platform.

Intel aims to deliver up to 100x reduction in the time to train a deep learning model over the next three years compared to GPU solutions.

Intel reinforces commitment to an open AI ecosystem through an array of developer tools built for ease of use and cross-compatibility, laying the foundation for greater innovation.

SAN FRANCISCO, Nov. 17, 2016 – Intel Corporation today announced a range of new products, technologies and investments from the edge to the data center to help expand and accelerate the growth of artificial intelligence (AI). Intel sees AI transforming the way businesses operate and how people engage with the world. Intel is assembling the broadest set of technology options to drive AI capabilities in everything from smart factories and drones to sports, fraud detection and autonomous cars.
At an industry gathering led by Intel CEO Brian Krzanich, Intel shared how both the promise and complexities of AI require an extensive set of leading technologies to choose from and an ecosystem that can scale beyond early adopters. As algorithms become complex and required data sets grow, Krzanich said Intel has the assets and know-how required to drive this computing transformation.

In a blog Krzanich said: "Intel is uniquely capable of enabling and accelerating the promise of AI. Intel is committed to AI and is making major investments in technology and developer resources to advance AI for business and society."

Intel's Robust AI Platform

Intel announced plans to usher in the industry's most comprehensive portfolio for AI – the Intel® Nervana™ platform. Built for speed and ease of use, the Intel Nervana portfolio is the foundation for highly optimized AI solutions, enabling more data professionals to solve the world's biggest challenges on industry standard technology.
Today, Intel powers 97 percent of data center servers running AI workloads and offers the most flexible, yet performance-optimized, portfolio of solutions. This includes Intel® Xeon® processors and Intel® Xeon Phi™ processors to more workload-optimized accelerators, including FPGAs (field-programmable gate arrays) and the technology innovations acquired from Nervana.

Press kit: Intel Artificial Intelligence: Unleashing the Next Wave

Intel also provided details of where the breakthrough technology from Nervana will be integrated into the product roadmap. Intel will test first silicon (code-named "Lake Crest") in the first half of 2017 and will make it available to key customers later in the year. In addition, Intel announced a new product (code-named "Knights Crest") on the roadmap that tightly integrates best-in-class Intel Xeon processors with the technology from Nervana. Lake Crest is optimized specifically for neural networks to deliver the highest performance for deep learning and offers unprecedented compute density with a high-bandwidth interconnect.

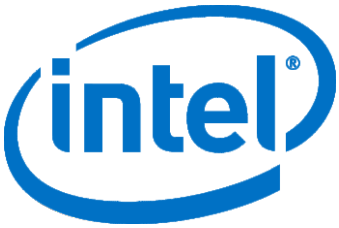
"We expect the Intel Nervana platform to produce breakthrough performance and dramatic reductions in the time to train complex neural networks," said Diane Bryant, executive vice president and general manager of the Data Center Group at Intel. "Before the end of the decade, Intel will deliver a 100-fold increase in performance that will turbocharge the pace of innovation in the emerging deep learning space."

Bryant also announced that Intel expects the next generation of Intel Xeon Phi processors (code-named "Knights Mill") will deliver up to 4x better performance1 than the previous generation for deep learning and will be available in 2017. In addition, Intel announced it is shipping a preliminary version of the next generation of Intel Xeon processors (code-named "Skylake") to select cloud service providers. With AVX-512, an integrated acceleration advancement, these Intel Xeon processors will significantly boost the performance of inference for machine learning workloads. Additional capabilities and configurations will be available when the platform family launches in mid-2017 to meet the full breadth of customer segments and requirements.

... to next page page

Intel Unveils Strategy for State-of-the-Art Artificial Intelligence

... from previous page



Enabling AI Everywhere and Cloud Alliance with Google*

Aside from silicon, Intel highlighted other AI assets, including Intel Saffron Technology™, a leading solution for customers looking for business insights. The Saffron Technology platform leverages memory-based reasoning techniques and transparent analysis of heterogeneous data. This technology is also particularly well-suited to small devices, making intelligent local analytics possible across IoT and helping advance state-of the-art collaborative AI.

To simplify deployment everywhere, Intel also delivers common, intelligent APIs that extend across Intel's distributed portfolio of processors from edge to cloud, as well as embedded technologies such as Intel® RealSense™ cameras and Movidius* vision processing units (VPUs).

Intel and Google announced a strategic alliance to help enterprise IT deliver an open, flexible and secure multi-cloud infrastructure for their businesses. The collaboration includes technology integrations focused on Kubernetes* (containers), machine learning, security and IoT.

To further AI research and strategy, Intel announced the formation of the Intel Nervana AI board, which will feature leading industry and academic thought leaders. Intel announced four founding members: Yoshua Bengio (University of Montreal), Bruno Olshausen (UC Berkeley), Jan Rabaey (UC Berkeley) and Ron Dror (Stanford University).

Additionally, Intel is working to make AI truly accessible. To help accomplish this, Intel has introduced the Intel Nervana AI Academy for broad developer access to training and tools. Intel also introduced the Intel Nervana Graph Compiler to accelerate deep learning frameworks on Intel silicon.

In conjunction with the AI Academy, Intel announced a partnership with global leading education provider Coursera* to provide a series of AI online courses to the academic community. Intel also launched a Kaggle Competition (coming in January) jointly with Mobile ODT* where the academic community can put their AI skills to the test to solve real-world socioeconomic problems, such as early detection for cervical cancer in developing countries through the use of AI for soft tissue imaging.

"Intel can offer crucial technologies to drive the AI revolution, but ultimately we must work together as an industry – and as a society – to achieve the ultimate potential of AI," said Doug Fisher, senior vice president and general manager of the Software and Services Group at Intel.

With the addition of the new edge and data center products, as well as the enablement programs, Intel has the full complement of technologies and ecosystem reach required to deliver the scale and promise of AI for everyone.

AI for the Betterment of Society

Lastly, Intel showcased some of the initiatives the company is investing in and partnering with to help maximize the positive impact of AI on the world. They include:
Intel is committing $25 million to the Broad Institute* to drive high-performance computing for genomics analytics. Through a five-year collaboration, researchers and software engineers at the Intel-Broad Center for Genomic Data Engineering will build, optimize and widely share new tools and infrastructure that will help scientists integrate and process genomic data. The project aims to optimize best practices in hardware and software for genome analytics to make it possible to access and use research data sets that reside on private, public and hybrid clouds.
Intel is a founding partner of Hack Harassment*, a cooperative effort with the mission of reducing the prevalence and severity of online harassment. The initiative is evaluating AI technology as a tool in this effort and is working to develop an intelligent algorithm to detect and deter online harassment. Over time, this capability will be released as an open source API that can be used in a variety of applications.
Intel is also a key partner of the National Center for Missing & Exploited Children* (NCMEC), a nonprofit whose mission is to help find missing children, reduce child sexual exploitation and prevent child victimization. Intel is providing AI technology and advising the center with the goal of accelerating the critical work of NCMEC's analysts to respond to reports of child sexual exploitation.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

Microsoft's CEO thinks AI should help people, not replace them

By James Walker, Jan 17, 2017 in Technology

Microsoft CEO Satya Nadella has spoken out on how artificial intelligence should not be seen as a way of replacing humans.

The leader made the comments as Microsoft enters new fields of AI research and prepares to 'democratise' the technology.

Nadella wants to see artificial intelligence aid people to do more with the time available to them. This isn't the same as having their job replaced by a robot, something he disagrees with both outside and within Microsoft. He cautioned against creating "parlor tricks" that risk demonstrating AI's power at the expense of human positions.

"The fundamental need of every person is to be able to use their time more effectively, not to say, 'let us replace you,'" Bloomberg reports Nadella said this week during an interview at the DLD conference in Munich. "This year and the next will be the key to democratizing AI. The most exciting thing to me is not just our own promise of AI as exhibited by these products, but to take that capability and put it in the hands of every developer and every organization."

The comments may reassure people concerned about losing their jobs to AI. A recent study by the World Economic Forum suggested 5.1 million positions could be taken over by technology within the next five years. This month, a Japanese insurance firm replaced 34 jobs with AI capable of thinking like a human. Other insurance companies have already introduced bots to their workplaces.

Nadella called for a set of guidelines to be developed that set out expectations around the use of artificial intelligence technology. He said this would help to preserve human positions while encouraging trust in new machines. It's still early days in the field of AI ethics though. Microsoft will need to work with its industry competitors to develop an AI code of conduct.

Like its rivals, Microsoft is already actively using AI in many of its flagship products. It powers its digital assistant Cortana and aids developments on its Azure cloud computing platform. The company has also introduced AI to Skype, Microsoft Translator and areas of Windows 10.

Recently, Microsoft hinted at its next moves in the field. It has acquired Maluuba, a Montreal-based startup that uses deep learning to aid natural language processing. Harry Shum, executive vice president of Microsoft's AI & Research group, said the company is "skating to where the puck will be next" through the acquisition.

"We've recently set new milestones for speech and image recognition using deep learning techniques, and with this acquisition we are, as Wayne Gretzky would say, skating to where the puck will be next — machine reading and writing," said Shum.

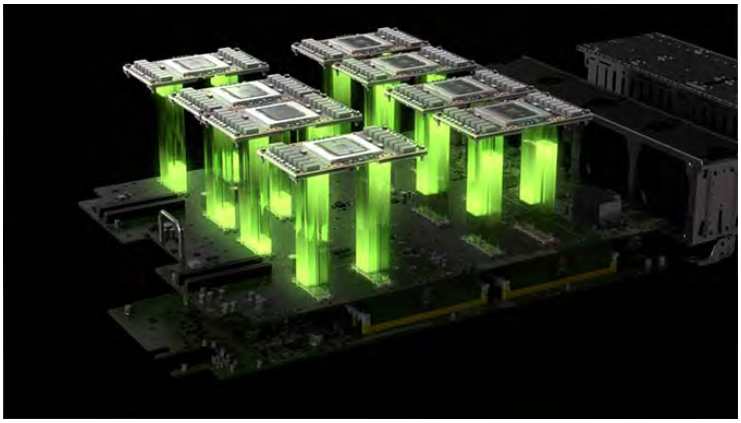
Nadella's desire to democratise AI could inspire his industry counterparts to begin their own investigations into AI ethics. A successful collaborative effort will be required to set the standards for what robots can do, allowing humans to approach emerging technologies optimistically without fearing for their jobs. Currently, development is still being prioritised ahead of thought, a trend which may need to change so AI can be accepted.

More about Microsoft, Satya Nadella, Artificial intelligence, Ai, machine learning Click Here

Nvidia Sees Bright Future for AI Supercomputing

By Tiffany Trader - November 23, 2016

Graphics chipmaker Nvidia made a strong showing at SC16 in Salt Lake City last week. Most prominent wins were achieving the number one spot on the Green500 list with new in-house DGX-1 supercomputer, SaturnV, and partnering with the National Cancer Institute, the U.S. Department of Energy (DOE) and several national laboratories to accelerate cancer research as part of the Cancer Moonshot initiative.



The company kicked off its SC activities with a press briefing on Monday (Nov. 14), during which CEO Jen-Hsun Huang characterized 2016 as a tipping point for the GPU computing approach popularized by Nvidia for over a decade.

Not surprisingly, Huang's main message was that the GPU computing era has arrived. Throughout the hour-long talk, Huang would revisit the theme of deep learning as both a supercomputing problem and a supercomputing opportunity.

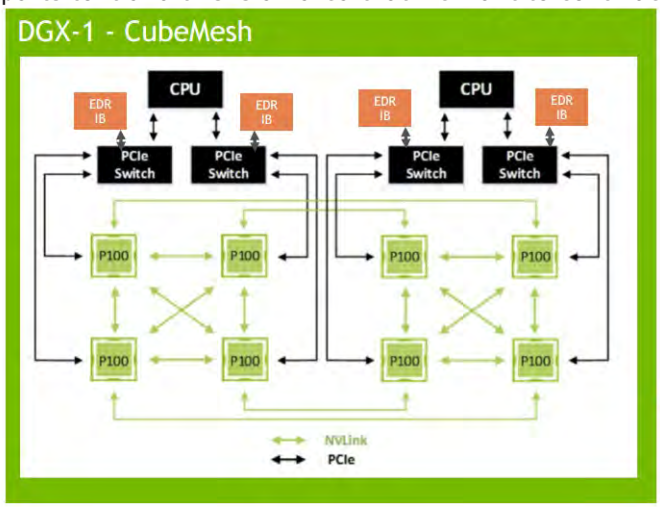
"We believe that supercomputers ought to be designed as AI supercomputers – meaning it has to be good at both computational science as well as data science – that building a machine that's only good at data science doesn't make sense and building a supercomputer that's only good at computational science doesn't make sense," he said.

"On the one hand, deep learning requires an enormous amount of data throughput processing – this way of developing software where the computers write software themselves inspired by a lot of data processing behind it is a very important approach to computing but it also has the wonderful opportunity to benefit supercomputing as well, solving problems for science that hasn't been possible before today," said Huang.

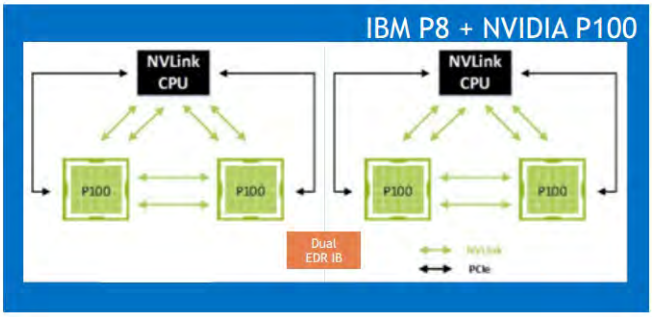
Huang's view is that traditional numerical HPC is not going anywhere, but will exist side by side with machine learning methods.

"I'm a big fan of using math when you can; we should use AI when you can't," he said. "For example what's the equation of a cat? It's probably very similar to the equation for a dog – two ears, four legs, a tail. And so there are a lot of areas where equations don't work and that's where I see AI – search problems, recommendation problems, likelihood problems, where there's either too much data, incomplete data, or no laws of physics that support it. So where do I feel like eating tonight – there's no laws of physics for that. There's a lot of these type of problems that we simply can't solve – I think that they're going to coexist."

While Nvidia is enabling parallel computing via thousands of CUDA cores combined with the CUDA programing framework, the CEO emphasized the necessity of a performant central processing unit. "Almost everything we do we start with a strong CPU," said Huang. "We still believe in Amdahl's law; we believe that code has a lot of single threaded parts to it and this is an area that we want to continue to be good at."



- 20GB/s per link
- 80GB/s + 80 GB/s in/out per GPU
- Very lightweight Protocol
- Used automatically with current GPU code
- Several systems being developed



NVDIA NVLINK

http://nvidia.com/nvlink

... to next page

Nvidia Sees Bright Future for AI Supercomputing

... from previous page

The two servers currently shipping with the NVLink P100 GPU – Nvidia's DGX-1 server and IBM's Minsky platform – speak to this goal. The DGX-1 connects eight NVLink'd Pascal P100s to two 20-core Intel Xeon E5-2698 v4 chips. The IBM Minsky server leverages two Power8 CPUs and four P100 GPUs connected by NVlink up to the CPUs.

Nvidia's 124-node supercomputer, SaturnV plays a crucial role in Nvidia's plans to usher in AI supercomputing. The machine debuted on the 48th TOP500 list at number 28 with 3.3 petaflops Linpack (4.9 petaflops peak). Even more impressively, it nabbed the number one spot on the Green500 list achieving more than 8.17 gigaflops/watt. That's a 42 percent improvement from the 6.67 gigaflops/watt delivered by the most efficient machine on the previous TOP500 list. Extrapolating to exascale gives us 105.7 MW. If we go with a semi-"relaxed" exascale power allowance of 30 MW (the original DARPA target was 20 MW), this is less than one-fourth the planned power consumption of US exascale systems. Three years ago, the extrapolated delta was over a 7X.

SaturnV – its name inspired by the original Moonshot – will be a critical part of the CANDLE (CANcer Distributed Learning Environment) project (covered here). Announced last month, CANDLE's mission is to exploit high performance computing (HPC), machine learning and data analytics technologies to advance precision oncology. Huang said the partners will be working together to develop "the world's first deep learning framework designed for exascale."

"It's going to be really hard," he added. "That's why we're working with the four DOE labs and have all standardized on the same architecture – SaturnV is the biggest one of them but we're all using exactly the same architecture and it's all GPU accelerated and we're going to develop a framework that allows us to scale to get to exascale."

Huang noted that when you apply deep learning FLOPS math – aka 16-bit floating point operations as opposed to the HPC norm of 64-bit FLOPS, exascale is not far away at all.

The [IBM/Nvidia] CORAL machines are on track for 2018 with 300 petaflops peak FP64, which comes out to 1,200 peak FP16, Huang pointed out. "For AI, FP16 is fine, now in some areas we need FP32, we need variable precision, but that's the point," he said. "I think CORAL is going to be the world's fastest AI supercomputer [and] I think that we didn't know it then but I believe that we are building an exascale machine already."

It's a fair point that dialing down the bits increases data throughput (boosting FLOPS), but as one analyst at the event said, "calling it exascale is changing the rules."

Lending more insight to Nvidia's plans was Solutions Architect Louis Capps, who presented at the Green500 BoF on November 16.

"This is completely a research platform," he said of SaturnV. "We're going to have academics using it. We're going to have partnerships, collaborations, and internally, we're working on our deep learning research and our HPC research."

Embedded, robotics, automotive, and hyperscale computing are all major focus areas, but Capps and Huang both were most effusive about the opportunities at the convergence of data science and HPC. "We're just now starting to bridge where real HPC work is converging with deep learning," said Capps.

SaturnV is organized into five 3U boxes per rack, with 15 kilowatt of power on each rack and some 25 racks total. While the press photo of SaturnV indicates 10 servers per rack, this is not reflective of what's inside. "We could not put that many in ours," said Capps. "We put this in a datacenter which is not HPC. It was an IT datacenter originally."

SaturnV was one of two systems on the newly published TOP500 list to employ the Pascal-based P100 GPUs. The number two greenest super, Piz Daint is using the PCIe variants. Installed at the Swiss National Supercomputing Centre, Piz Daint delivers an energy-efficiency rating of 7.45 gigaflops/watt. Refreshed with the new P100 hardware, Piz Daint achieved 9.8 petaflops on the Linpack benchmark, securing it the eighth spot on the latest list.

Notably, every single one of the top ten systems on the Green500 list is using some flavor of acceleration or manycore. There is no pure-play traditional x86 in the bunch.

... to next page

Nvidia Sees Bright Future for AI Supercomputing

... from previous page

A compelling testament to this approach came from Thomas Schulthess, director of the Swiss National Supercomputing Centre, where Nvidia K80 GPUs have been used for operational weather forecasting for over a year now. "I know the HPC community has a problem with the heterogeneous approach," he said. "We've done a lot of analysis on this issue. We asked, what would the goals we have at exascale look like if we build a homogeneous Xeon-based system, and there's no way that you will run significant problems that are significantly bigger and faster than we do today in 5-6 years at exascale if you build it based on a Xeon system.

"The message to the application folks is, you've had time to think about it now, but now there is no more choice. If you want to run at exascale, it is going to be on Xeon Phi or GPU-accelerated or the lightweight core, almost Cell-like architectures that we see on TaihuLight."

Green500 List for November 2016

Listed below are the November 2016 The Green500's energy-efficient supercomputers ranked from 1 to 10.

Green500 Rank	MFLOPS/W	Site	System	Total Power (kW)
1	9462.1	NVIDIA Corporation	NVIDIA DGX-1, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100	349.5
2	7453.5	Swiss National Supercomputing Centre (CSCS)	Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100	1312
3	6673.8	Advanced Center for Computing and Communication, RIKEN	ZettaScaler-1.6, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband FDR, PEZY-SCnp	150.0
4	6051.3	National Supercomputing Center in Wuxi	Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway	15371
5	5806.3	Fujitsu Technology Solutions GmbH	PRIMERGY CX1640 M1, Intel Xeon Phi 7210 64C 1.3GHz, Intel Omni-Path	77
6	4985.7	Joint Center for Advanced High Performance Computing	PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path	2718.7
7	4688.0	DOE/SC/Argonne National Laboratory	Cray XC40, Intel Xeon Phi 7230 64C 1.3GHz, Aries interconnect	1087
8	4112.1	Stanford Research Computing Center	Cray CS-Storm, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, Nvidia K80	190
9	4086.8	Academic Center for Computing and Media Studies (ACCMS), Kyoto University	Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect	748.1
10	3836.6	Thomas Jefferson National Accelerator Facility	KOI Cluster, Intel Xeon Phi 7230 64C 1.3GHz, Intel Omni-Path	111

Source: Top500/Green500

-- End --

EDITOR'S NOTE

At the NVIDIA GTC (GPU Technical Conference) in Amsterdam on 28-Sep-2016, Jen-Hsun Huang, CEO at NVIDIA presented the opening keynote during more than 2 Hours, I had the pleasure to talk to him and several members of his team.

This conference was one of the very best I had in my career of over 3 decades, if you are involved in those technologies I recommand you to attend one of the next editions.

Daniel Dierickx

IBM and NVIDIA Team Up on World's Fastest Deep Learning Enterprise Solution

New IBM PowerAI Software Toolkit Paired with NVIDIA NVLink and GPUDL Libraries Optimized for IBM Power Architecture Helps Enable 2X Performance Breakthroughs on AlexNet with Caffe

NOTE: *this article as an uptade to the IBM Press Conference in Amsterdam on 02-Sep-2016 and co-located with the NVIDIA GPU Conference (see page 14, 15 & 16). Participants include: Sumit Gupta VP High Performance Computing a Analitics at IBM Systems in San Jose CA, Adel El-Hallak Director OpenPower Marketing at IBM Systams in Somers NY, Yves Van Seters Leader External Relations IBM Benelux, Daniel Dierickx Editor e2mos (see page 2) and many international attendees.*

SALT LAKE CITY, UT - 14 Nov 2016: IBM (IBM: NYSE) and NVIDIA (NASDAQ: NVDA) today announced collaboration on a new deep learning tool optimized for the latest IBM and NVIDIA technologies to help train computers to think and learn in more human-like ways at a faster pace.

Deep learning is a fast growing machine learning method that extracts information by crunching through millions of pieces of data to detect and rank the most important aspects from the data. Publicly supported among leading consumer web and mobile application companies, deep learning is quickly being adopted by more traditional business enterprises.

Deep learning and other artificial intelligence capabilities are being used across a wide range of industry sectors; in banking to advance fraud detection through facial recognition; in automotive for self-driving automobiles and in retail for fully automated call centers with computers that can better understand speech and answer questions.

A new deep learning software toolkit available today called IBM PowerAI runs on the recently announced IBM server built for artificial intelligence that features NVIDIA® NVLink™ interconnect technology optimized for IBM's Power architecture. The hardware-software solution provides more than 2X performance over comparable servers with 4 GPUs running AlexNet with Caffe.[1] The same 4-GPU Power-based configuration running Alexnet with BVLC Caffe can also outperform 8 M40 GPU-based x86 configurations [2], making it the world's fastest commercially available enterprise systems platform on two versions of a key deep learning framework.

Caffe is a widely-used deep learning framework developed by Berkeley Vision and Learning Center (BVLC) and is recognized within the technology industry as one of the most popular deep learning community applications. Caffe is one of five deep learning software frameworks available in the IBM PowerAI toolkit. The toolkit leverages NVIDIA GPUDL libraries including cuDNN, cuBLAS and NCCL as part of NVIDIA SDKs to deliver multi-GPU acceleration on IBM servers.

IBM PowerAI is designed to run on IBM's highest performing server in its OpenPOWER LC lineup, the IBM Power S822LC for High Performance Computing (HPC), which features NVIDIA NVLink technology optimized for the Power architecture and NVIDIA's latest GPU technology. The new solution supports emerging computing methods of artificial intelligence, particularly deep learning. IBM PowerAI also provides a continued path for Watson, IBM's cognitive solutions platform, to extend its artificial intelligence expertise in the enterprise by using several deep learning methods to train Watson.

"PowerAI democratizes deep learning and other advanced analytic technologies by giving enterprise data scientists and research scientists alike an easy to deploy platform to rapidly advance their journey on AI," said Ken King, General Manager, OpenPOWER. "Coupled with our high performance computing servers built for AI, IBM provides what we believe is the best platform for enterprises building AI-based software, whether it's chatbots for customer engagement, or real-time analysis of social media data."

"Our innovation with IBM on NVIDIA NVLink has created new opportunities for POWER in the deep learning and analytics market," said Ian Buck, VP and GM of Accelerated Computing Group. "NVIDIA's GPUDL libraries in PowerAI will provide world class high-performance tools to power GPU-accelerated deep learning applications."

IBM PowerAI is available immediately at no charge to customers of IBM's Power S822LC for HPC server. PowerAI is designed to run on a single S822LC server and also to scale to large scale supercomputing clusters consisting of dozens, hundreds or thousands of servers.

... to next page

IBM and NVIDIA Team Up on World's Fastest Deep Learning Enterprise Solution

... from previous page

The NVLink Advantage

PowerAI is a set of binary distributions of popular deep learning frameworks including Caffe, Torch and Theano. Additional distributions include the IBM and NVIDIA versions of the Caffe deep learning frameworks, IBM-Caffe and NVCaffe. IBM has optimized each of the distributions to take advantage of the recently announced IBM POWER8 chip with the NVIDIA NVLink interface featured on the IBM Power S822LC for HPC server.

The POWER8 with NVIDIA NVLink chip is a technology-leading processor design that is the result of open collaboration between OpenPOWER Foundation members IBM and NVIDIA. The new chip enables tight integration between IBM's POWER8 CPU server architecture and the new Pascal architecture NVIDIA's Tesla P100 GPU accelerators. The CPUs and GPUs integrated into the Power S822LC for HPC are connected to each other via the high-speed NVIDIA NVLink interconnect. This industry-unique interface between the CPUs and GPUs, and also between the GPUs, removes potential bottlenecks created by the PCIe interface found in most Intel x86-based servers. The PowerAI toolkit of deep learning applications takes advantage of this new NVLink-based server architecture to optimize performance of the leading artificial intelligence, deep learning and machine learning applications.

Growing Momentum for the Power S822LC for High Performance Computing

The hardware pairing for PowerAI, the IBM Power S822LC for HPC server, was launched in early September. There was immediate interest in the server, equipped with raw performance advantages, among leading research institutions, cloud service providers and business enterprises. This led to very strong demand in the third quarter, contributing to Power's 2x year to year growth in Linux systems revenues.

Initial client uses for the new IBM Power S822LC for HPC servers include:

- **Human Brain Project** – In support of the Human Brain Project, a research project funded by the European Commission to advance understanding of the human brain, IBM, and NVIDIA deployed a pilot system at the Juelich Supercomputing Centre as part of the Pre-Commercial Procurement process. Called JURON, the new supercomputer leverages Power S822LC for HPC systems.
- **Cloud provider Nimbix** – HPC cloud platform provider, Nimbix expanded its cloud supercomputing offerings this month, putting IBM Power S822LC for HPC systems with PowerAI in the hands of developers and data scientists to achieve enhanced performance.
- **City of Yachay, Ecuador** – Ecuador's "City of Knowledge," Yachay, is a planned city designed to push the nation's economy away from commodities and towards knowledge-based innovation. Last week the city announced it is using a cluster of Power S822LC servers to build the country's first supercomputer for the purpose of creating new forms of energy, predicting climates, and pioneering food genomics.
- **SC3 Electronics** – A leading cloud supercomputing center in Turkey, SC3 Electronics announced last month at the OpenPOWER Summit Europe that it is creating the largest HPC cluster in the Middle East and North Africa region based on Power S822LC for HPC servers.

To download IBM PowerAI, go to www.ibm.biz/powerai.

To learn more about the IBM Power S822LC for High Performance Computing server, go to: www.ibm.biz/s822lc-hpc.

To learn more about NVIDIA deep learning, go to www.nvidia.com/deeplearning.

Xilinx to Dive in Hyperscale Race Going after Intel/Altera and Nvidia

Junko Yoshida

TOKYO — Xilinx is jumping into the increasingly hot race for hyperscale data centers.

The FPGA company is pursuing Altera (now a part of Intel), who took an early lead in the growing market by getting designed into Microsoft's Project Catapult, the technology behind Microsoft's hyperscale acceleration fabric. Xilinx also plans to go after the inference side of machine learning, an area in which the company is confident in beating Nvidia, as the GPU company is more focused on the training side of machine learning.

At the SC16 in Salt Lake City, Xilinx is rolling out a new suite of technology designed to enable the world's largest cloud service providers to rapidly develop and deploy acceleration platforms.

The move reflects Xilinx' commitment to capitalize on a recent big design win scored with **Chinese Web giant Baidu**. Baidu designed a software-defined accelerator based on an eight-lane PCIe 3.0 card using a 20nm Xilinx KU115 FPGA and 8-32 Gbytes memory.

Baidu's system parses SQL constructs into five types, enabling hardware processing elements in the FPGA to speed up from 8-55x depending on the application. This was discussed at the Hot Chips even last August.

Andy Walsh, director of strategic market development for cloud computing at Xilinx, told EE Times that in addition to the Baidu design win, "two other hyperscale companies are testing Xilinx' FPGA" for its integration into their accelerators.

Hyperscale computing is in high demand. Companies such as Facebook, Google, Microsoft, Amazon and Baidu are seeking a new data center architecture that can scale appropriately as systems face increased demand.

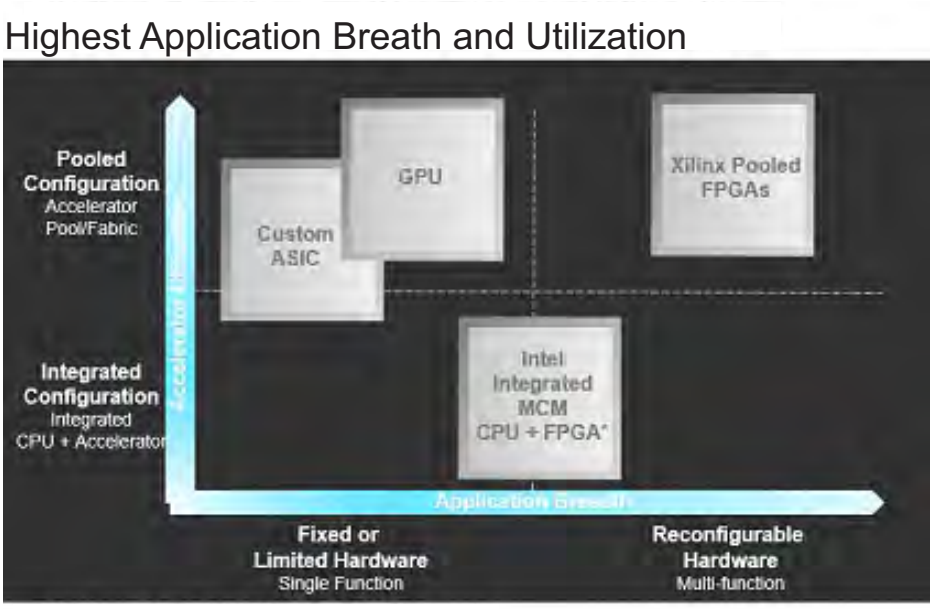
Walsh joined Xilinx last year after spending twelve years at Nvidia. He stressed that the very nature of cloud data centers has dramatically changed over the last several years when compared with the days when banks and insurance companies bought up data servers based on Intel CPUs.

Hyperscale companies are looking for flexible and scalable solutions for accelerators. They want "reconfigurable acceleration stack that can handle everything from machine learning, transcoding to data analytics, networking and storage," he explained.

In response, Xilinx is launching its FPGA-powered Xilinx Reconfigurable Acceleration Stack, which includes libraries, framework integrations, developer boards, and OpenStack support.

Noting that Xilinx enables silicon optimization for the broadest set of performance-demanding workloads, Walsh said, "These workload optimizations can be done in milliseconds by swapping in the most optimal design bitstream."

Compared with x86 server CPUs, Xilinx's FPGAs can provide "the fastest path to realize 40x better compute efficiency," the company claimed. Xilinx also boasted that the company can offer "up to six times the compute efficiency over competitive FPGAs."



* Limited in size, flexibility and choice due to MCM power density

The race for getting reconfigurable (Source: Xilinx)

Please Click Here to continue reading this excellent article from Junko Yoshida.

It includes 2 more pages:
- Machine learning
- Architecture advantages

Cavium to Demonstrate Leading Datacenter, HPC and Next-generation Cloud Infrastructure Solutions at Red Hat Summit 2017

Cavium, Inc. (NASDAQ: CAVM), a leading provider of semiconductor products that enable secure and intelligent processing for enterprise, datacenter, cloud, wired and wireless networking, will demonstrate leading datacenter, HPC and next-generation cloud infrastructure solutions with ThunderX® running on Red Hat operating systems and applications at Red Hat Summit 2017.

Red Hat Summit is the premier open source technology event to showcase the latest and greatest in cloud computing, platform, virtualization, middleware, storage, and systems management technologies.

The ThunderX product family is Cavium's 64-bit ARMv8-A server processor for datacenter and cloud applications, and features high-performance custom cores, single- and dual-socket configurations, high memory bandwidth and large memory capacity. The product family also includes integrated hardware accelerators, integrated feature-rich high bandwidth network and storage IO, fully virtualized core and IO, and scalable high bandwidth, low latency Ethernet fabric, which affords ThunderX best-in-class ARMv8-A performance per dollar. They are fully compliant with ARMv8-A architecture specifications as well as ARM's SBSA and SBBR standards, and widely supported by industry-leading OS, Hypervisor and Software tool and application vendors.

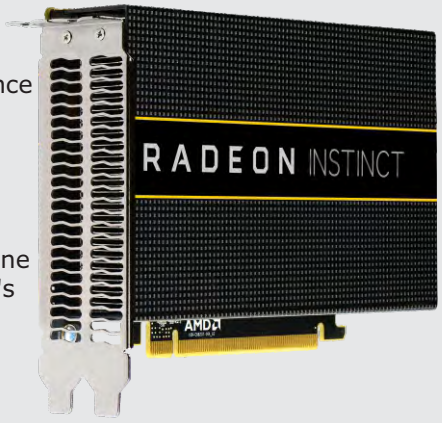
Cavium will present the following product demonstrations in the ARM Ecosystem Showcase booth #426:

• ThunderX & ThunderX2™: 64–bit ARMv8–based SoC family of workload–optimized processors with a range of SKUs and form factors optimized for scale out workloads including volume compute, storage, secure compute and networking running Red Hat Enterprise Linux OS and key cloud workloads.
• Cavium OEM and Cloud partners will be presenting ThunderX platform demonstrations and ThunderX Web Hosting Solutions.
• Cavium Open Source and Application partners will be demonstrating software applications and expanding ARM server software ecosystem availability and maturity.

Cavium will also be participating at the Red Hat hosted panel session on ARM-based datacenters. A number of Cavium partners will also be participating at this panel including Packet, Linaro and ARM. **MORE:** http://cavium.com/

AMD introduces Radeon Instinct: Accelerating Machine Intelligence

AMD speeds deep learning inference and training with high-performance Radeon Instinct accelerators and MIOpen open-source GPU-accelerated library



AMD (NASDAQ: AMD) today unveiled its strategy to accelerate the machine intelligence era in server computing through a new suite of hardware and open-source software offerings designed to dramatically increase performance, efficiency, and ease of implementation of deep learning workloads. New Radeon™ Instinct accelerators will offer organizations powerful GPU-based solutions for deep learning inference and training. Along with the new hardware offerings, AMD announced MIOpen, a free, open-source library for GPU accelerators intended to enable high-performance machine intelligence implementations, and new, optimized deep learning frameworks on AMD's ROCm software to build the foundation of the next evolution of machine intelligence workloads.

Inexpensive high-capacity storage, an abundance of sensor driven data, and the exponential growth of user-generated content are driving exabytes of data globally. Recent advances in machine intelligence algorithms mapped to high-performance GPUs are enabling orders of magnitude acceleration of the processing and understanding of that data, producing insights in near real time. Radeon Instinct is a blueprint for an open software ecosystem for machine intelligence, helping to speed inference insights and algorithm training.

"Radeon Instinct is set to dramatically advance the pace of machine intelligence through an approach built on high-performance GPU accelerators, and free, open-source software in MIOpen and ROCm," said AMD President and CEO, Dr. Lisa Su. "With the combination of our high-performance compute and graphics capabilities and the strength of our multi-generational roadmap, we are the only company with the GPU and x86 silicon expertise to address the broad needs of the datacenter and help advance the proliferation of machine intelligence."

At the AMD Technology Summit held last week, customers and partners from 1026 Labs, Inventec, SuperMicro, University of Toronto's CHIME radio telescope project and Xilinx praised the launch of Radeon Instinct, discussed how they're making use of AMD's machine intelligence and deep learning technologies today, and how they can benefit from Radeon Instinct. For more information, visit Radeon.com/Instinct.