Modern Machine Learning Technology*

Prof. Dr. Victor David Sánchez, Ph.D.
Brilliant Brains. Palo Alto, California

December, 2017

Abstract

Artificial Intelligence (AI) and Machine Learning (ML) are showing tremendous potential in industrial applications ranging from internet, cloud, media, entertainment, medicine, biology, security, defense, to autonomous vehicles, driving and flying. The status quo is reviewed providing concrete insight of this fascinating and continuously evolving technology which is currently being massively deployed in multiple industrial segments and which continues to forcefully advance to its next stages of development.

The author has been committed to advancing AI and ML since the early 1980's including pioneering breakthrough applications, industrial and in research [10], advanced learning / training methods [11] as well as the early use of neurochips and advanced development environments [6]. After being heavily involved in AI for VLSI, i.e., built an AI- tool at the Karlsruhe Institute of Technology (KIT) to design ASICs which we then used at Siemens AG to build custom processors for automation, I founded with around 15 experts a committee on VLSI for AI in 1985 in Berlin.

Conscious of the demanding time and memory complexity requirements, with some of them I built later a real-time, scalable, heterogeneous, parallel distributed supercomputer at the German NASA (DLR), the fastest of its time worldwide, on which intelligent, autonomous systems could be implemented for space and terrestrial applications, e.g., at Daimler Benz (Mercedes) for its autonomous vehicle development and which we used in a NASA-ESA-DLR spaceshuttle-spacelab mission demonstrating fully integrated, real-time AI, computer vision, augmented reality, analytics, teleoperation, shared control, and full autonomy for specific tasks, i.e., in brief what only decades later, today, the rest of the world is pursuing.

REFERENCES

[1] A. Krizhevsky et al. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS25*, 2012.

[2] C. Szegedy et al. Rethinking the Inception Architecture for Computer Vision. In *CVPR29*, 2016.

[3] J. Deng et al. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[4] K. He et al. Deep Residual Learning for Image Recognition. In *CVPR29*, 2016.

[5] K. Simonyan et al. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR03*, 2015.

[6] V.D. Sánchez et al. Maschinenmarkt–On the Way to Intelligence, Structure and Function of Artificial Neural Networks using Supervised Learning; The Grey Cells as Example, Analyzed Neural Operations can be realized by VLSI Components (in German)–, 96[46], 97[3]; Chip Plus–ANSpec, A Specification Language (in German)–, 7; Technische Rundschau–Neurocomputers in Industrial Applications (in German)–, 82[65]; Neurocomputing–The Design of a Real-Time Neurocomputer Based on RBF Networks–, 20, 1990, 1991, 1990, 1990, 1998.

[7] Google. Cloud TPU. `https://cloud.google.com/tpu/`.

[8] Intel. AI Academy. `https://software.intel.com/en-us/ai-academy`.

[9] Nvidia. Deep Learning and AI. `https://www.nvidia.com/en-us/deep-learning-ai/`.

*This abstract has been granted permission for public release.

[10] V.D. Sánchez. Neurocomputing–Research and Applications–; Increasing the Autonomy of Space Robots; Intelligent BioSystems; Modeling Dynamics for Communications, Navigation, Guidance and Control Applications. German Research Center for Anthropotechnics, Wachtenberg-Werthoven, Germany; NASA Ames Research Center, Moffett Field, CA; KAIST Department of BioSystems, Daejeon, South Korea; Rockwell Collins, Advanced Technology Center (ATC), Cedar Rapids, IA, 1990, 2002, 2003, 2011.

[11] V.D. Sánchez. Neurocomputing–Special Issue on Backpropagation, parts I-IV–, 5[4–6], 6[1–2]; –Special issue on RBF Networks, parts I-II–, 19[1–3], 20[1–3]; Advanced Support Vector Machines and Kernel Methods, 55[1–2], 1993–1994,1998,2003.

The main industrial players include currently multiple large corporations like Google, Microsoft, IBM, Amazon, Intel, specialized startups, either acquired or still growing, like Movidius, Argo AI, Scyfer, Zebra Medical Vision to provide either the basic infrastructure to develop deep learning applications or even highly specific vertical applications as well as multiple end users, investors like J.P. Morgan in financial services, Illumina in genomics, Tesla in the automotive industry, Lockheed Martin in cybersecurity to mention just very few. Here we showcase a few key representatives to provide direct insight into the developing systems being used to perform inference and training of machine learning models in practice.

Since the inference, i.e., running an already trained model, is computationally less challenging, our focus is on the training (in case of supervised learning) of deep learning models. Using benchmarks like Imagenet [3], Deep Neural Networks (DNN), e.g., AlexNet [1], RedNet-152 [4], InceptionV3 [2], VGG-16 [5] to mention a few, are often tested and compared, cf. some DNN example architectures in Figure 7, shown in (a) is the AlexNet and in (b) VGG-19 and ResNet-34 in comparison.

Figure 1 shows in (a) an Intel's Xeon Phi processor often built in servers which offer DNN training and inference in the cloud, in (b) planned evolution of the Xeon Phi product family including Knight Hills in 10 nm technology, in (c) members of Intel's family of processors for AI and Deep Learning (DL) datacenter workloads used for inference and training, and in (d) the Knight Hills System-on-Chip (SoC). Microsoft's new cloud acceleration framework, called hardware microservices model, designed also to accelerate DNNs, utilizes hybrid servers that mix Intel's Xeon processors (CPUs) with Intel-Altera's Field Programmable Gate Arrays (FPGAs). Fairly recently, Intel has announced work towards the next exascale microarchitecture which should bring among others Cray to build the U.S. DoE's Aurora Supercomputer to provide EFlop performance.

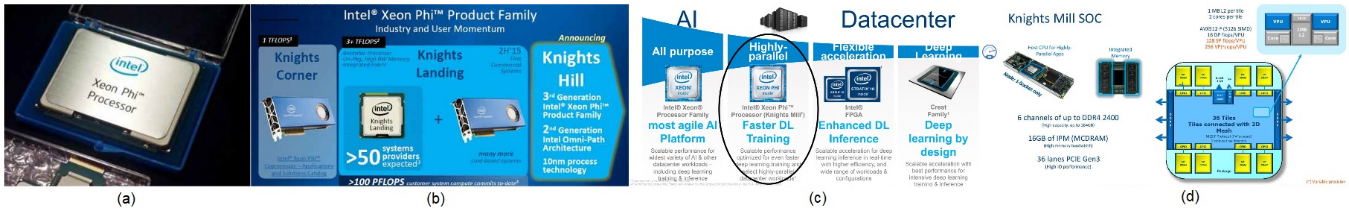

Figure 1: Intel's Xeon Phi Processor Product Family for Datacenter AI and ML Acceleration [8]

Figure 2 shows in (a) a so called Google's TPU (Tensor Processing Unit) pod composed of 64 2nd generation TPUs providing 11.5 PFlops and 4TB of memory interconnected using a 2D toroidal mesh network, in (b) a board hosting 4 2nd generation TPUs, each chip providing 45 TFlops, in (c) a 1st generation TPU used only for inference, and in (d) 1st generation TPUs deployed at a Google datacenter. Using TensorFlow, an open-source symbolic math library used for machine learning in research and production at Google, stateful dataflow graphs are used to express computations, e.g., operations on tensors performed by neural networks, that can then run in accelerated fashion on multiple CPUs, GPUs, and Google' s TPUs, i.e., ASICs used as high throughput programmable AI accelerators for inference and in its 2nd generation also for training.

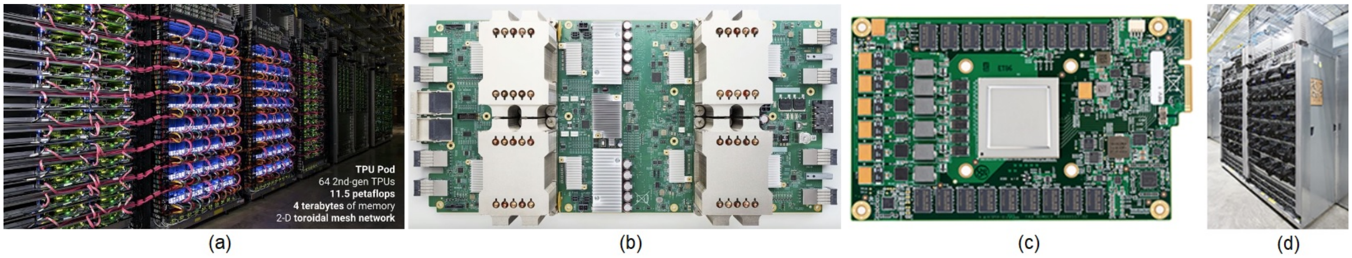

Figure 2: Google's AI Accelerating Supercomputers based on 1st and 2nd generation TPUs [7]

Figure 3 shows in (a) a 4 PFlop supercomputer Fujitsu is building for the Japanese Riken research center composed of 32 Fujitsu's Intel Xeon-based servers and 24 Nvidia's DGX-1 AI accelerator systems, in (b) the Nvidia's 170 TFlop FP16 first generation DGX-1 deep learning supercomputer based on 8 Tesla P100, and in (c) the Tesla P100 GPU accelerator based on the

Pascal GPU architecture hosting 3584 CUDA cores, 21.2 TFlops FP16, 16 GB HMB2 GPU memory, 732 GB/s memory bandwidth, and the NVLink high-speed interconnect to tightly integrate CPUs and GPUs. APIs supported include Cuda, OpenCL among others.

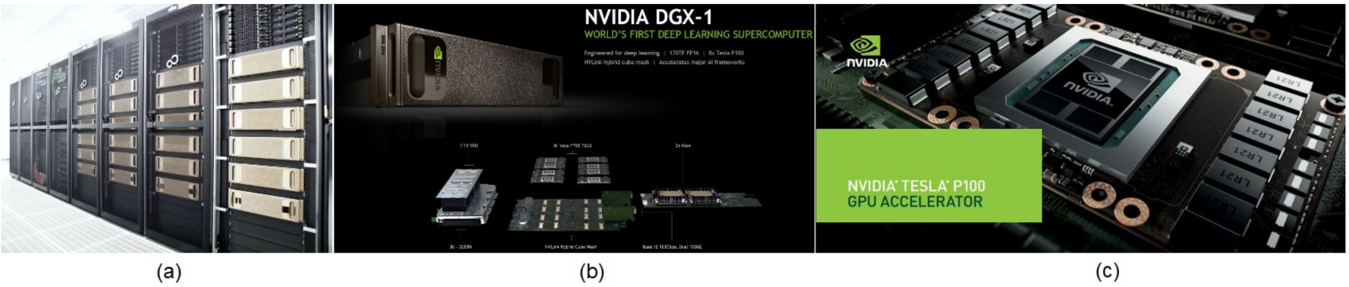

Figure 3: Deep Learning Supercomputers based on Nvidia's DGX-1 and Tesla P100 GPU Accelerator [9]

Figure 4 shows in (a) the Nvidia's deep learning Software Development Kit (SDK), basis for diverse frameworks and applications, in (b) the Nvidia's interactive deep learning GPU training system DIGITS to process data, configure DNNs, monitor progress and visualize layers, and in (c) the DGX Stack, a fully integrated deep learning platform composed of the GPU-optimized Linux server OS, GPU compute driver software, the containerization tool NVDocker, the deep learning libraries cuDNN and NCCL, and on top DIGITS. Figure 5 shows in (a) the Nvidia's Titan V card based on the Nvidia's supercomputing GPU architecture named Volta, 110 TFlops, 12 GB HBM2 3D stacked memory, 652.8 GB/s memory bandwidth, 640 and 5120 Tensor and CUDA cores respectively (tensor cores are programmable matrix-multiply-and-accumulate units that provide a substantial boost to convolutions and matrix operations), in (b) a Nvidia's Tesla V100 SXM2 mezzanine card, 120 TFlops, 16 GB HBM2 GPU memory, 900 GB/s memory bandwidth, 640 and 5120 Tensor and CUDA cores respectively, 300 GB/s interconnect bandwidth, in (c) Nvidia's TensorRT, a high performance neural network inference optimizer and runtime engine for production deployment, and in (d) faster inference throughput results on ResNet-50 when using a Nvidia's V100 and TensorRT: 5707 images/sec and 6.97 ms real-time latency, vehemently beating all other three alternatives: a Xeon Broadwell-E CPU-only, a V100 and TensorFlow, and a P100 and TensorRT.

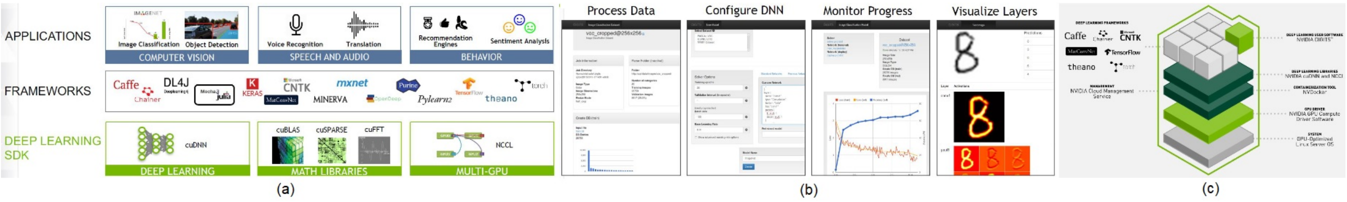

Figure 4: Nvidia's Deep Learning Platform: SDK, Digits, and DGX Stack [9]

The largest AI supercomputing system in Japan will soon be built also by Fujitsu for the Japanese National Institute of Advanced Industrial Science and Technology (AIST) for its AI Bridging Cloud Infrastructure (ABCI) capable of 550 PFlops consisting of 1,088 Fujitsu Primergy CX2570 M4 servers, each server hosting 2 Intel Xeon Gold processors and 4 Nvidia Tesla V100 GPU accelerators. Figure 6 shows in (a) Nvidia Docker to create containerized applications, i.e., to build and run Docker containers leveraging the deployed Nvidia's GPUs, in (b) how to deploy TensorRT as a microservice using the GPU REST Engine (GRE) SDK, and in (c) an example of this procedure with a classification microservice deployed to CPUs/GPUs.

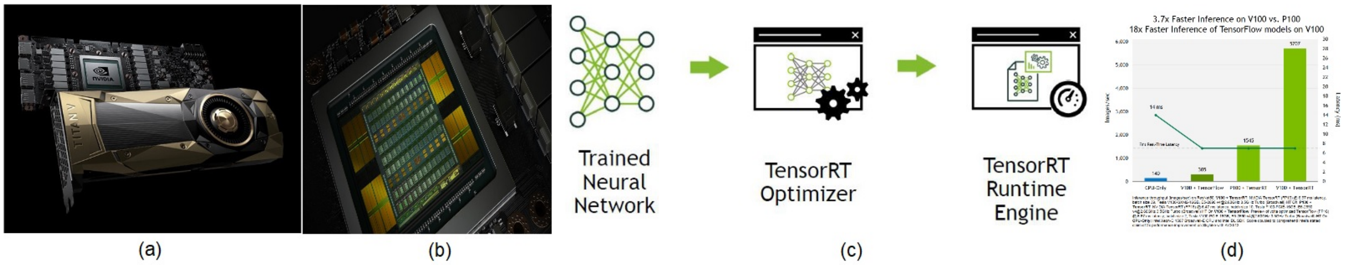

Figure 5: Volta Architecture, Tesla V100 GPU Accelerator, Acceleration Results using TensorRT [9]

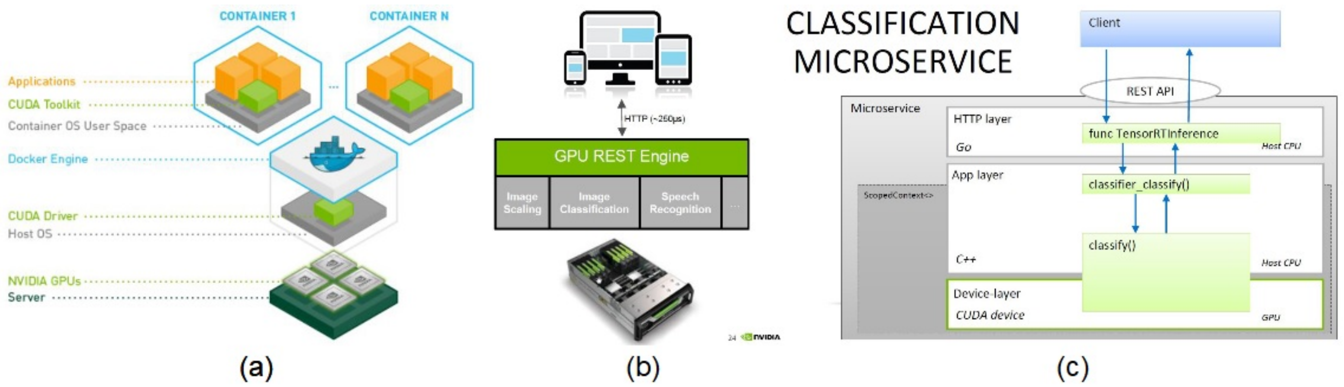

Figure 6: Application Containerization and Production Deployment of TensorRT Microservices [9]

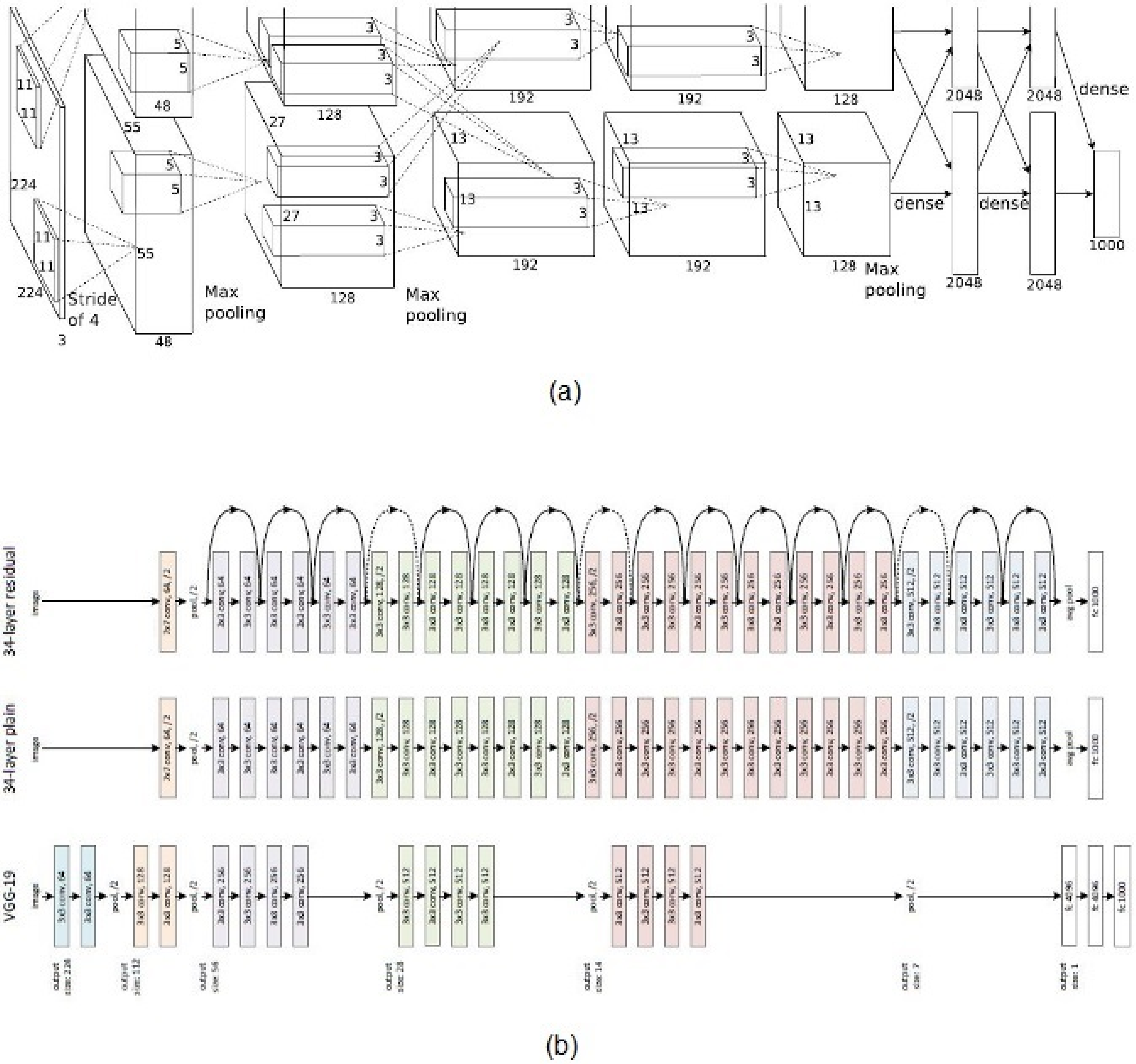

Figure 7: Some Deep Neural Network (DNN) Architecture Examples [1, 4]