Tools and Platforms for Data Analytics, Deep Learning, and Visualisation

Werner Scholz, 28. Nov. 2017
XENON Systems, CTO and Head of R&D
werners@xenon.com.au

XENON
High Performance Computing
www.xenon.com.au

XENON SYSTEMS – WHO WE ARE

Australian company established in 1996.

Direct Relationship with all major component manufacturers to lower cost and speed up support

- Scientific / Academic Research
- Oil and Gas
- Cloud

- Defence
- Education
- Broadcast

- Finance
- Telecommunication

In-house technical ability to build low volume custom designed servers

Focused on innovation and investment in Research & Development

Global Onsite hardware support and installation services in over 80 countries

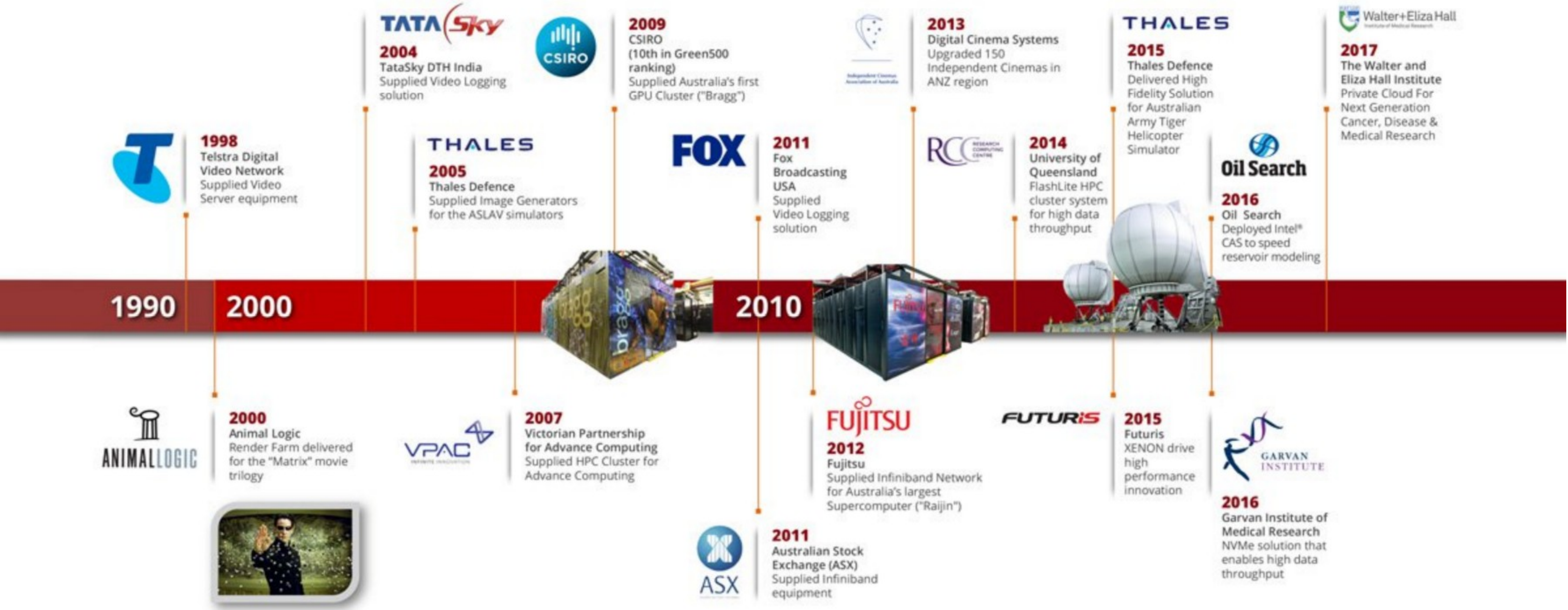
Subsidiaries:

Mediaproxy
Global leader in compliance logging and transport stream monitoring for broadcast and TV industries.

XDT/Catapult
Software for film and post production industries.

XENOptics
Fibre automation solutions for SDN in data centres

XENON SYSTEMS – A HISTORY OF HPC AND GPU SOLUTIONS



1998
Telstra Digital Video Network
Supplied Video Server equipment

2000
Animal Logic
Render Farm delivered for the "Matrix" movie trilogy

2004
TataSky DTH India
Supplied Video Logging solution

2005
Thales Defence
Supplied Image Generators for the ASLAV simulators

2007
Victorian Partnership for Advance Computing
Supplied HPC Cluster for Advance Computing

2009
CSIRO (10th in Green500 ranking)
Supplied Australia's first GPU Cluster ("Bragg")

2011
Fox Broadcasting USA
Supplied Video Logging solution

2011
Australian Stock Exchange (ASX)
Supplied Infiniband equipment

2012
Fujitsu
Supplied Infiniband Network for Australia's largest Supercomputer ("Raijin")

2013
Digital Cinema Systems
Upgraded 150 Independent Cinemas in ANZ region

2014
University of Queensland
FlashLite HPC cluster system for high data throughput

2015
Thales Defence
Delivered High Fidelity Solution for Australian Army Tiger Helicopter Simulator

2015
Futuris
XENON drive high performance innovation

2016
Oil Search
Deployed Intel® CAS to speed reservoir modeling

2016
Garvan Institute of Medical Research
NVMe solution that enables high data throughput

2017
The Walter and Eliza Hall Institute
Private Cloud For Next Generation Cancer, Disease & Medical Research

ACCELERATORS IN WEATHER FORECASTING

GPU accelerated WRF code (UCAR)

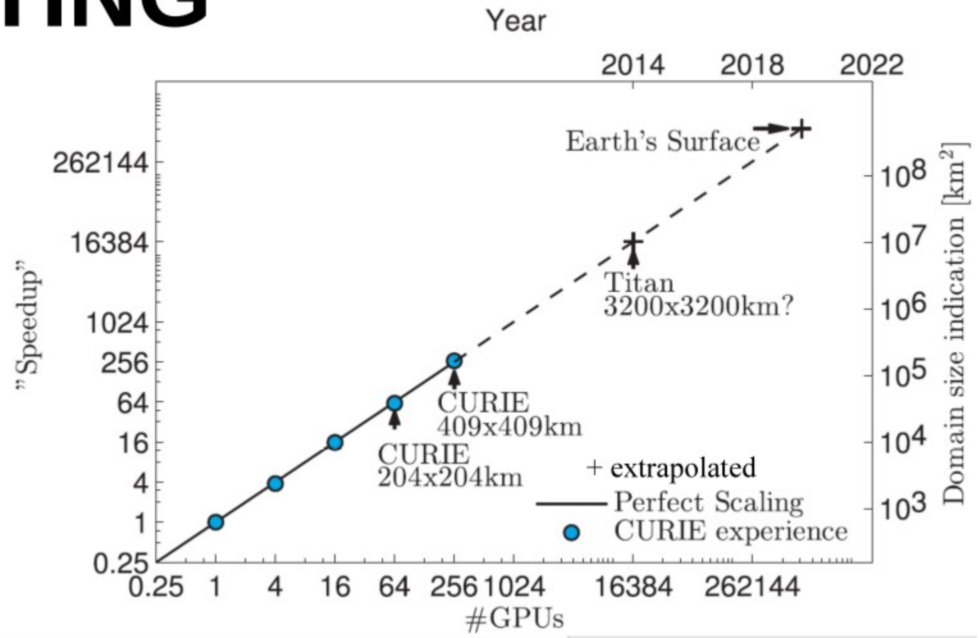
http://www.nvidia.com/object/weather.html

COSMO Weather Model

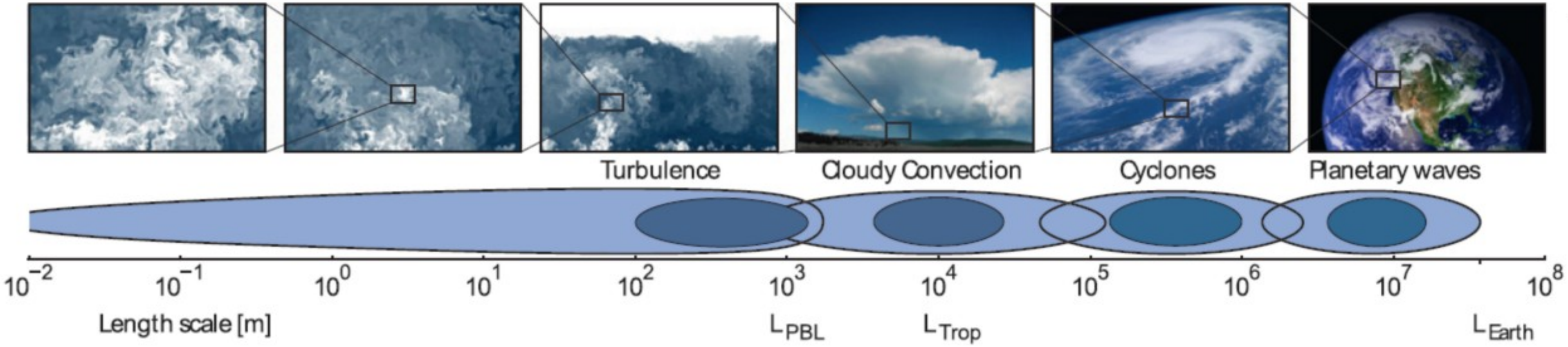
GPU accelerated version
http://www.cosmo-model.org/

Weather Forecasting Using GPU-Based Large-Eddy Simulations

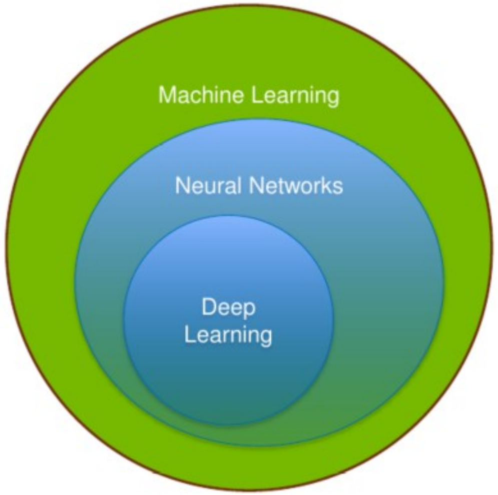
https://doi.org/10.1175/BAMS-D-14-00114.1 [1]



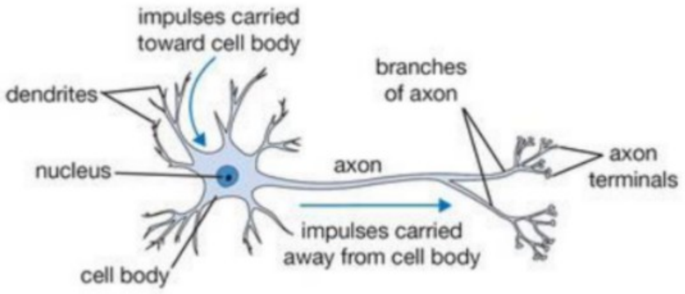
[1]



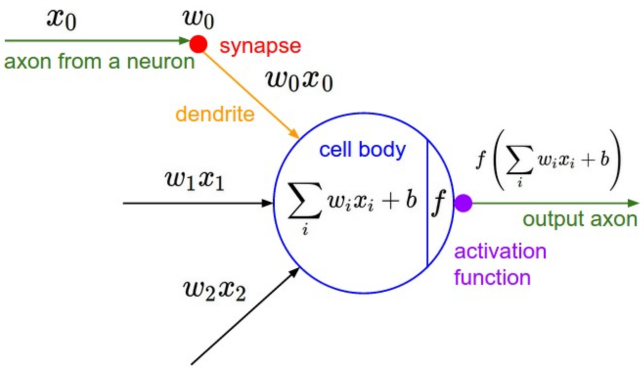
DEEP LEARNING AND AI

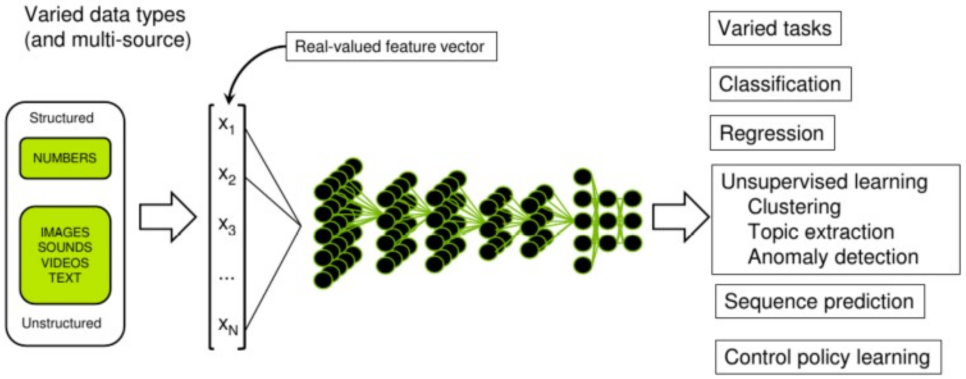


Biological neuron

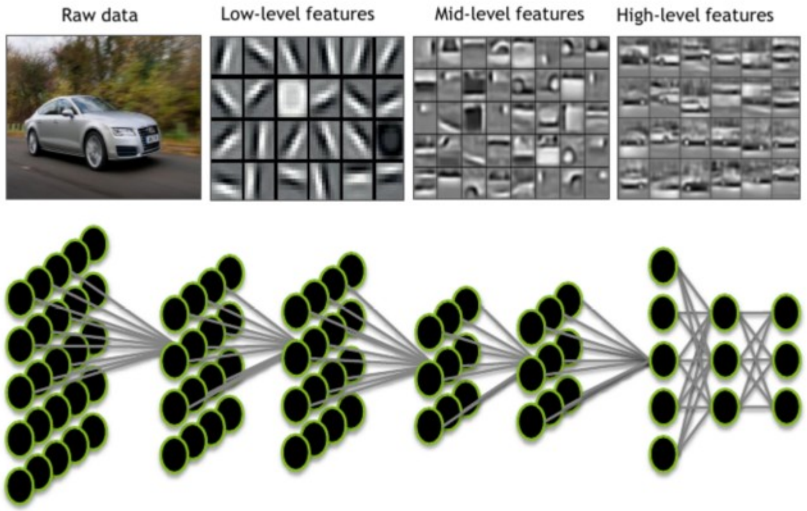

impulses carried toward cell body
dendrites
branches of axon
nucleus
axon
axon terminals
impulses carried away from cell body
cell body

From Stanford cs231n lecture notes



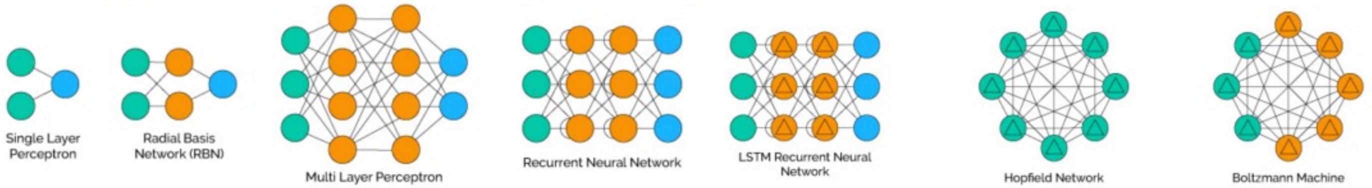


Constants: Big (high dimensional) Data + a complex function to learn



ENABLERS FOR DEEP LEARNING

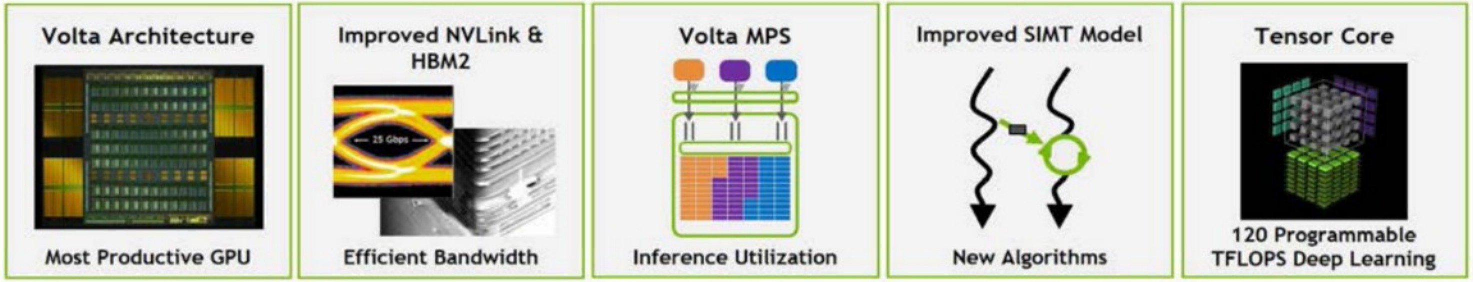
New approaches to the design and optimisation of neural networks



- Supervised and unsupervised Learning
- Adversarial Neural Networks

Development and publication of a variety of open source Deep Learning frameworks

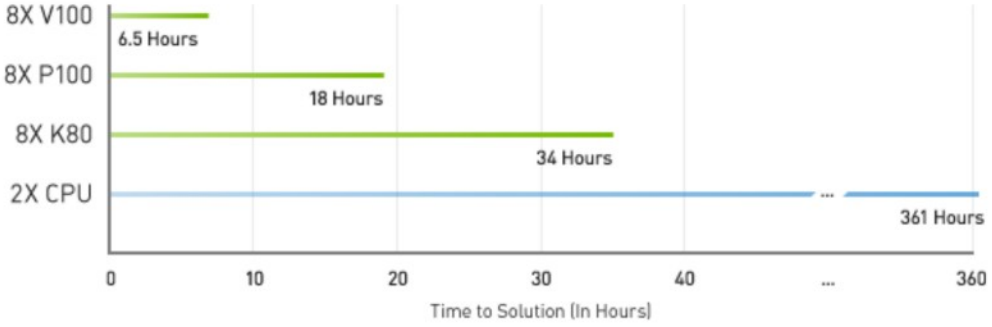
Capabilities of modern accelerator designs from NVIDIA, Intel, etc.



NVIDIA TESLA V100 (VOLTA ARCHITECTURE)

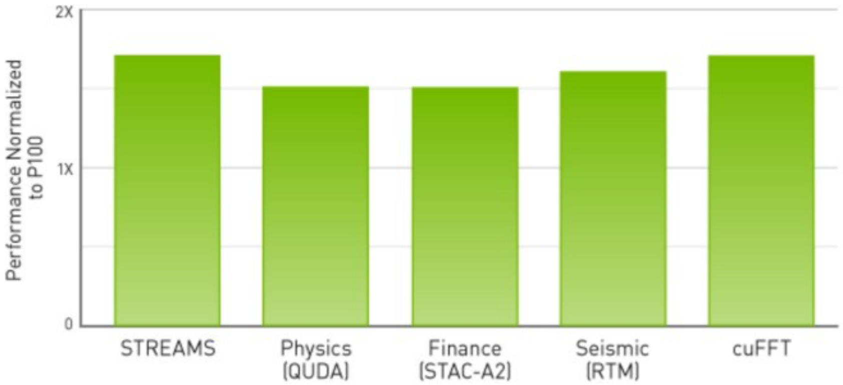
- TSMC 12nm FINFET process
- 21 Billion transistors
- >5000 compute units
- 15 TFLOPS DP
- 640 Tensor Cores
- 120 TFlops tensor operations
- 20MB register file
- 16MB cache
- 900 GB/s memory bandwidth
- 300 GB/s NVLINK2

3X Faster on Deep Learning Training



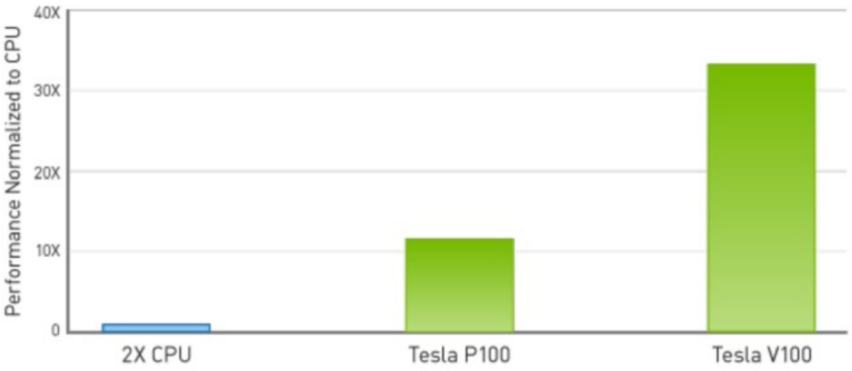
CPU Server: Dual Xeon E5-2699 v4, 2.6GHz | GPU Servers add 8X Tesla K80, Tesla P100 or Tesla V100 | V100 measured on pre-production hardware | Workload: NMT, 13 epochs to solution.

1.5X HPC Performance in One Year



CPU System: 2X Xeon E5-2660 v4 @ 2GHz | GPU System: NVIDIA® Tesla® P100 or V100 at 150W | V100 measured on pre-production hardware | Workload: ResNet-50

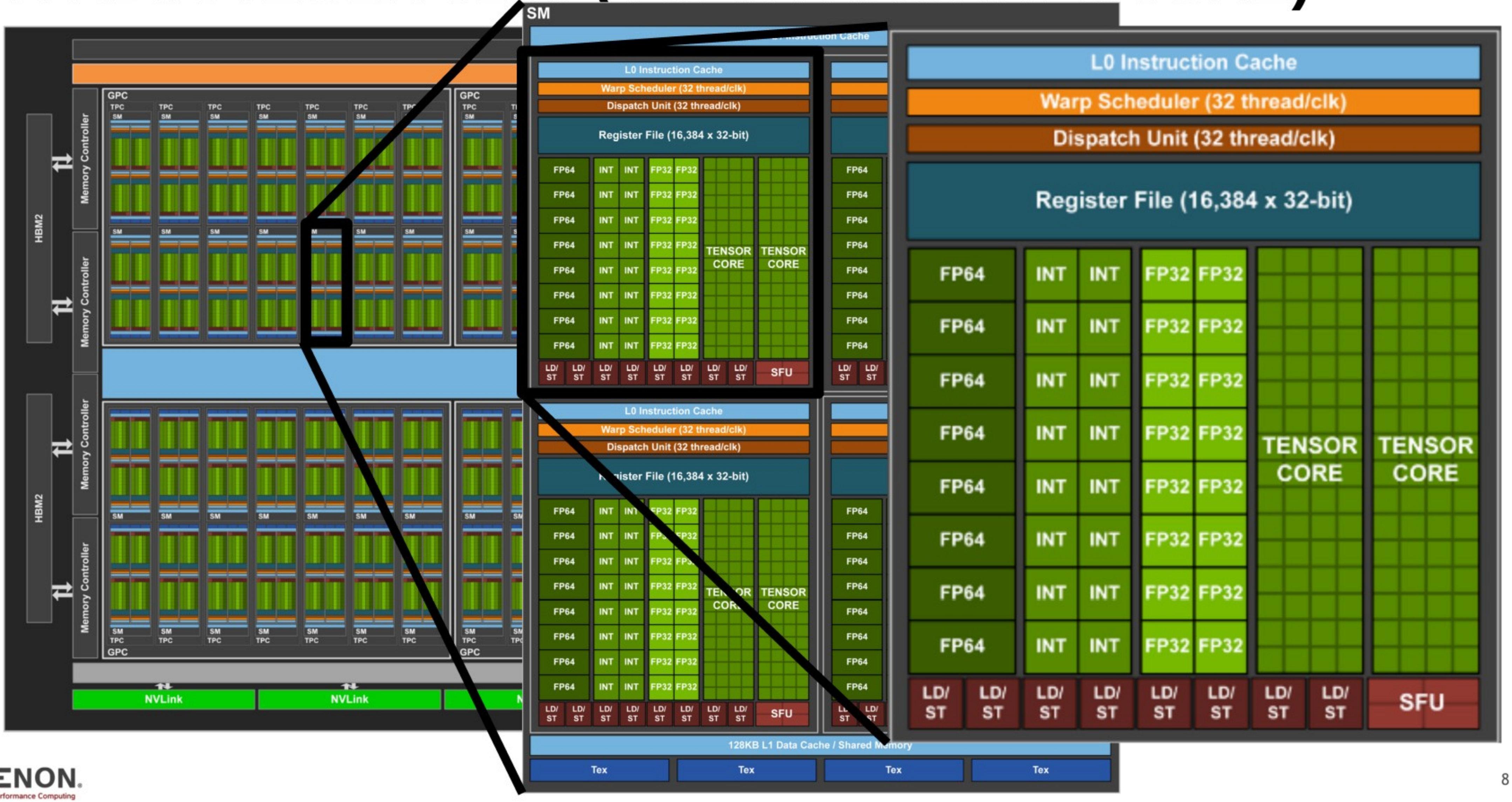
30X Higher Throughput than CPU Server on Deep Learning Inference



Workload: ResNet-50 | CPU: 2X Xeon E5-2660 v4, 2GHz | GPU: add 1X Tesla P100 or V100 at 150W | V100 measured on pre-production hardware.

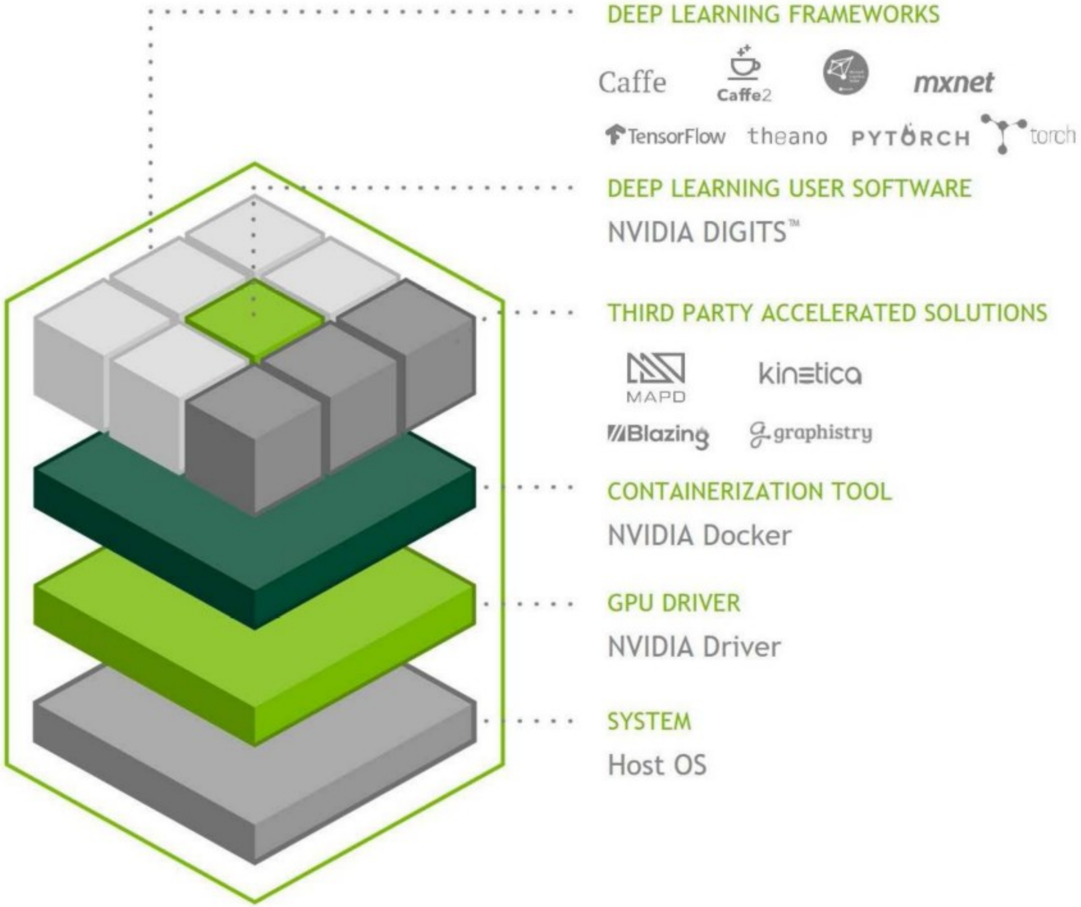
Ref.: https://www.nvidia.com/en-us/data-center/tesla-v100/

NVIDIA TESLA V100 (VOLTA ARCHITECTURE)


HBM2
Memory Controller
GPC
TPC
SM
NVLink
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
FP64
INT
FP32
TENSOR CORE
LD/ST
SFU
128KB L1 Data Cache / Shared Memory
Tex

Deep Learning Software Stack

High Performance GPU-Acceleration for Deep Learning



DEEP LEARNING FRAMEWORKS

Caffe Caffe2 mxnet

TensorFlow theano PYTORCH torch

DEEP LEARNING USER SOFTWARE

NVIDIA DIGITS™

THIRD PARTY ACCELERATED SOLUTIONS

MAPD kinetica

Blazing graphistry

CONTAINERIZATION TOOL

NVIDIA Docker

GPU DRIVER

NVIDIA Driver

SYSTEM

Host OS

- **Same software stack from workstation to Supercomputer**
- **Now available on all XENON GPU solutions**

DEEP LEARNING PLATFORMS - OVERVIEW

Workloads

- Dev and Test
- Training
- Inference

Technologies

- CPU
- GPU
- GPUs for DL (Tensor Cores), single prec., half prec.
- FPGA
- ASICS: TPU, etc.

On-premise

- GPU servers: x86, ARM,
  IBM 922SL: Power9 + V100 + NVLINK2
- NVIDIA DIGITS, IBM PowerAI

Cloud

- CPU, GPU, FPGA instances
- HWaaS: Softlayer
- DLaaS: Watson, “Tensorflow”aaS

New Services

- IBM Aivision, DLInsight

WHAT DOES IT RUN ON?

- XENON workstations with NVIDIA GPUs
- XENON DEVCUBE
- XENON Radon rack servers
  - 1U high density servers (up to 4 GPUs)
  - 4U 8-GPU servers
  - 4U 10-GPU servers
- NVIDIA DGX-1 and DGX Station
- IBM Power9 + V100 + NVLINK
- AMD platforms
- ARM (coming soon)

XENON SOLUTIONS

XENON Server Solutions

Performance and Reliability for the most demanding graphics, engineering, digital arts workloads.

GPU Computing

High performance **acceleration solutions** leveraging NVIDIA Tesla technology and the CUDA ecosystem

Virtualisation

End-to-end virtualisation solutions for compute, storage, networking, and desktop.

Storage

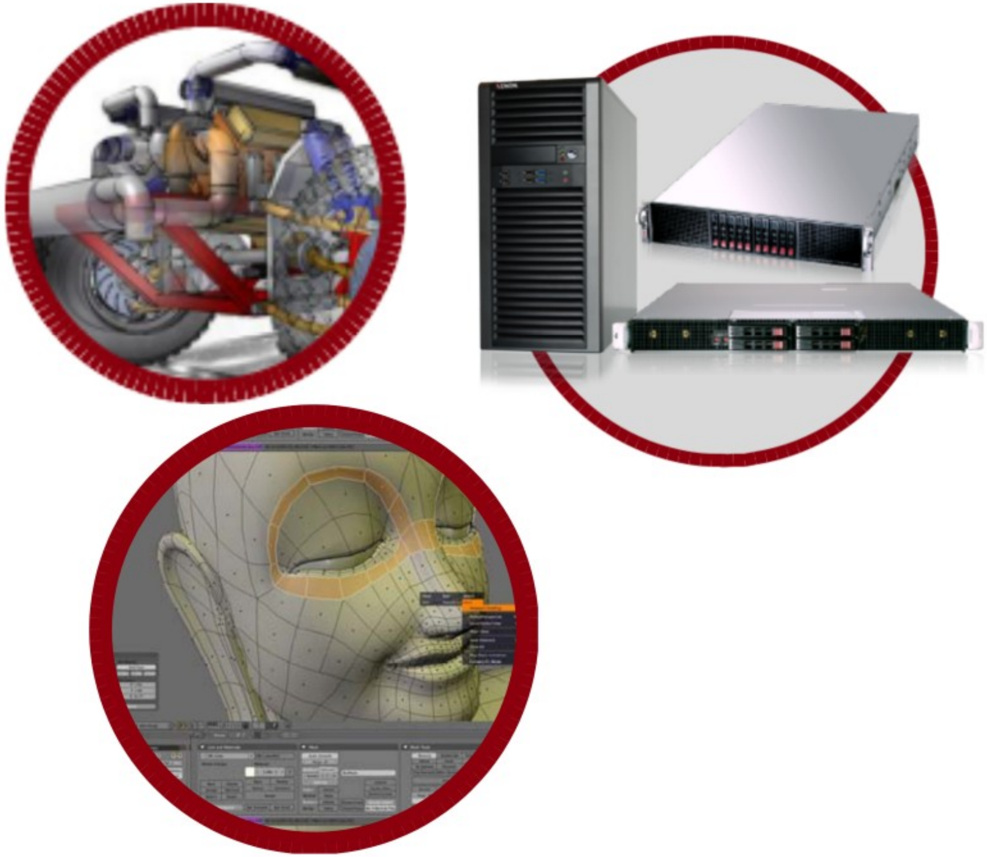
High performance parallel file systems,
e.g. IBM Spectrum Scale

Networking

High performance Infiniband and Ethernet solutions

Consulting

Implementation of DL/AI solutions



Thank you!

Werner Scholz, 28. Nov. 2017
XENON Systems, CTO and Head of R&D
werners@xenon.com.au

XENON
High Performance Computing
www.xenon.com.au