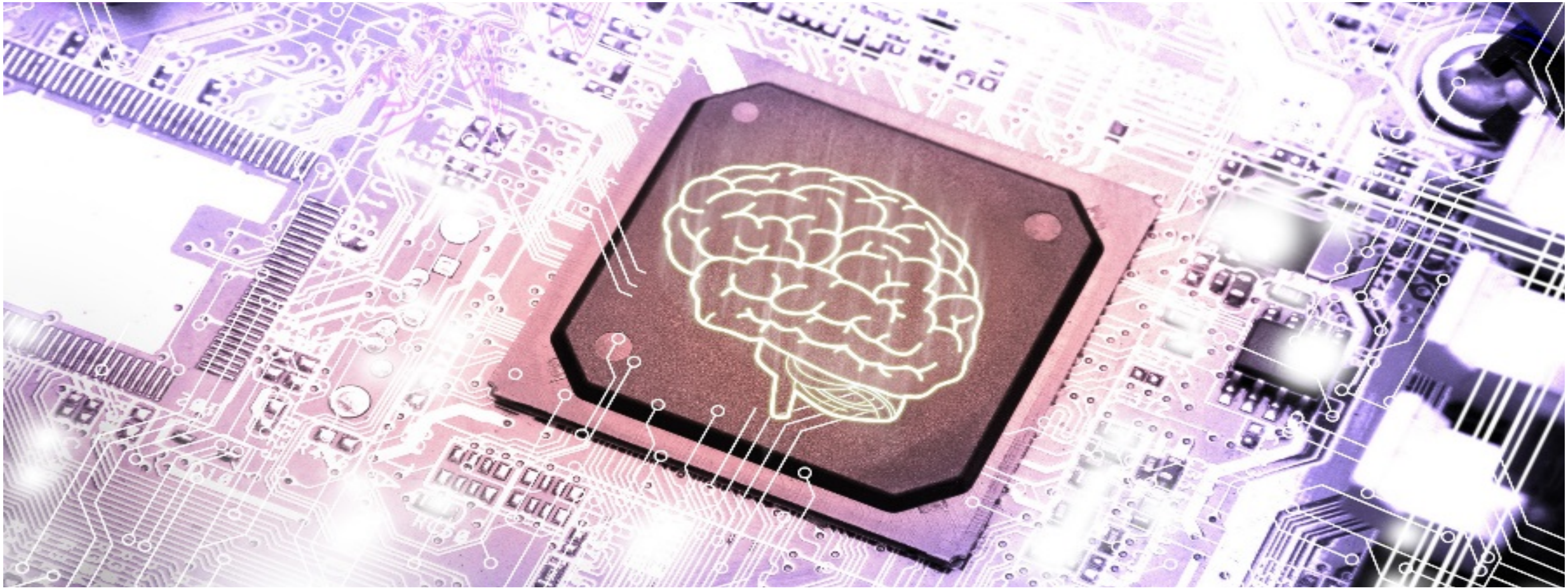


Few hints for the post exascale architectures

Marc Duranton and Denis Dutoit
Commissariat à l'énergie atomique et aux énergies alternatives

May 23th, 2019
Aristote seminar: "En route vers l'exascale!"

This presentation is based on the work done in ETP4HPC and in HiPEAC

A blueprint for the new Strategic Research Agenda for High Performance Computing

April 2019

https://www.etp4hpc.eu/pujades/files/Blueprint%20document_20190429.pdf

Common document between **ETP4HPC, BDVA & HiPEAC**

Contributors

Gabriel Antoniu, INRIA (**BDVA**)

Marc Asch, U-PICARDIE (**BDEC-2**)

Peter Bauer, ECMWF

Costas Bekas, IBM

Pascale Bernier-Bruna, **ETP4HPC**

Francois Bodin, IRISA

Laurent Cargemel, Atos

Paul Carpenter, BSC

Marc Duranton, CEA (**HiPEAC**)

Maike Gilliot, ETP4HPC

Hans-Christian Hoppe, INTEL

Jens Krueger, ITWM-FRAUNHOFER

Julian Kunkel, Univ. of Reading

Erwin Laure, KTH

Jean-Francois Lavignon, TECHNOLOGY-STRATEGY

Guy Lonsdale, SCAPOS

Michael Malms, **ETP4HPC**

Fabio Martinelli, CNR (**ECSO**)

Sai Narasimhamurthy, SEGATE

Marcin Ostasz, BSC

Maria Perez, UPM (**BDVA**)

Dirk Pleiter, JSC

Andrea Reale, IBM (**BDVA**)

Pascale Rosse-Laurent, Atos

This presentation is based on the work done in ETP4HPC and in HiPEAC

A blueprint for the new
Strategic Research Agenda
for High Performance Computing

ETP4HPC

exdci

April 2019

https://www.etp4hpc.eu/pujades/files/Blueprint%20document_20190429.pdf

HiPEAC

HiPEAC Vision 2019

HIGH PERFORMANCE AND EMBEDDED ARCHITECTURE AND COMPILATION

Editorial board:
Marc Duranton, Koen De Bosschere, Bart Coppens,
Christian Gamrat, Madeleine Gray, Harm Munk, Emre Ozer,
Tullio Vardanega, Olivier Zendra

https://www.hipeac.net/roadmap

Outline

1) Evolution of application scope: the continuum
2) Hardware heterogeneity and orchestration
3) Software?

Outline

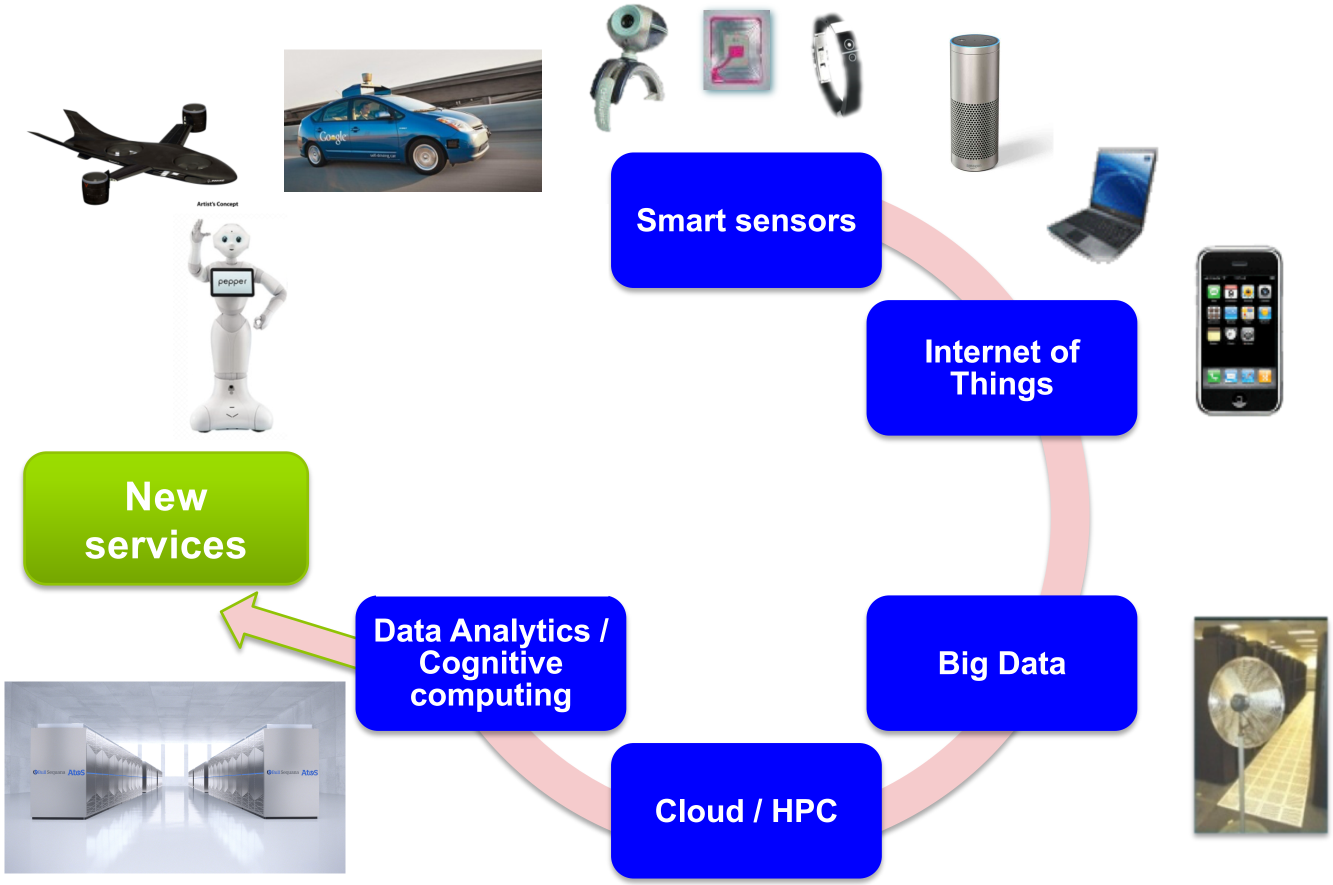
1) Evolution of application scope: the continuum
2) Hardware heterogeneity and orchestration
3) Software?

Outline

1) Evolution of application scope: the continuum
    1) From smart sensors to HPC
    2) Artificial Intelligence (Deep Learning) loads
    3) Implications for the architecture
2) Hardware heterogeneity and orchestration
3) Software?

FROM SMART SENSORS TO HPC

Mainstream “business” model



Smart sensors

Internet of Things

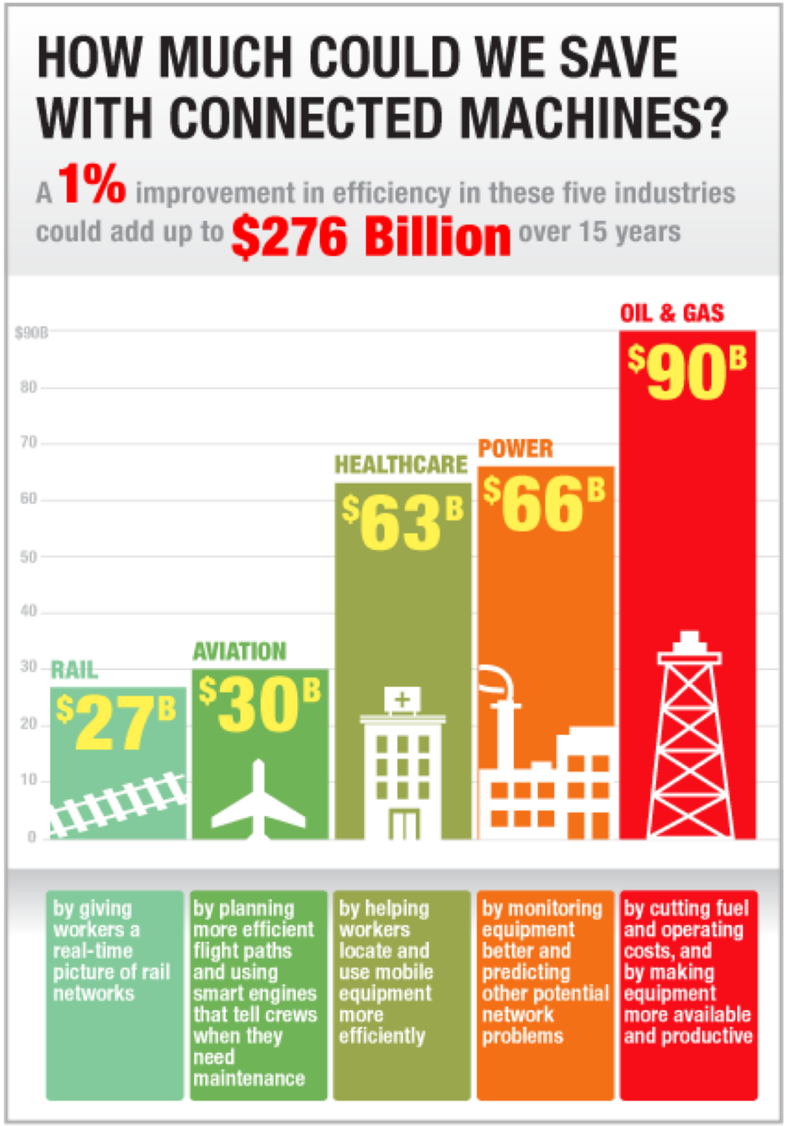
Big Data

Cloud / HPC

Data Analytics / Cognitive computing

New services

ECONOMICAL DRIVE OF CONNECTED THINGS: BETTER EFFICIENCY IN RESOURCES AND ENERGY

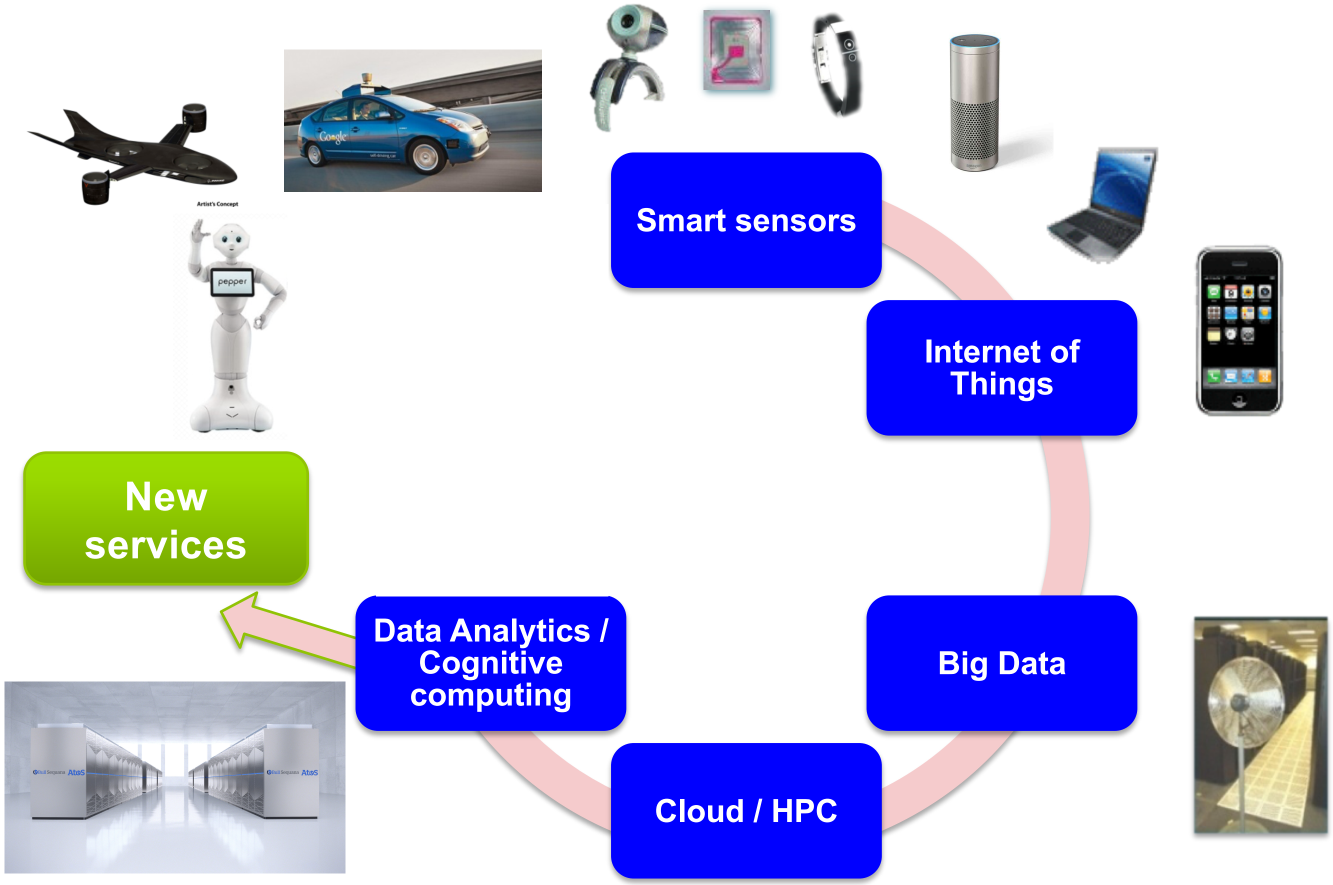


HOW MUCH COULD WE SAVE WITH CONNECTED MACHINES?

A 1% improvement in efficiency in these five industries could add up to $276 Billion over 15 years

RAIL	AVIATION	HEALTHCARE	POWER	OIL & GAS
by giving workers a real-time picture of rail networks	by planning more efficient flight paths and using smart engines that tell crews when they need maintenance	by helping workers locate and use mobile equipment more efficiently	by monitoring equipment better and predicting other potential network problems	by cutting fuel and operating costs, and by making equipment more available and productive

Mainstream “business” model



Smart sensors

Internet of Things

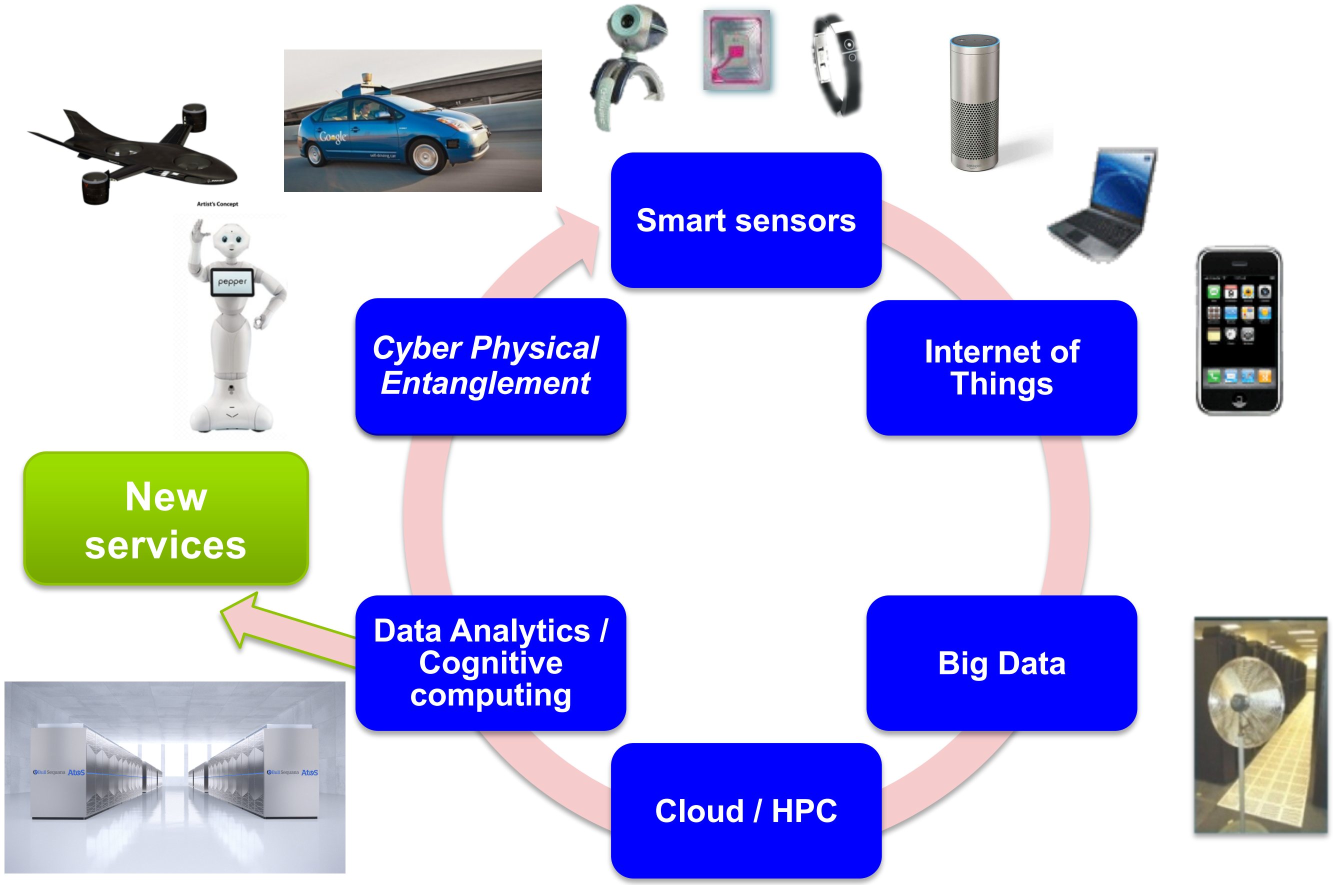
Big Data

Cloud / HPC

Data Analytics / Cognitive computing

New services

HPC in the loop



Smart sensors

Internet of Things

Big Data

Cloud / HPC

Data Analytics / Cognitive computing

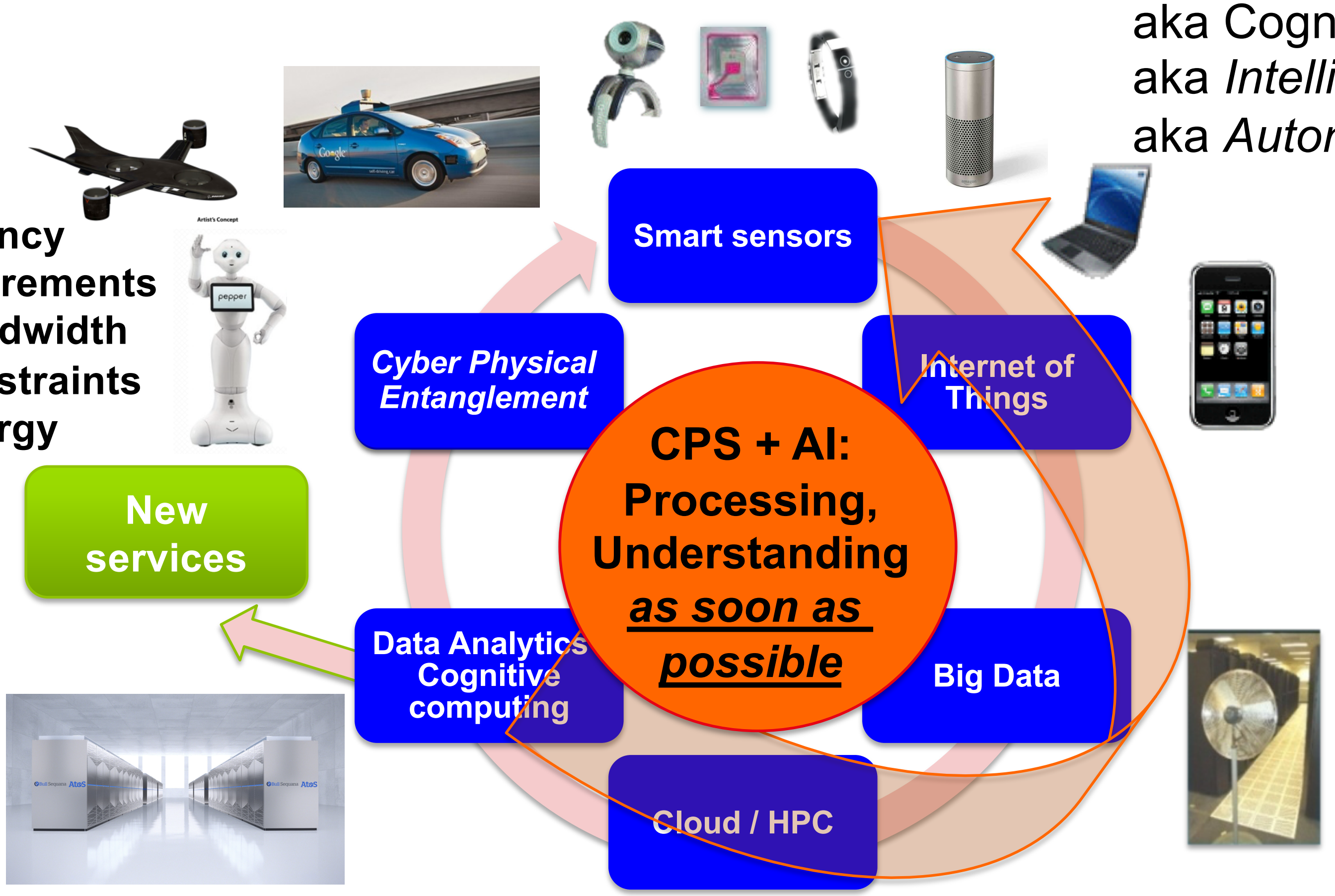
Cyber Physical Entanglement

New services

HIGH PERFORMANCE SYSTEMS IN THE LOOP

aka Cognitive CPS
aka *Intelligent Embedded Systems*
aka *Autonomous CPS (ACPS)*

Reduce latency
Safety requirements
Reduce bandwidth
Privacy constraints
Reduce energy

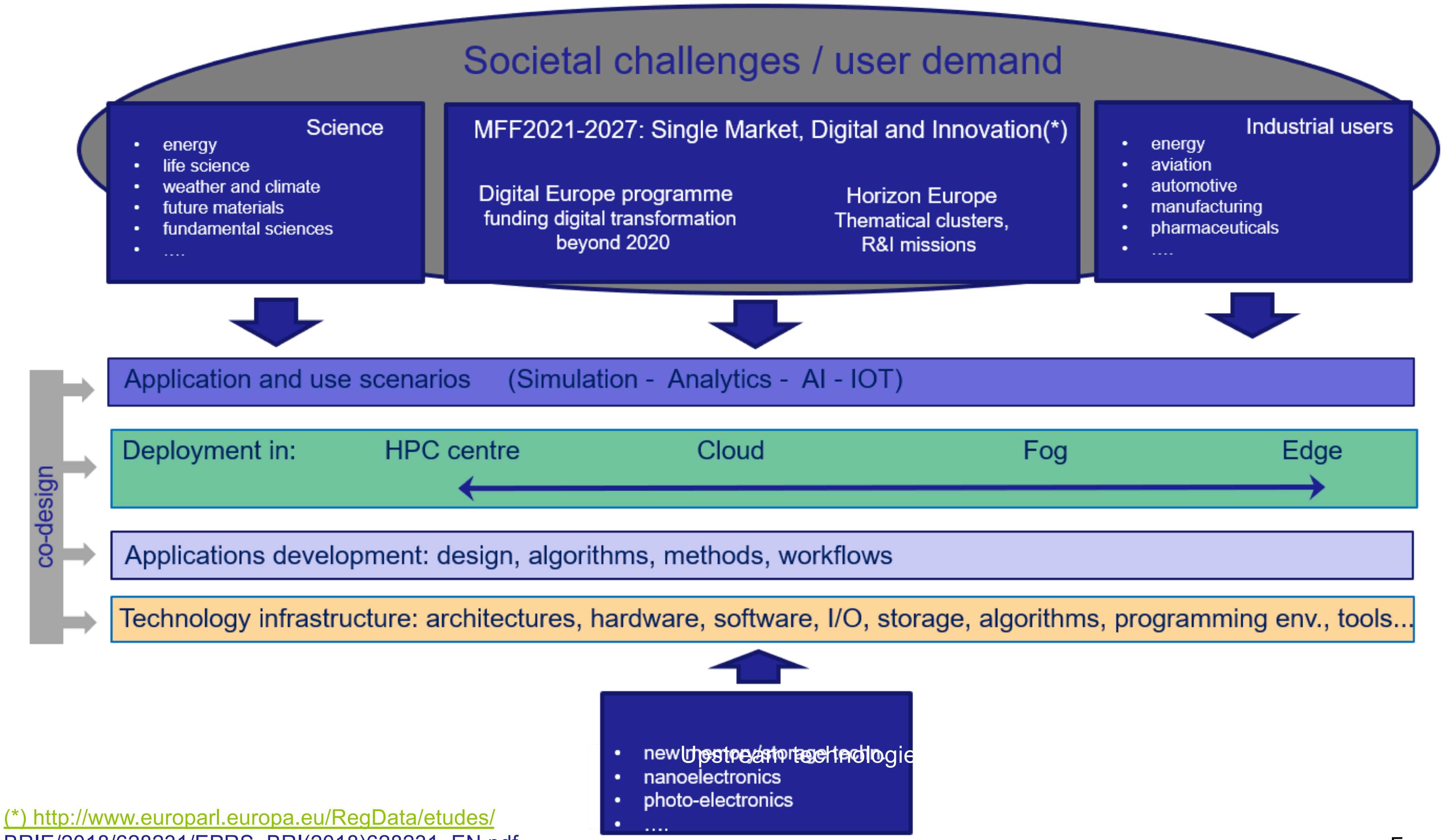

Artist's Concept
pepper
Smart sensors
Cyber Physical Entanglement
Internet of Things
CPS + AI: Processing, Understanding as soon as possible
Data Analytics Cognitive computing
Big Data
Cloud / HPC
New services

Enabling Intelligent data processing at the edge:
Fog computing
Edge computing
Stream analytics
Fast data…

True collaboration between edge devices and the HPC/cloud
➪ creating a
continuum

ENABLING EDGE INTELLIGENCE
Transforming **data** into ***information as early as possible***

SRA-4: THE INCREASING INTERPLAY OF SIMULATION, AI, IOT AND ANALYTICS



Societal challenges / user demand

Science

- energy
- life science
- weather and climate
- future materials
- fundamental sciences
- ….

MFF2021-2027: Single Market, Digital and Innovation(*)

Digital Europe programme
funding digital transformation
beyond 2020

Horizon Europe
Thematical clusters,
R&I missions

Industrial users

- energy
- aviation
- automotive
- manufacturing
- pharmaceuticals
- ….

Application and use scenarios (Simulation - Analytics - AI - IOT)

Deployment in: HPC centre Cloud Fog Edge

Applications development: design, algorithms, methods, workflows

Technology infrastructure: architectures, hardware, software, I/O, storage, algorithms, programming env., tools…

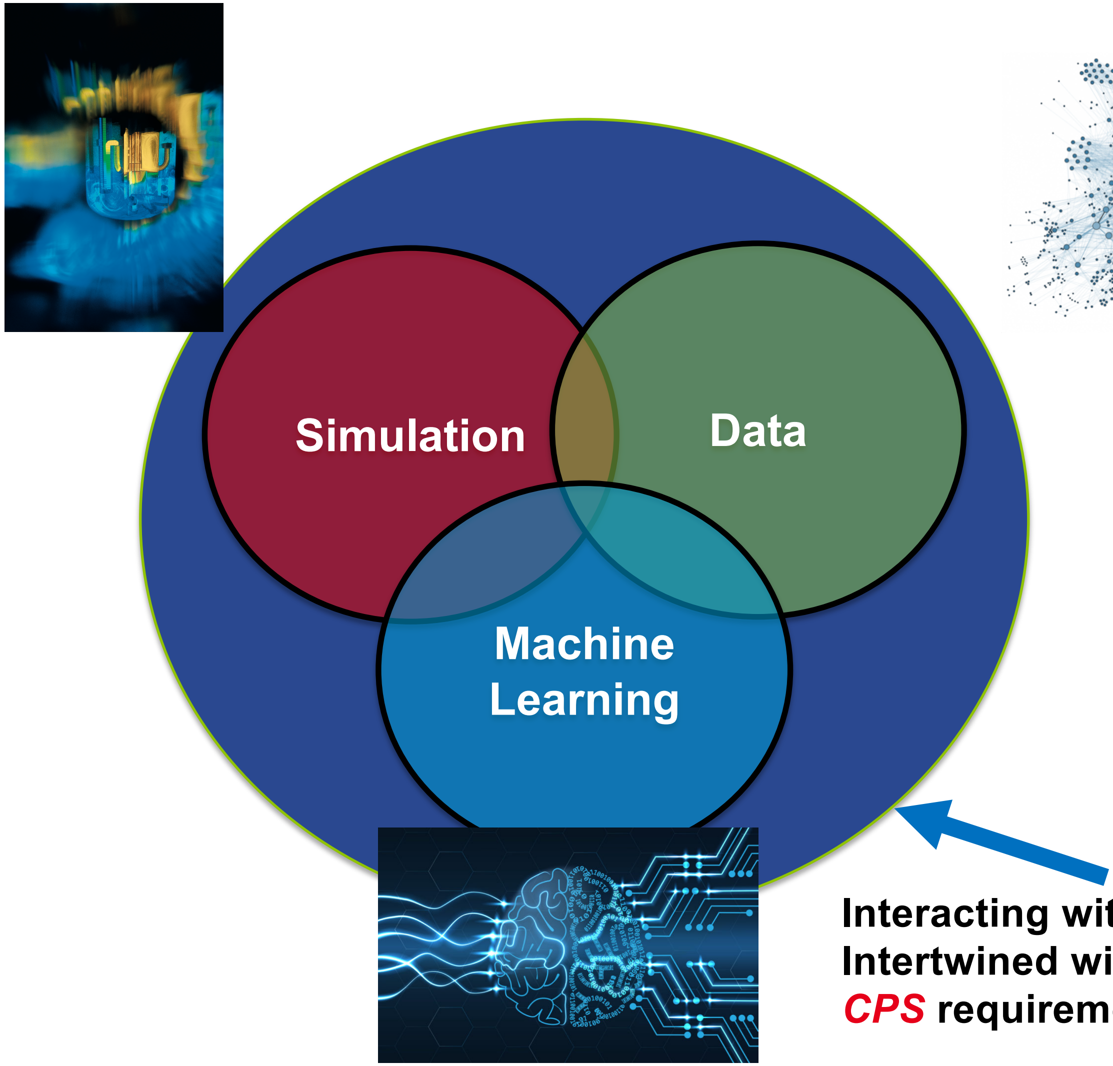
co-design

Upstream technologie

- new memory/storage techn
- nanoelectronics
- photo-electronics
- ….

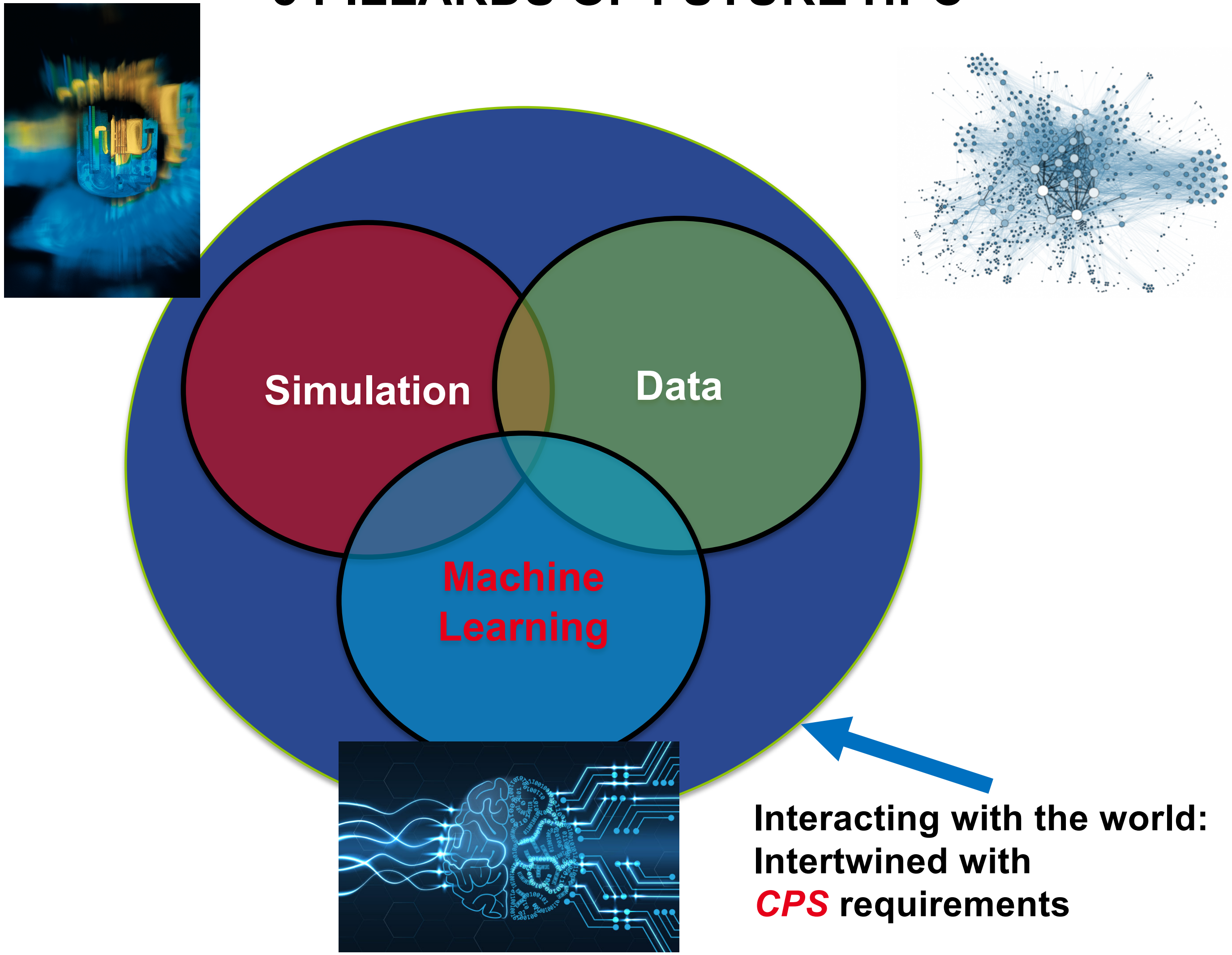
(*) http://www.europarl.europa.eu/RegData/etudes/BRIE/2018/628231/EPRS_BRI(2018)628231_EN.pdf

From ETP4HPC



Interacting with the world:
Intertwined with
***CPS* requirements**

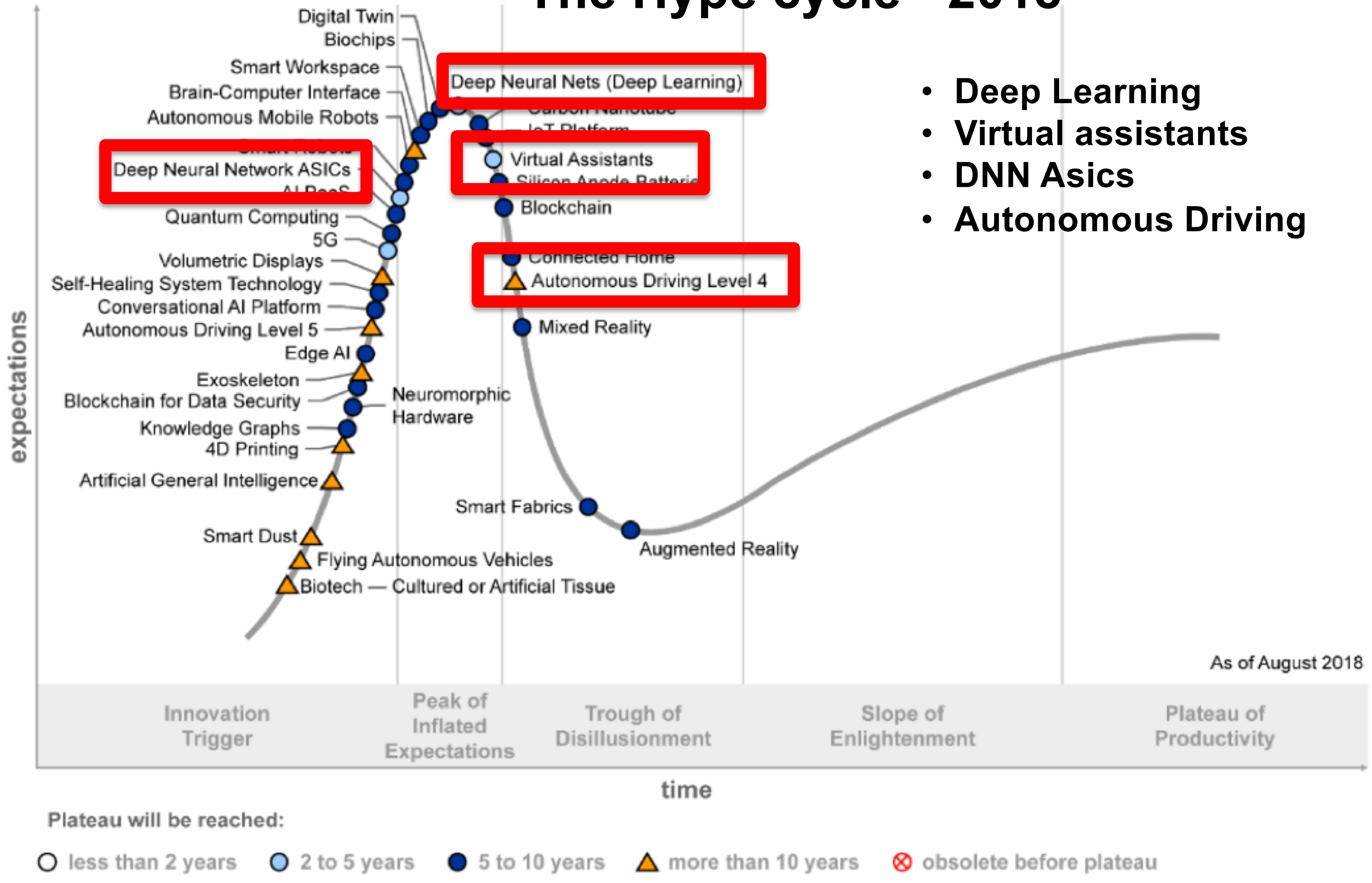
3 PILLARDS OF FUTURE HPC



Interacting with the world:
Intertwined with
***CPS* requirements**

ARTIFICIAL INTELLIGENCE (DEEP LEARNING) LOADS

The Hype cycle - 2018



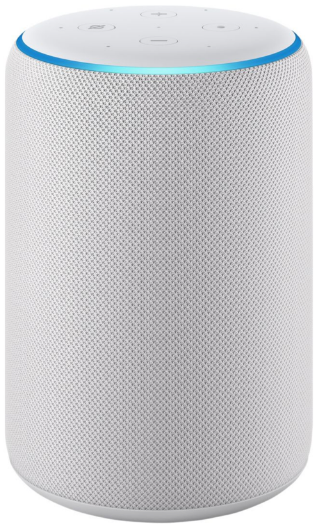
© 2018 Gartner, Inc.

- **Deep Learning**
- **Virtual assistants**
- **DNN Asics**
- **Autonomous Driving**

ONE ASPECT OF AI: PERSONAL ASSISTANTS....

Google Assistant
(1 billon devices)

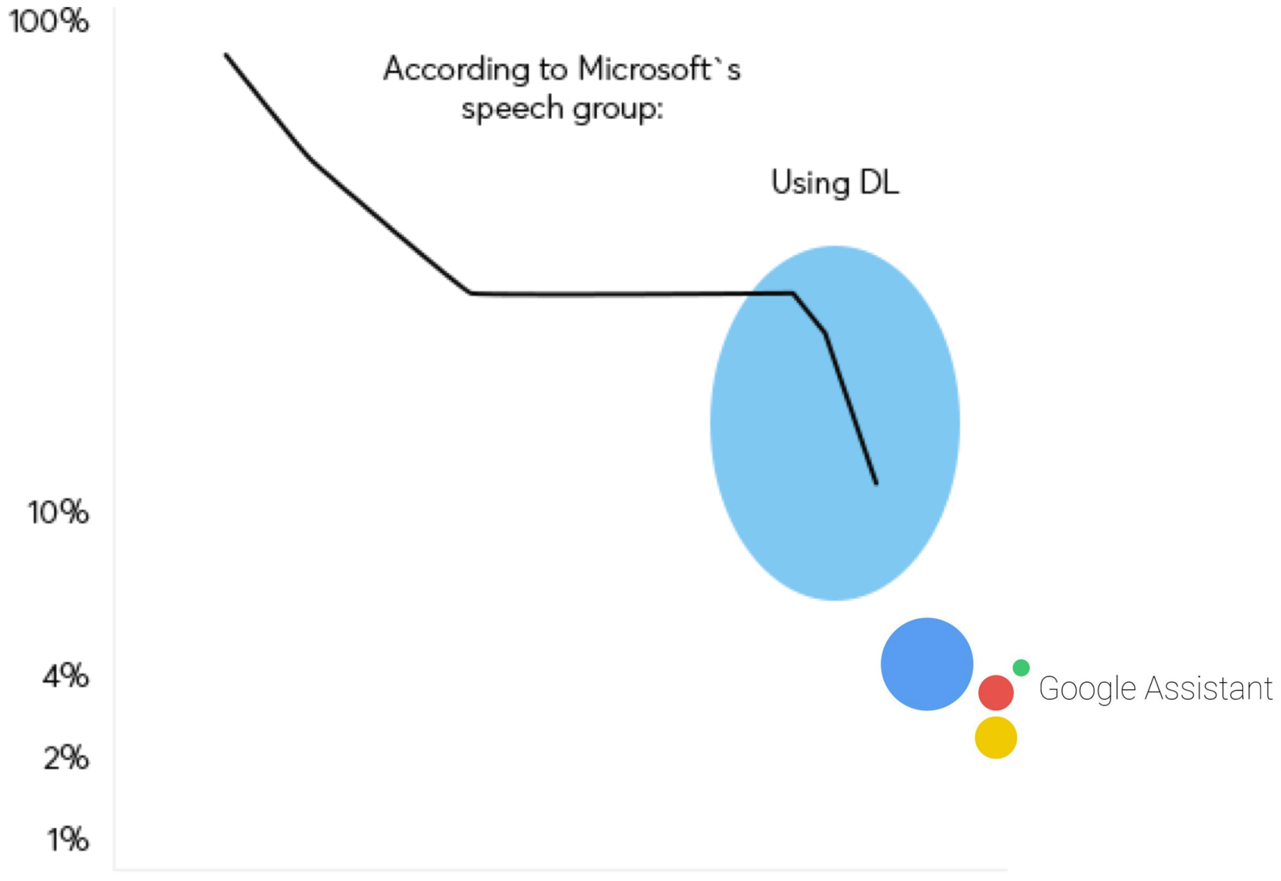
Apple Siri
(+500 millions devices)



Amazon Alexa
(+100 millions devices)

Baidu’s DuerOS
(+100 millions devices)

DEEP LEARNING AND VOICE RECOGNITION



Deep Learning in **Speech Recognition**

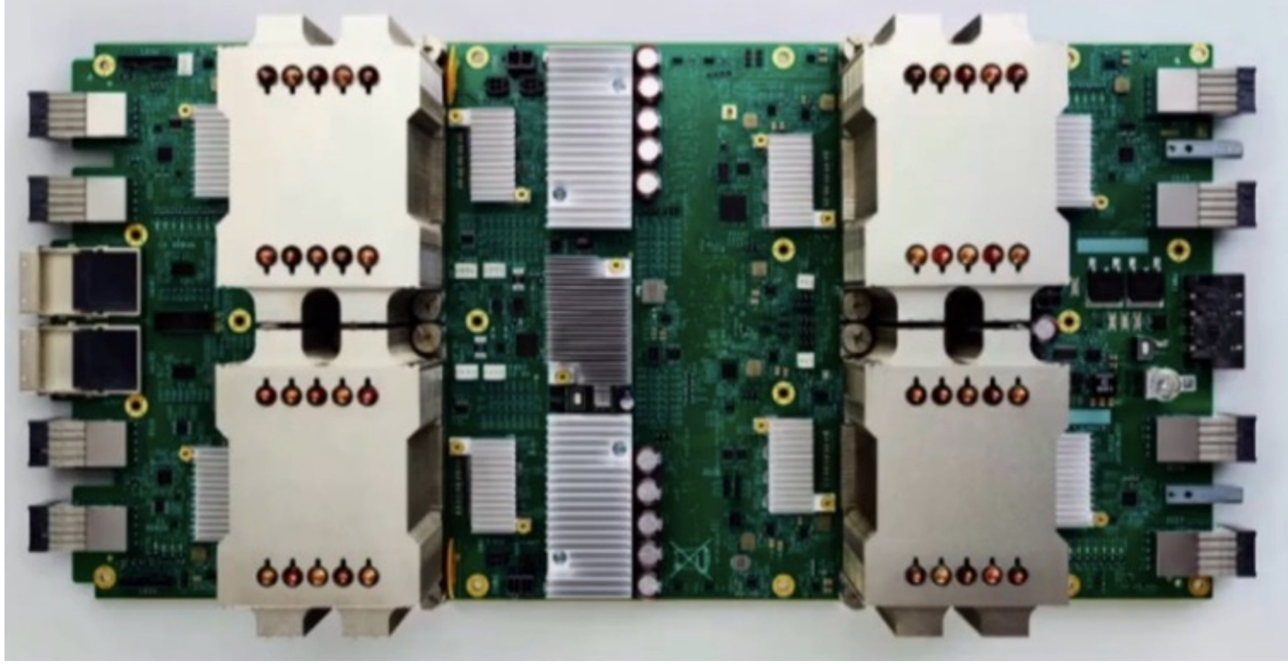
DEEP LEARNING AND VOICE RECOGNITION

" The need for TPUs really emerged about six years ago, when we started using computationally expensive deep learning models in more and more places throughout our products. The computational expense of using these models had us worried. If we considered a scenario where **people use Google voice search for just three minutes a day** and we ran deep neural nets for our speech recognition system on the processing units we were using, we would have had to ***double the number of Google data centers*****!"**

[https://cloudplatform.googleblog.com/2017/04/quantifying-the-performance-of-the-TPU-our-first-machine-learning-chip.html]

GOOGLE'S CUSTOMIZED HARDWARE…

… required to increase energy efficiency
with **accuracy adapted to the use (e.g. float 16)**



Google's TPU2 : training and inference in a **180 teraflops$_{16}$ (180 x 10^{12} Flops$_{16}$)** board (over 200W per TPU2 chip according to the size of the heat sink)

GOOGLE'S CUSTOMIZED TPU (V2) HARDWARE…

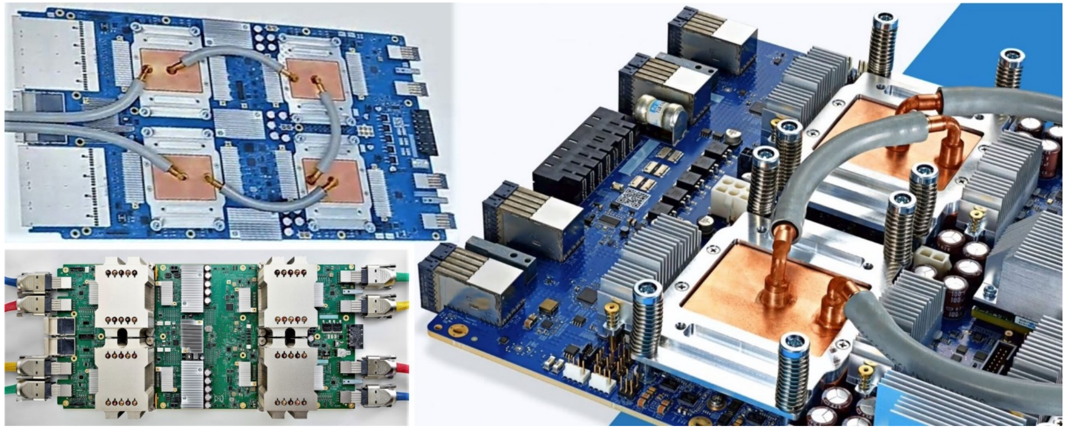
… required to increase energy efficiency
with accuracy adapted to the use (e.g. float 16)

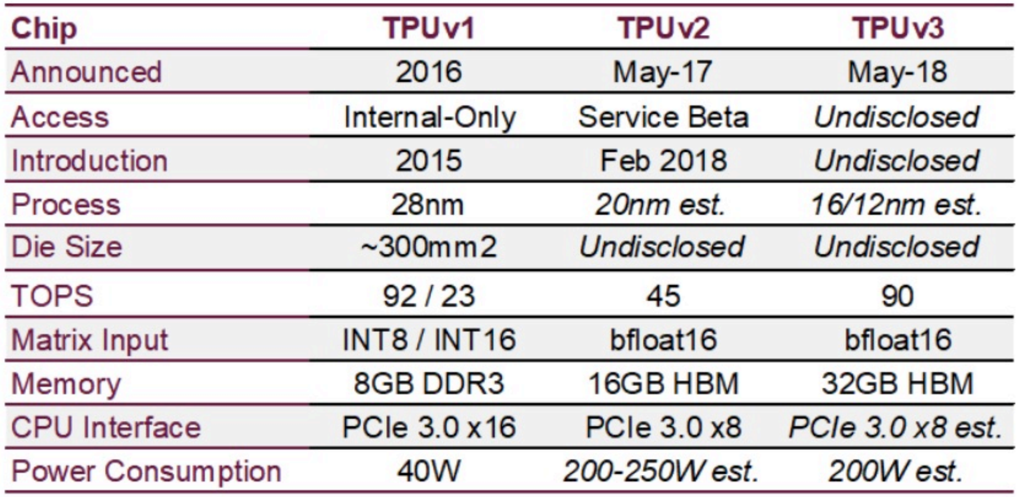
Google's TPU2 : 11.5 petaflops$_{16}$ of machine learning number crunching
(and guessing about **400+ KW**…, 100+ GFlops$_{16}$/W)

From Google

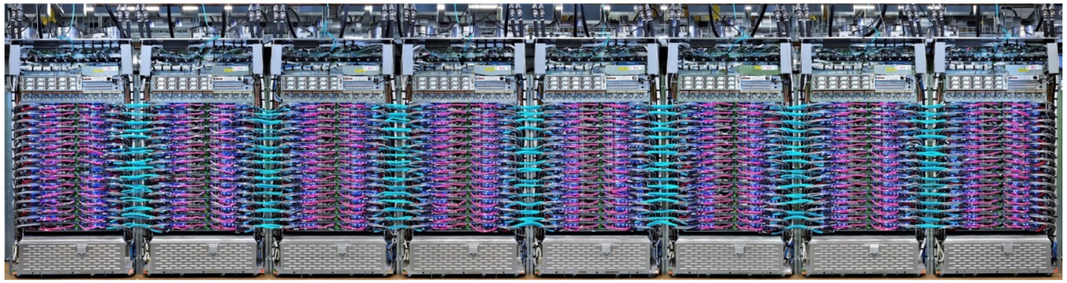
Peta = 10^{15} = million of milliard

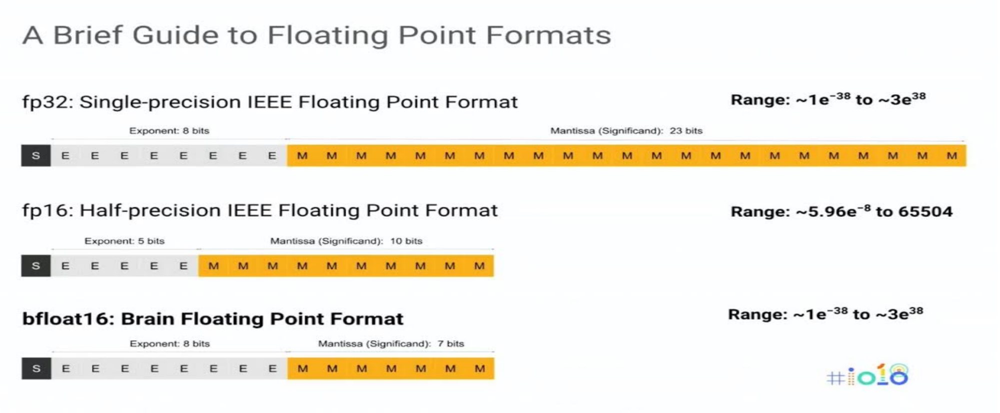
GOOGLE'S CUSTOMIZED TPU (V3) HARDWARE…





Chip	TPUv1	TPUv2	TPUv3
Announced	2016	May-17	May-18
Access	Internal-Only	Service Beta	*Undisclosed*
Introduction	2015	Feb 2018	*Undisclosed*
Process	28nm	*20nm est.*	*16/12nm est.*
Die Size	~300mm2	*Undisclosed*	*Undisclosed*
TOPS	92 / 23	45	90
Matrix Input	INT8 / INT16	bfloat16	bfloat16
Memory	8GB DDR3	16GB HBM	32GB HBM
CPU Interface	PCIe 3.0 x16	PCIe 3.0 x8	*PCIe 3.0 x8 est.*
Power Consumption	40W	*200-250W est.*	*200W est.*





A Brief Guide to Floating Point Formats

fp32: Single-precision IEEE Floating Point Format — **Range: ~1e^{-38} to ~3e^{38}**

Exponent: 8 bits
Mantissa (Significand): 23 bits
S E E E E E E E E M M M M M M M M M M M M M M M M M M M M M M M

fp16: Half-precision IEEE Floating Point Format — **Range: ~5.96e^{-8} to 65504**

Exponent: 5 bits
Mantissa (Significand): 10 bits
S E E E E E M M M M M M M M M M

bfloat16: Brain Floating Point Format — **Range: ~1e^{-38} to ~3e^{38}**

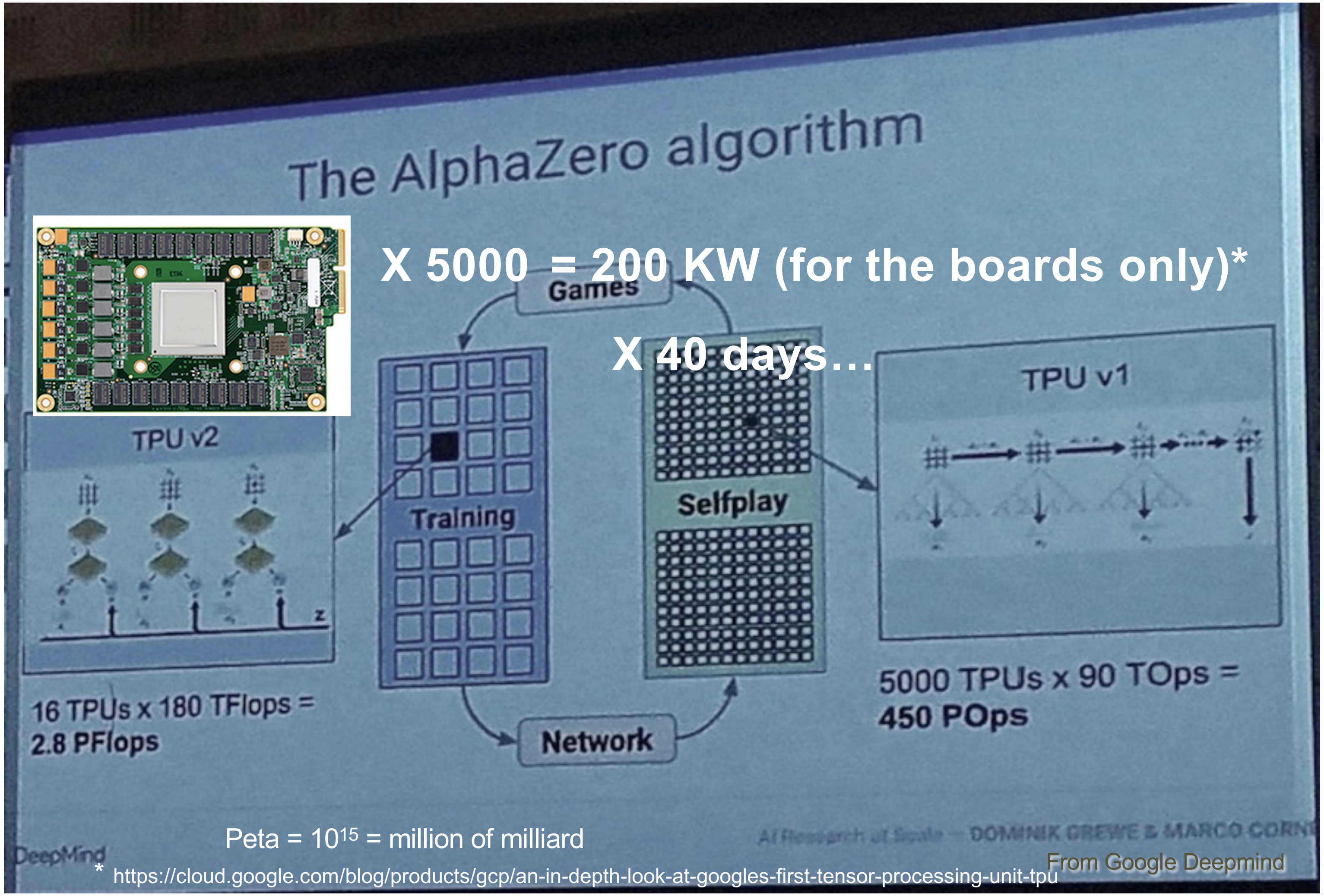
Exponent: 8 bits
Mantissa (Significand): 7 bits
S E E E E E E E E M M M M M M M
#io18

From https://www.nextplatform.com/2018/05/10/tearing-apart-googles-tpu-3-0-ai-coprocessor/

ALPHAGO ZERO: SELF-PLAYING TO

From doi:10.1038/nature24270 (Received 07 April 2017)

ALPHAZERO FROM DEEP MIND : COMPUTING RESOURCES



The AlphaZero algorithm

X 5000 = 200 KW (for the boards only)*

X 40 days…

Games

Training

Selfplay

Network

TPU v2

TPU v1

16 TPUs x 180 TFlops = 2.8 PFlops

5000 TPUs x 90 TOps = 450 POps

DeepMind

AI Research at Scale — DOMINIK GREWE & MARCO CORN

Peta = 10^{15} = million of milliard

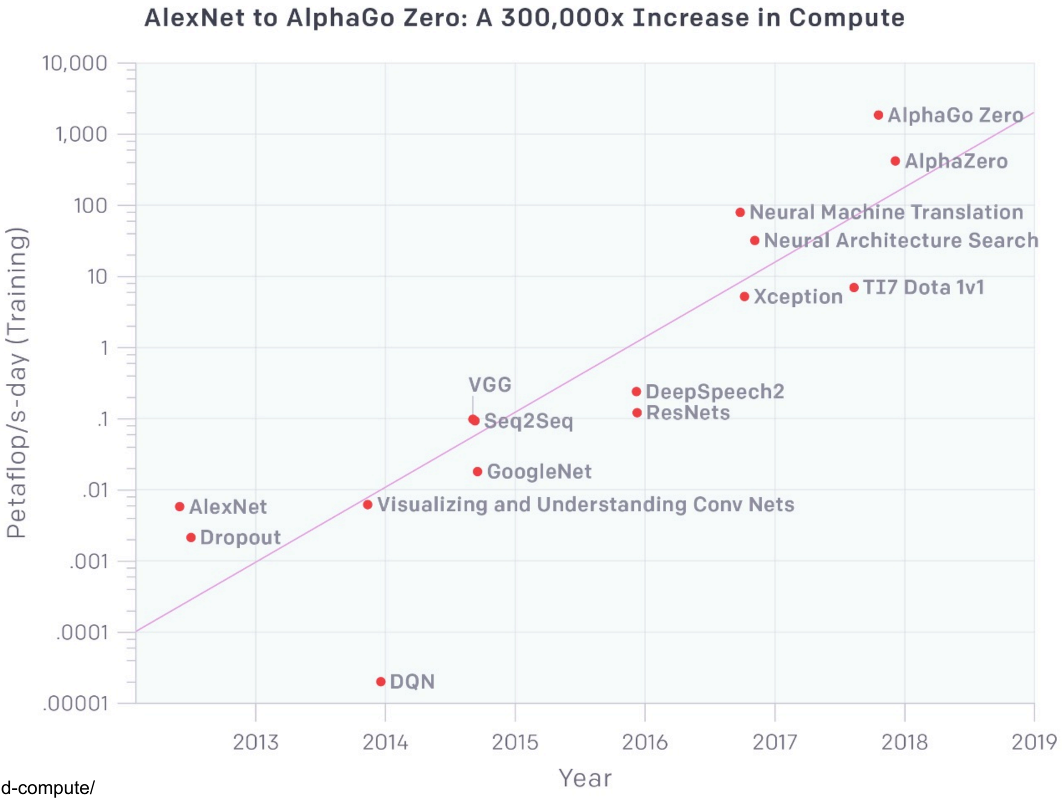
From Google Deepmind

* https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

EXPONENTIAL INCREASE OF COMPUTING POWER FOR AI TRAINING

"Since 2012, the amount of compute used in the largest AI training runs has been increasing exponentially with a 3.5 month-doubling time…

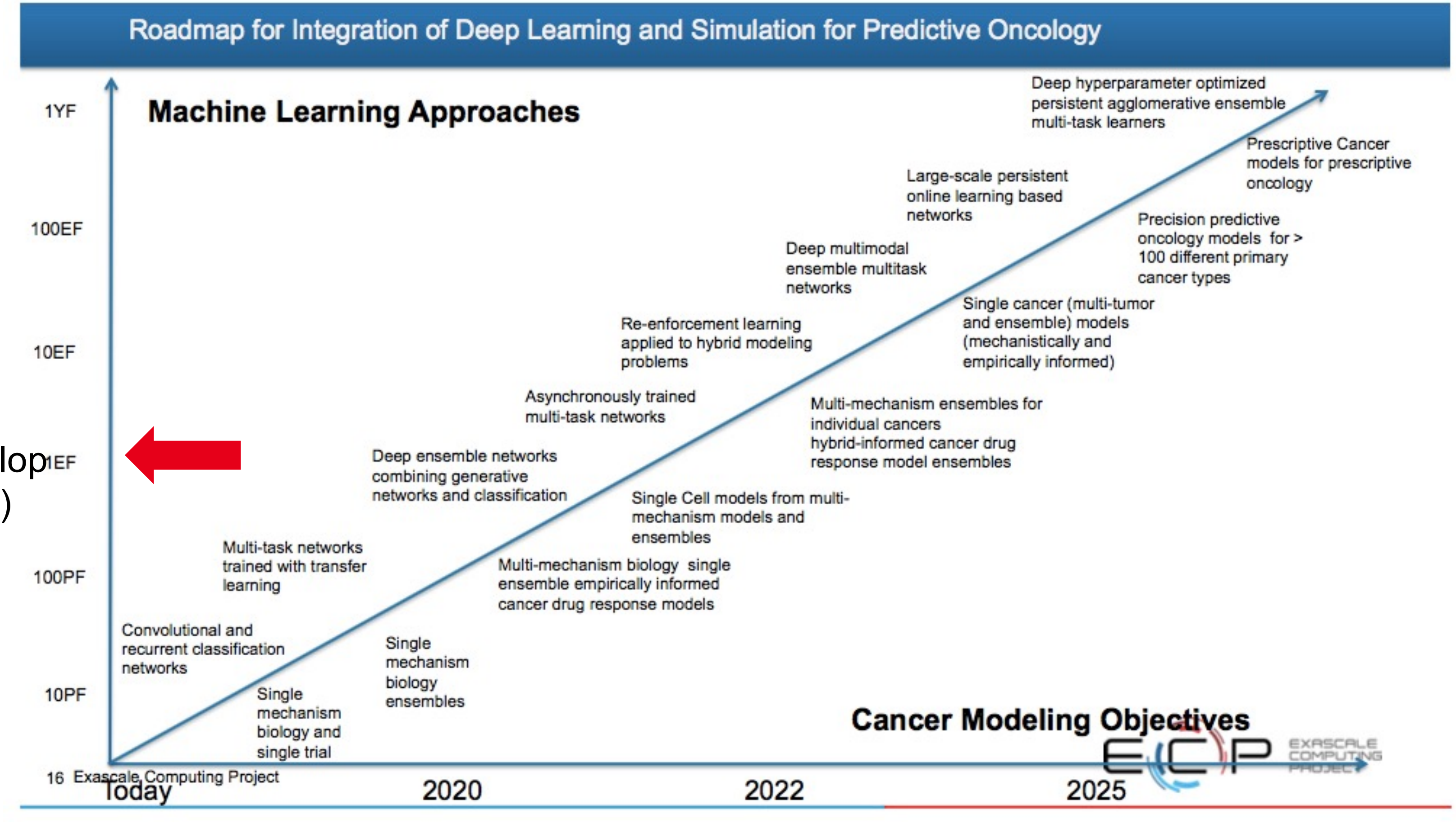
(by comparison, Moore's Law had an 18-month doubling period)*"



Peta= 10^{15}

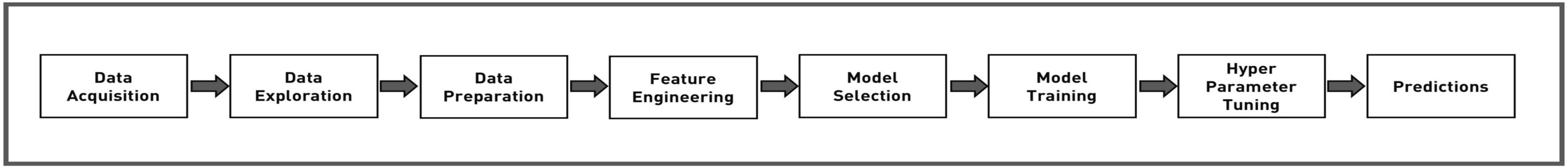
* https://blog.openai.com/ai-and-compute/

ALWAYS MORE COMPUTING RESOURCES


Roadmap for Integration of Deep Learning and Simulation for Predictive Oncology
Machine Learning Approaches
1YF
100EF
10EF
1EF
100PF
10PF
Deep hyperparameter optimized persistent agglomerative ensemble multi-task learners
Prescriptive Cancer models for prescriptive oncology
Large-scale persistent online learning based networks
Precision predictive oncology models for > 100 different primary cancer types
Deep multimodal ensemble multitask networks
Single cancer (multi-tumor and ensemble) models (mechanistically and empirically informed)
Re-enforcement learning applied to hybrid modeling problems
Asynchronously trained multi-task networks
Multi-mechanism ensembles for individual cancers hybrid-informed cancer drug response model ensembles
Deep ensemble networks combining generative networks and classification
Single Cell models from multi-mechanism models and ensembles
Multi-task networks trained with transfer learning
Multi-mechanism biology single ensemble empirically informed cancer drug response models
Convolutional and recurrent classification networks
Single mechanism biology ensembles
Single mechanism biology and single trial
Cancer Modeling Objectives
16 Exascale Computing Project
Today
2020
2022
2025
ECP EXASCALE COMPUTING PROJECT

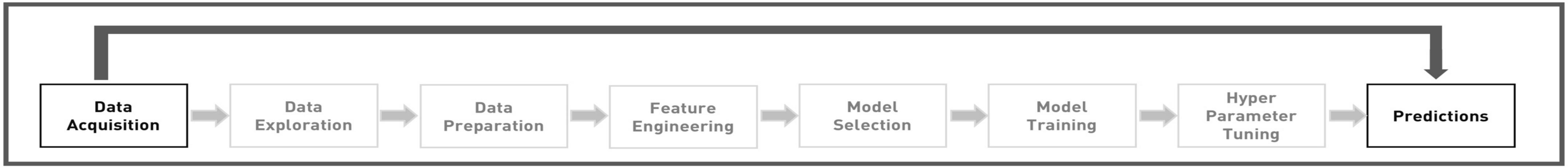
HPC: exaflop (10^{18} flops)

From Paul Messina, Argonne National Laboratory



Data Acquisition → Data Exploration → Data Preparation → Feature Engineering → Model Selection → Model Training → Hyper Parameter Tuning → Predictions

Traditional Machine Learning Workflow



Data Acquisition → Data Exploration → Data Preparation → Feature Engineering → Model Selection → Model Training → Hyper Parameter Tuning → Predictions

AutoML Workflow

From Forbes

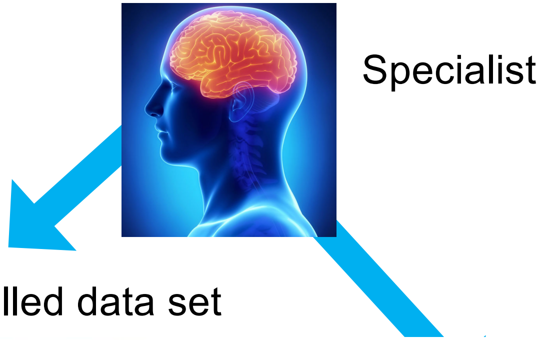
Auto-ML uses optimization approaches to select a "good" set of parameters "*automagically*"

- **It is generally very computing expansive (configuration space search)**
- **Use clever algorithms to avoid exploring all the configuration space**

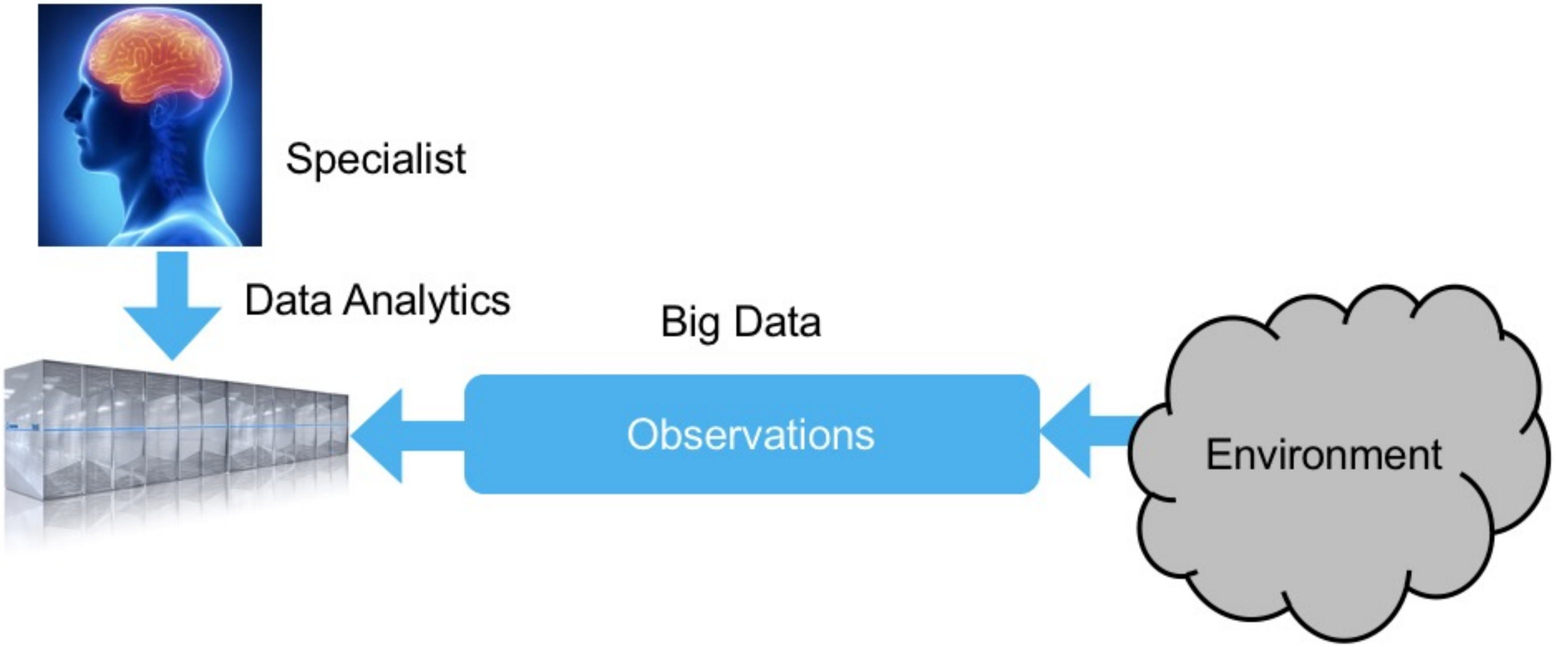
More details for example in http://automl.org

CHARACTERIZATION OF DEEP LEARNING LOADS

MACHINE LEARNING (DEEP LEARNING) LEARNING PHASE

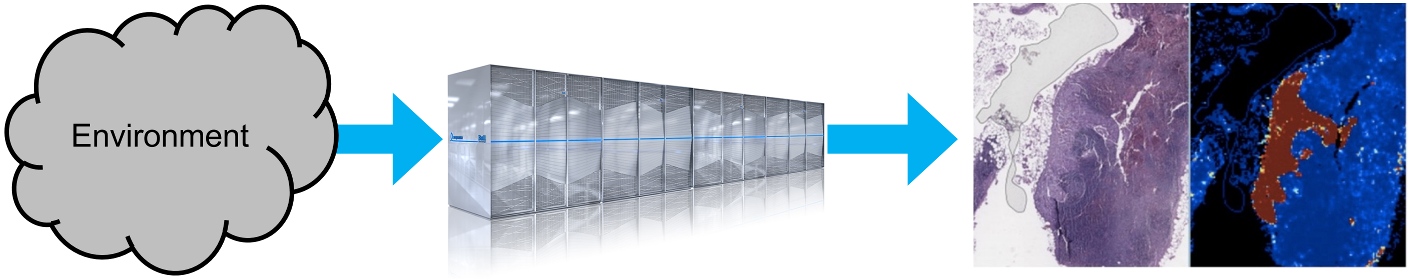


- Human defines the learning data set, not the algorithm
- Large set of input data for learning phase
- Low precision floating point
- Large number of operations
- (Stochastic) gradient descent



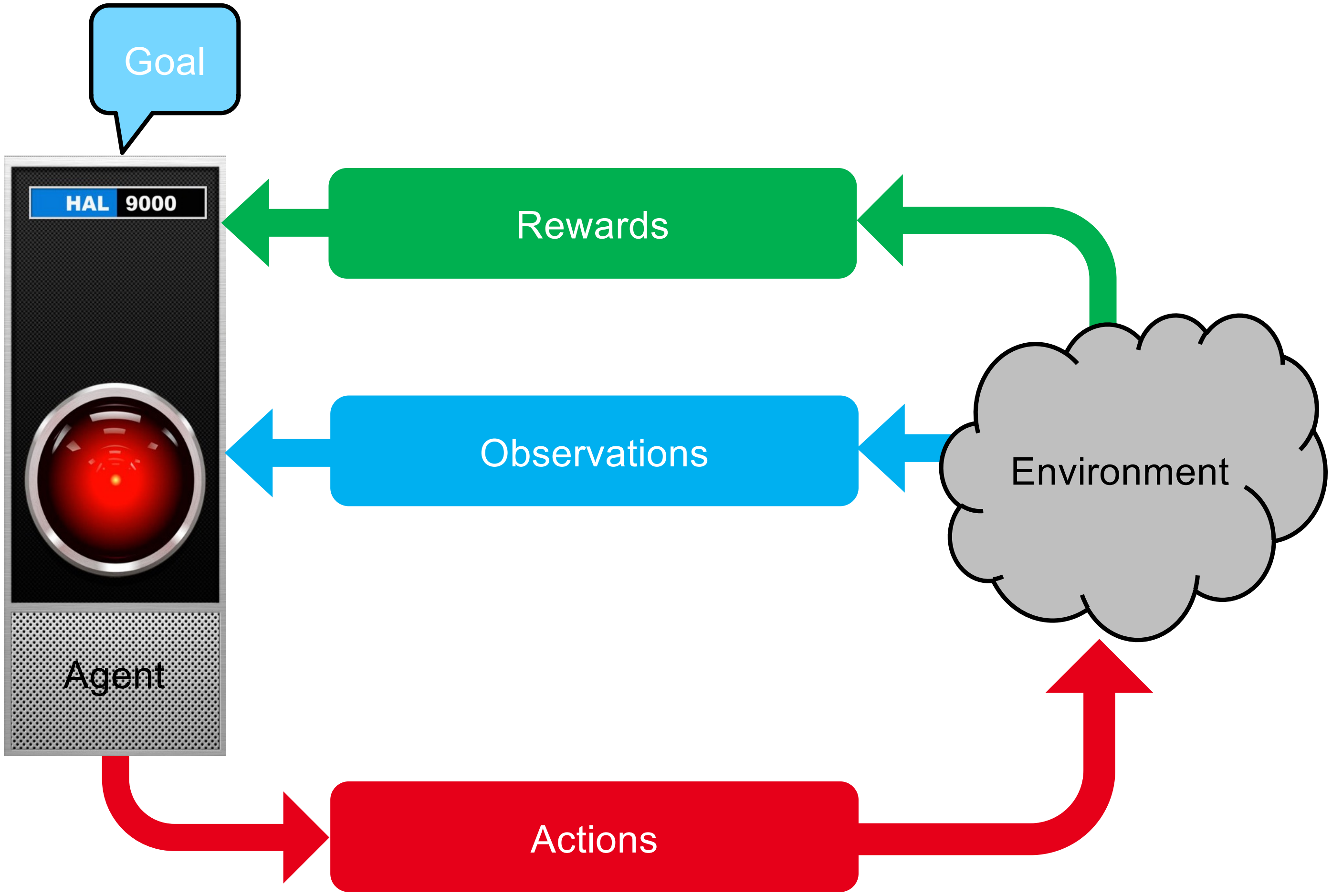
MACHINE LEARNING (DEEP LEARNING) INFERENCE PHASE

- Low precision arithmetic
- Medium to low number of operations
- Co-location computing and storage (“*computing in memory*”)
- Should satisfy the application non-functional requirements



But for large number of inferences (users)
-> more cloud like structure, high throughput

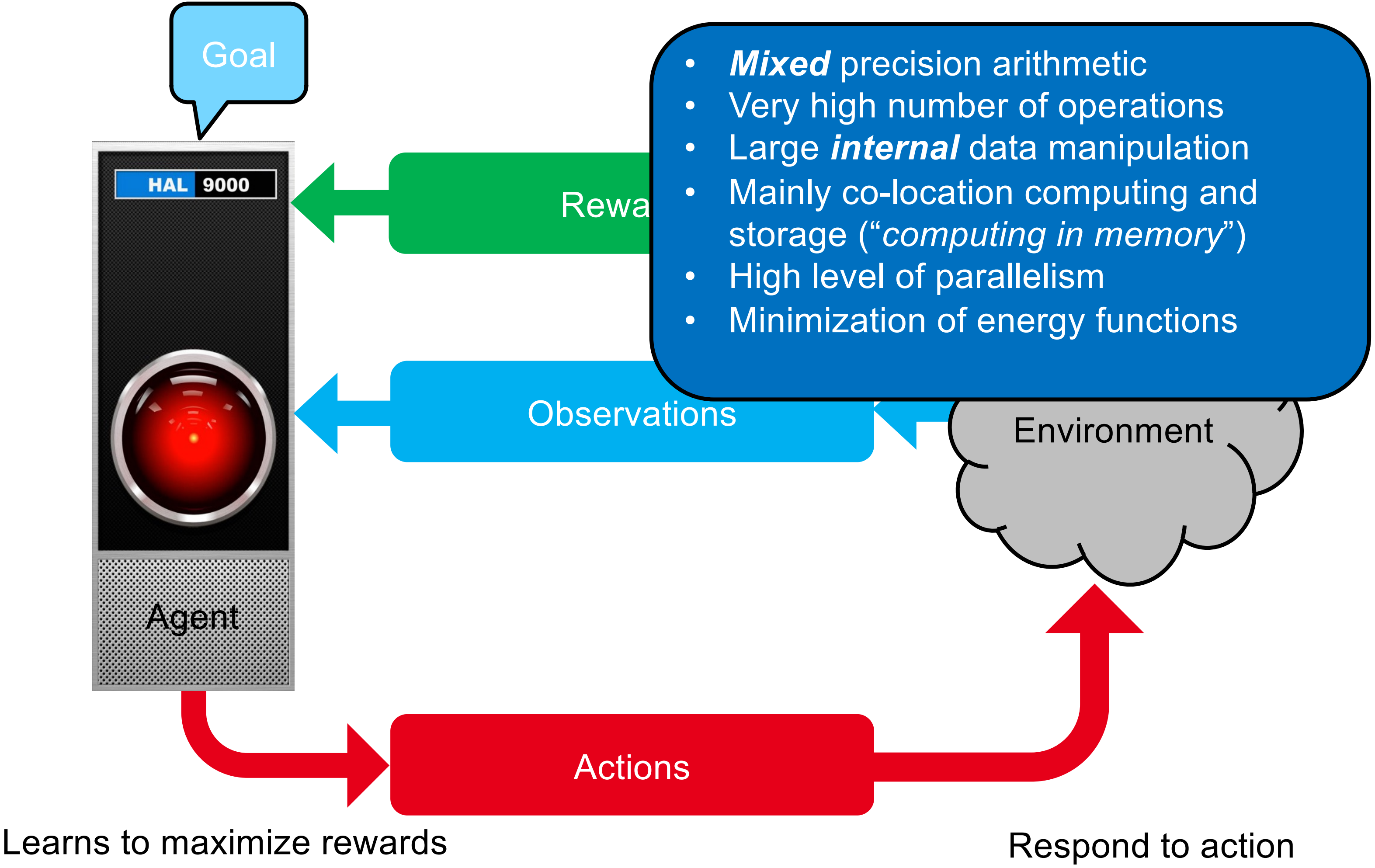
REINFORCEMENT LEARNING:
DYNAMIC PROGRAMMING + DEEP LEARNING



Learns to maximize rewards

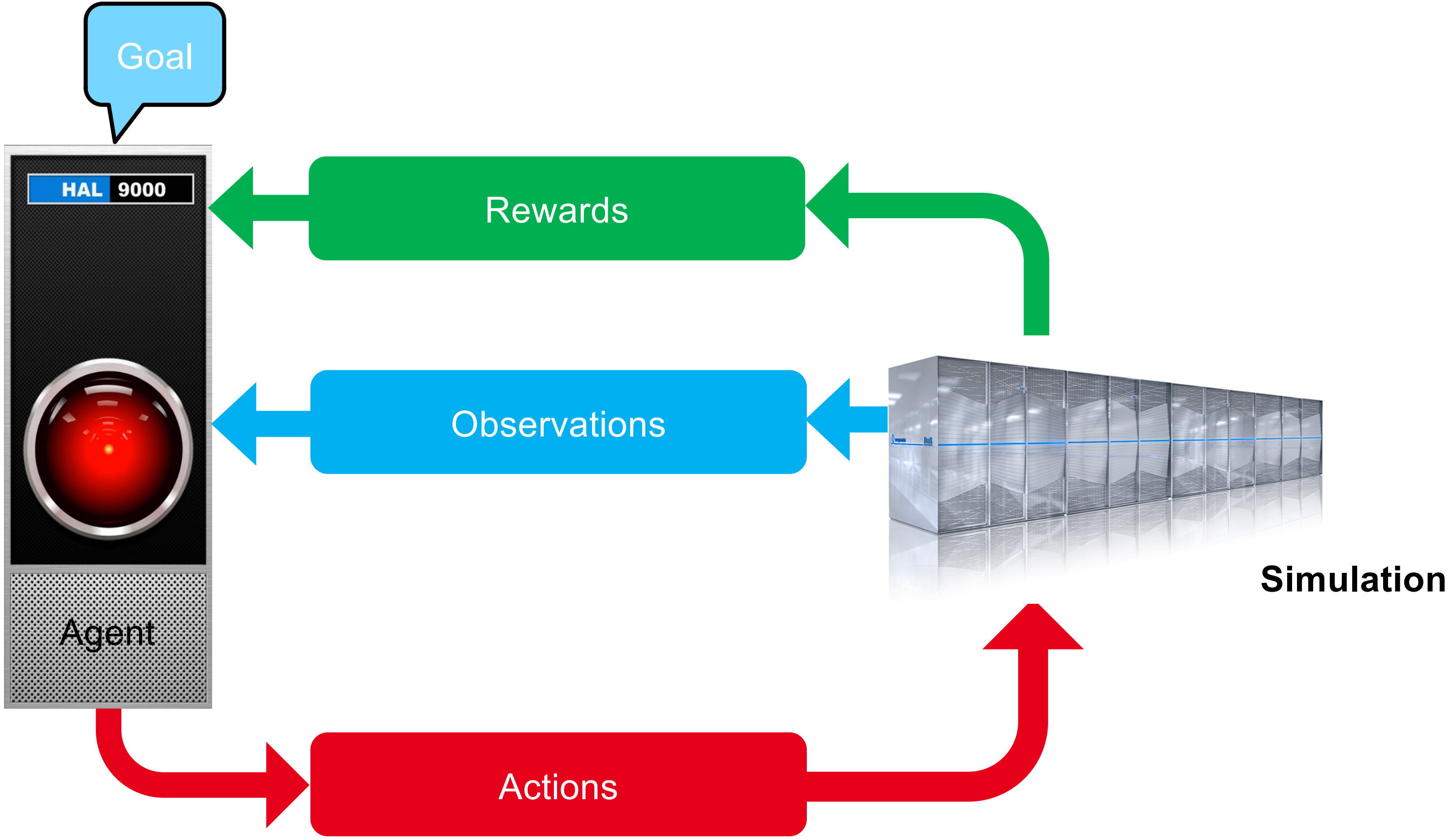
Respond to action

REINFORCEMENT LEARNING:
DYNAMIC PROGRAMMING + DEEP LEARNING



- ***Mixed*** precision arithmetic
- Very high number of operations
- Large ***internal*** data manipulation
- Mainly co-location computing and storage (“*computing in memory*”)
- High level of parallelism
- Minimization of energy functions

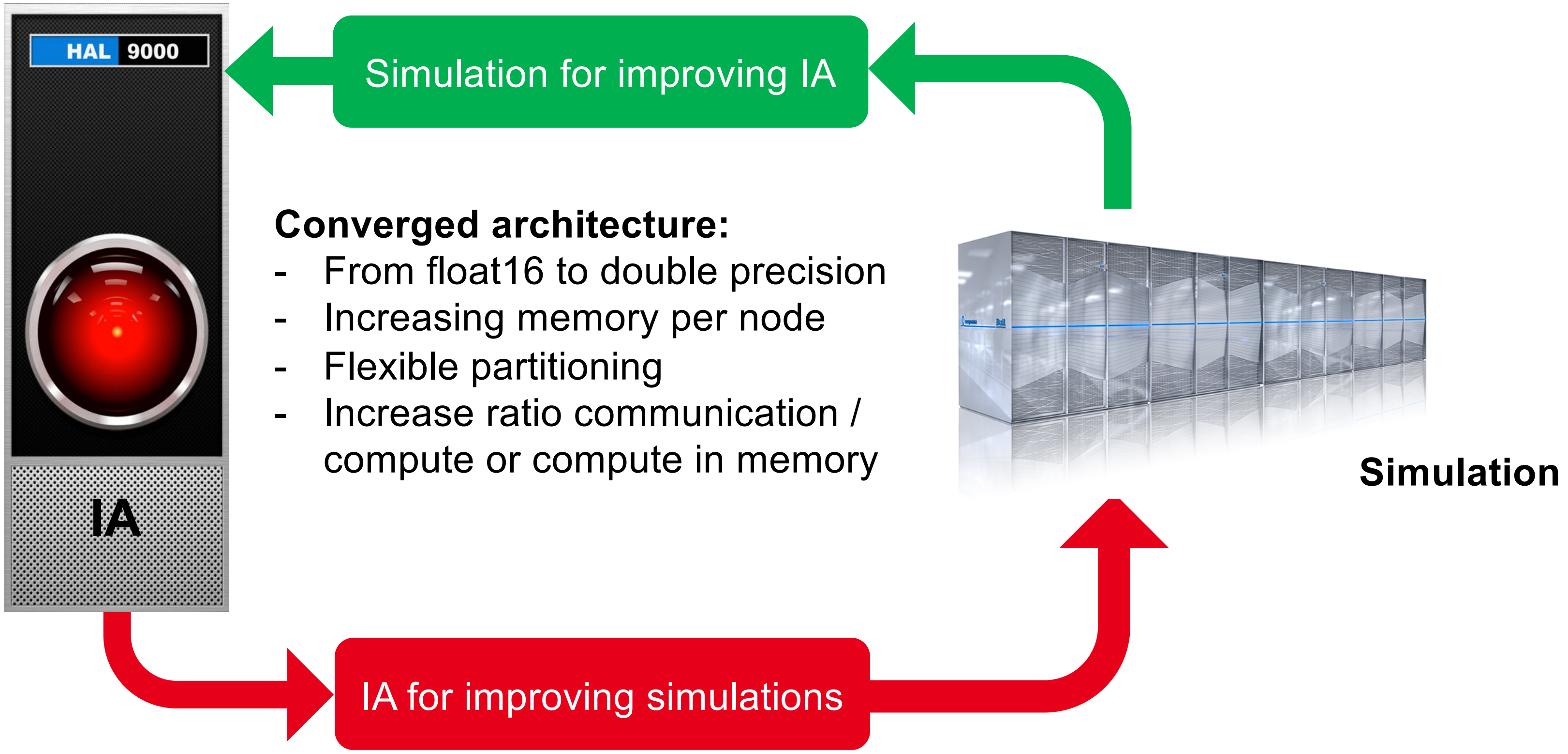
REINFORCEMENT LEARNING:
DYNAMIC PROGRAMMING + DEEP LEARNING



Learns to maximize rewards

Respond to action

COMPLEMENTARITY OF SIMULATION AND IA TECHNIQUES



Simulation for improving IA

Converged architecture:

- From float16 to double precision
- Increasing memory per node
- Flexible partitioning
- Increase ratio communication / compute or compute in memory

Simulation

IA for improving simulations

Outline

1) Evolution of application scope: the continuum

2) Hardware heterogeneity and orchestration

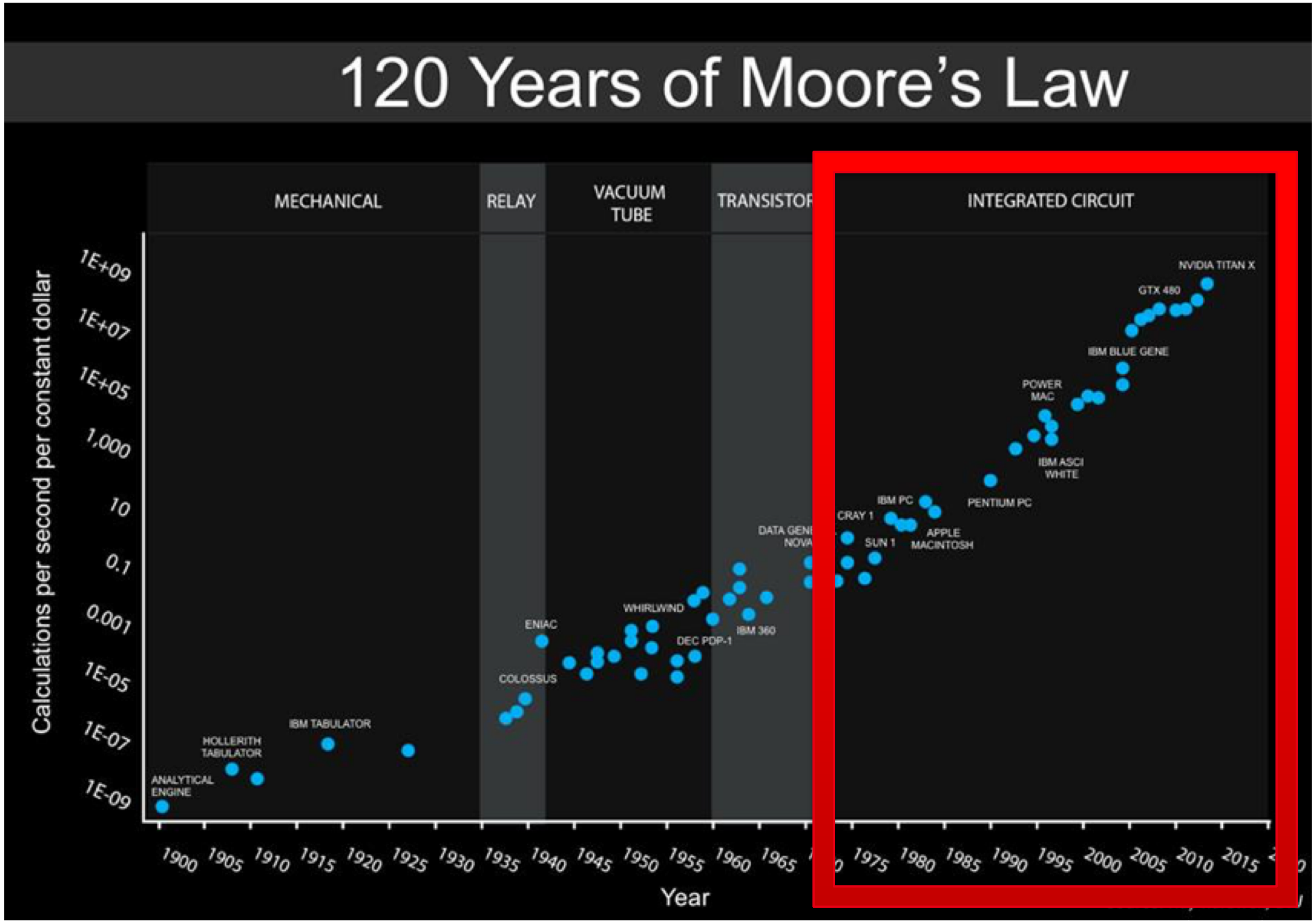
3) Software?

Outline

1) Evolution of application scope: the continuum
2) Hardware heterogeneity and orchestration
   1) End of Dennard’s scaling
   2) Heterogeneous accelerators
   3) Heterogeneous integration
3) Software?

END OF DENNARD’S SCALING

WHAT WILL BE THE NEXT TECHNOLOGY?



And after CMOS?

Exponential increase of performances in 33 years

To infinity and beyond…

Production car of 1985
Lamborghini Countach 5000QV
Max speed 300 Km/h

→

X 100 000 000
in 33 years

Star Trek Enterprise
Year: about 2290
27 times the speed of light?

THE END OF ~~MOORE'S LAW~~ DENNARD SCALING

Parameter (scale factor = a)	Classic Scaling
Dimensions	1/a
Voltage	1/a
Current	1/a
Capacitance	1/a
Power/Circuit	$1/a^2$
Power Density	1
Delay/Circuit	1/a

Everything was easy:

- Wait for the next technology node
- Increase frequency
- Decrease Vdd

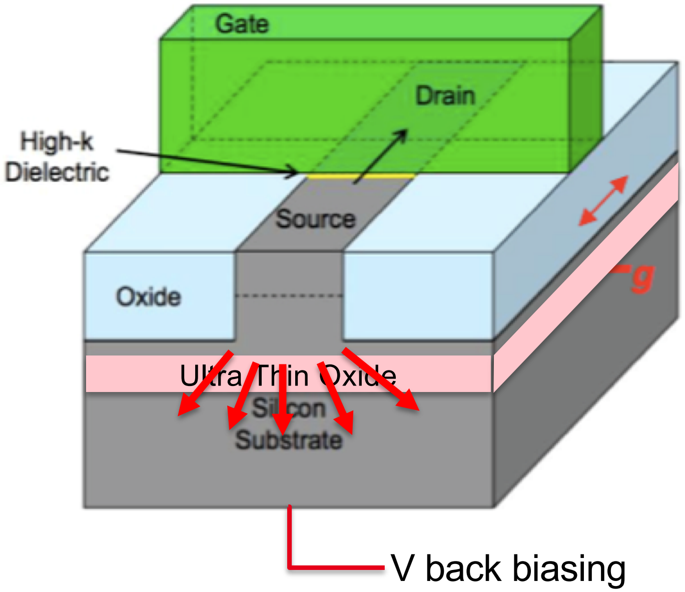
⇒ Similar increase of sequential performance

⇒ No need to recompile (except if architectural improvements)

Source: Krisztián Flautner "From niche to mainstream: can critical systems make the transition?"

Technology evolution

Transistor 2D



Fully Depleted Silicon on Insulator (FDSOI) Transistor

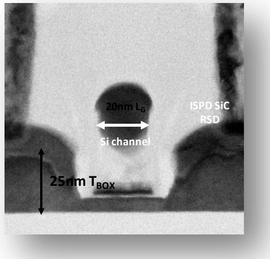
Technology evolution

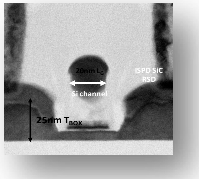
FDSOI

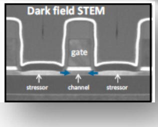
22FD

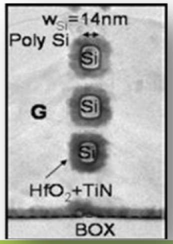
12FD

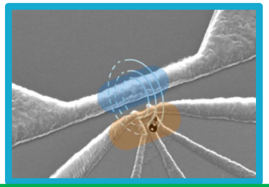
Next Gen











Silicon Quantum bits

Non planar / trigate / stacked Nanowires

FinFET

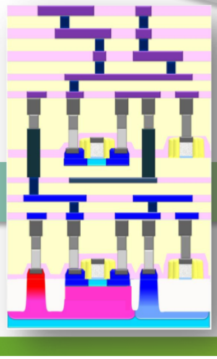
28nm · 14nm · 10nm · 2017 · 2019 · 7nm · 5nm

Disruptive scaling

Alternative to scaling and diversification

Hybrid logic

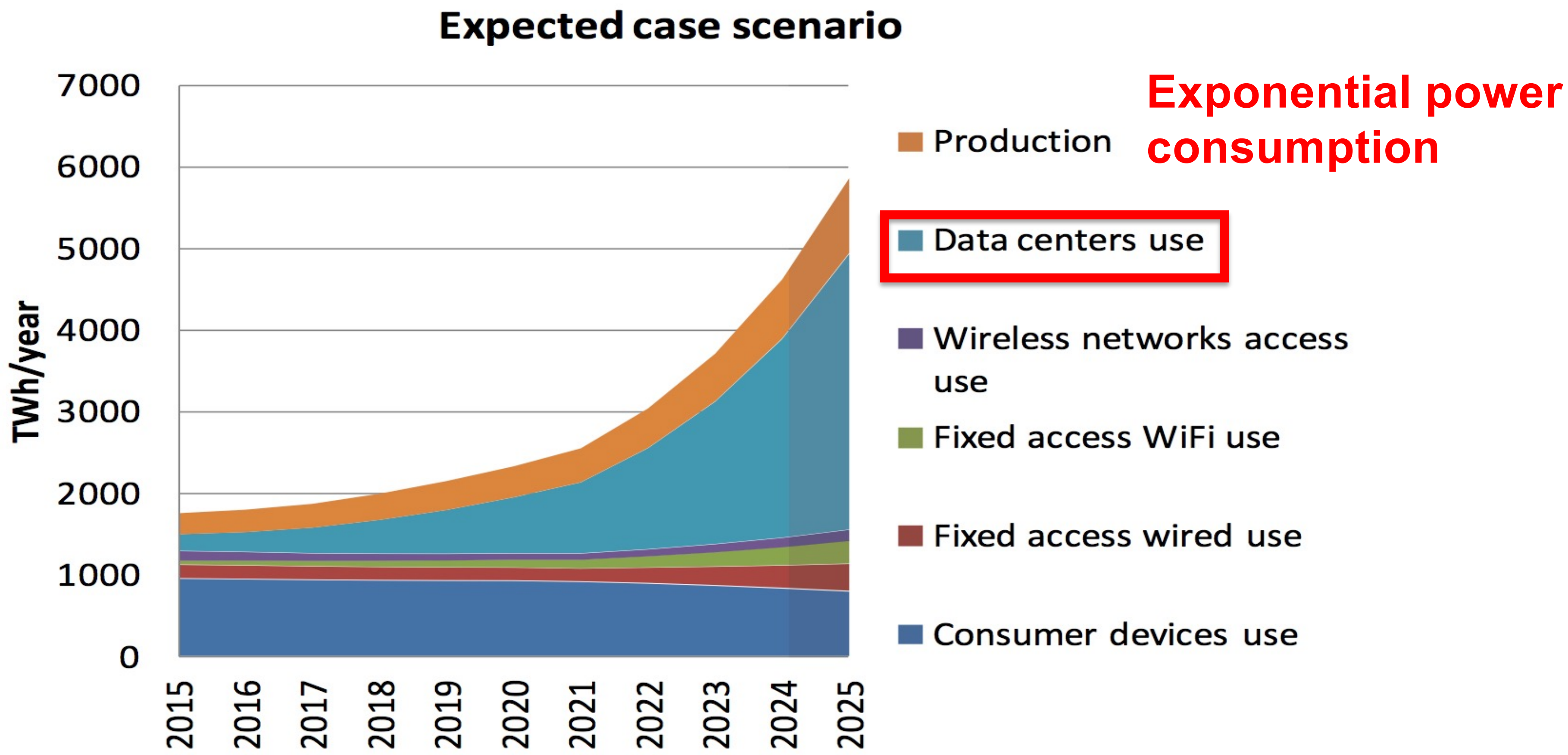
- Steep slope devices
- Mechanical switches
- Si Quantum bits



Monolithic 3D for 3D VLSI

HETEROGENEOUS ACCELERATORS

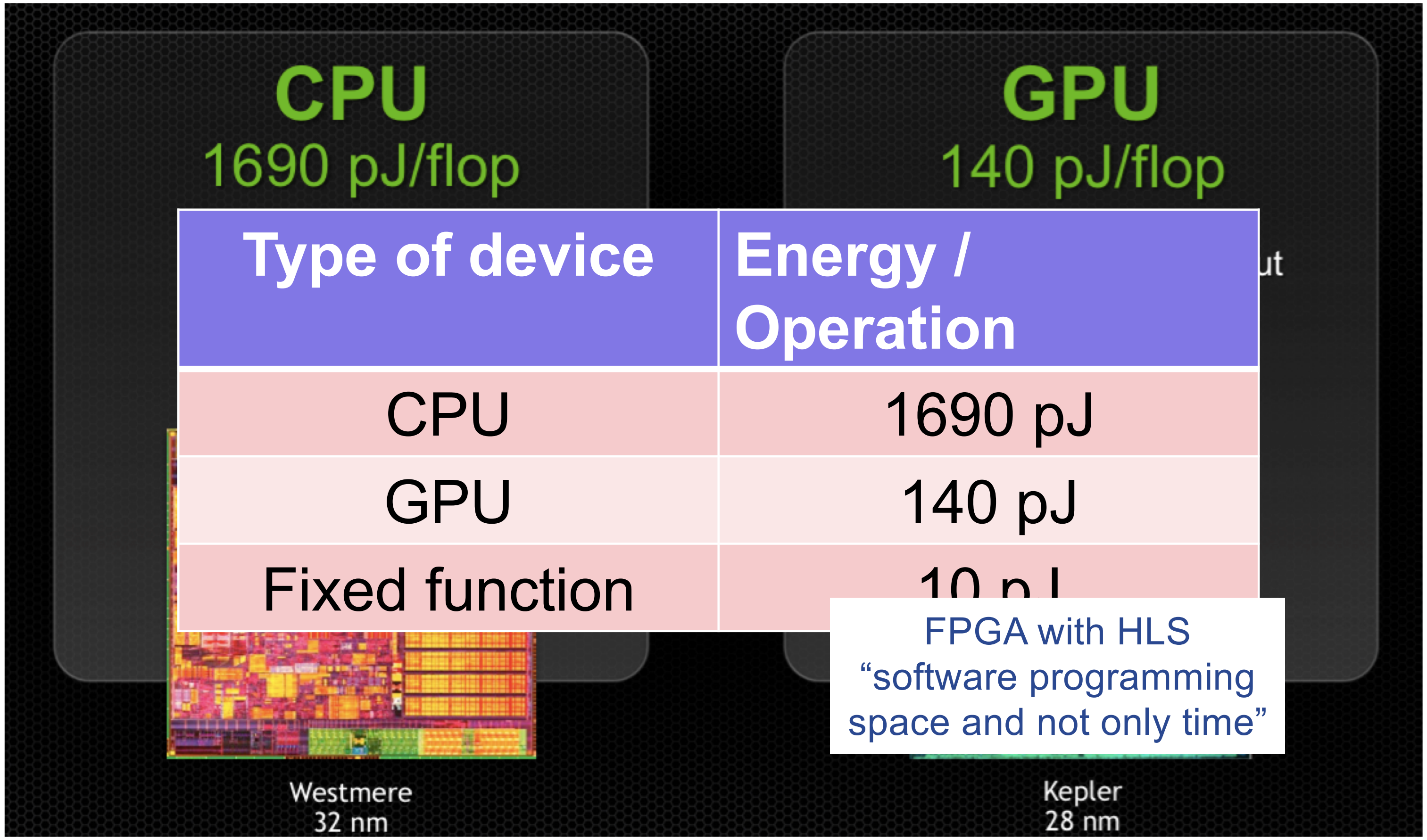
The problem:



Exponential power consumption

From "Total Consumer Power Consumption Forecast", Anders S.G. Andrae, October 2017

SPECIALIZATION LEADS TO MORE EFFICIENCY EFFICIENCY

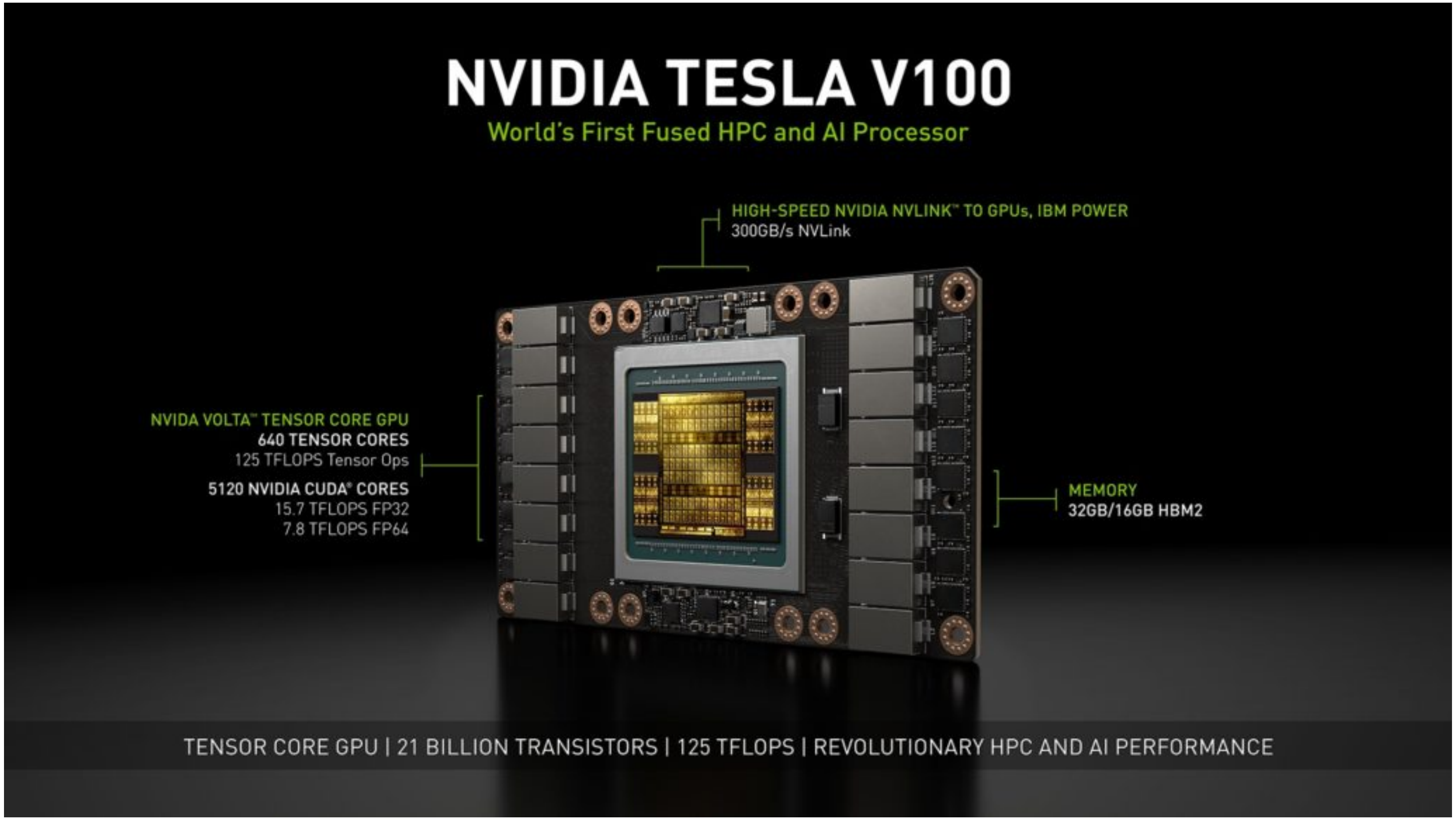


Type of device	Energy / Operation
CPU	1690 pJ
GPU	140 pJ
Fixed function	10 pJ

FPGA with HLS
“software programming
space and not only time”

Source from Bill Dally (nVidia) « Challenges for Future Computing Systems »
HiPEAC conference 2015

TODAY’S HPC ARE HETEREGENEOUS

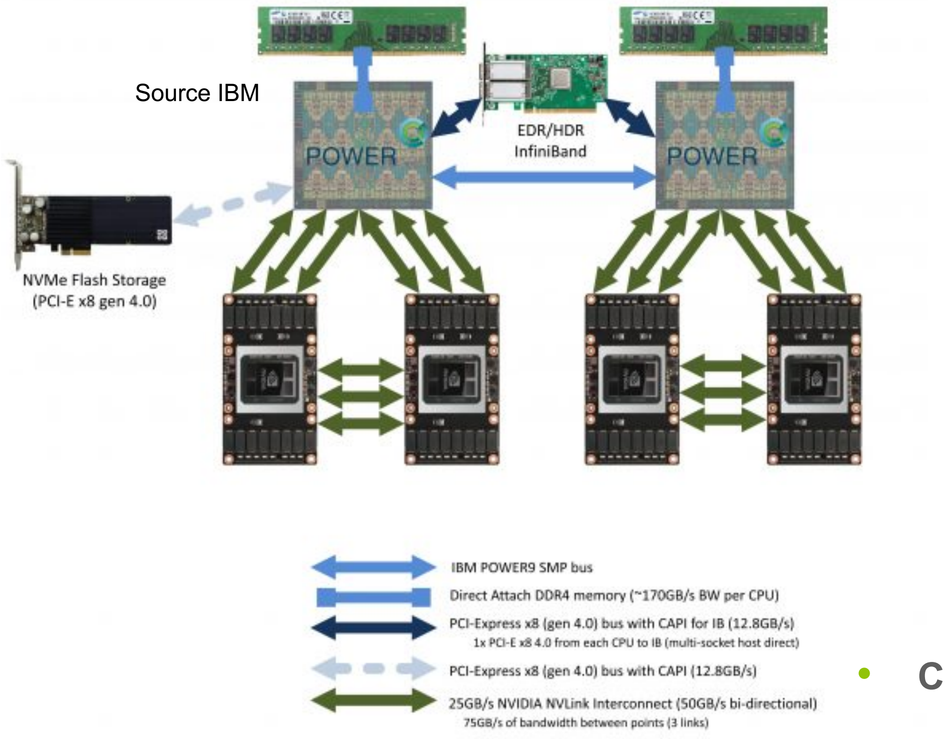

NVIDIA TESLA V100
World’s First Fused HPC and AI Processor
HIGH-SPEED NVIDIA NVLINK™ TO GPUs, IBM POWER
300GB/s NVLink
NVIDA VOLTA™ TENSOR CORE GPU
640 TENSOR CORES
125 TFLOPS Tensor Ops
5120 NVIDIA CUDA® CORES
15.7 TFLOPS FP32
7.8 TFLOPS FP64
MEMORY
32GB/16GB HBM2
TENSOR CORE GPU | 21 BILLION TRANSISTORS | 125 TFLOPS | REVOLUTIONARY HPC AND AI PERFORMANCE

- 3.3 peak exaops for emerging AI workloads
- 4,608 compute nodes, each containing two 22-core IBM Power9 processors and **six Nvidia Tesla V100 GPUs**
- Interconnected with dual-rail Mellanox EDR 100Gb/s InfiniBand.

TOP500 #1 & #2: NVIDIA TESLA V100 GPU + IBM POWER9 CPU

Server Block Diagram

Power Systems AC922 with NVIDIA Tesla V100 with Enhanced NVLink GPUs



Source IBM

- IBM POWER9 SMP bus
- Direct Attach DDR4 memory (~170GB/s BW per CPU)
- PCI-Express x8 (gen 4.0) bus with CAPI for IB (12.8GB/s)
  1x PCI-E x8 4.0 from each CPU to IB (multi-socket host direct)
- PCI-Express x8 (gen 4.0) bus with CAPI (12.8GB/s)
- 25GB/s NVIDIA NVLink Interconnect (50GB/s bi-directional)
  75GB/s of bandwidth between points (3 links)

NVIDIA TESLA V100 SPECIFICATIONS

		Tesla V100 for NVLink
PERFORMANCE with NVIDIA GPU Boost™	DOUBLE-PRECISION	7.8 teraFLOPS
	SINGLE-PRECISION	15.7 teraFLOPS
	DEEP LEARNING	125 teraFLOPS
INTERCONNECT BANDWIDTH Bi-Directional	NVLINK	300 GB/s
MEMORY CoWoS Stacked HBM2	CAPACITY	32/16 GB HBM2
	BANDWIDTH	900 GB/s

Source NVIDIA

- Compute performance from GPU

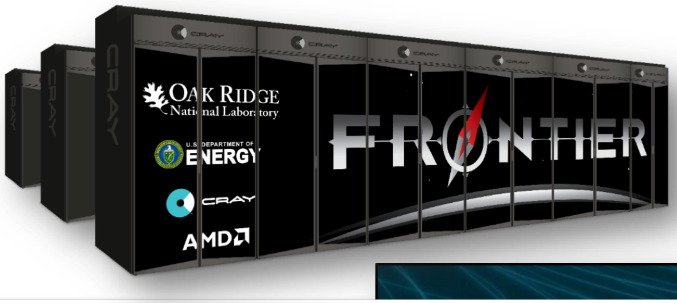
➡ Heterogeneous integration driven by compute energy efficiency

From Denis Dutoit

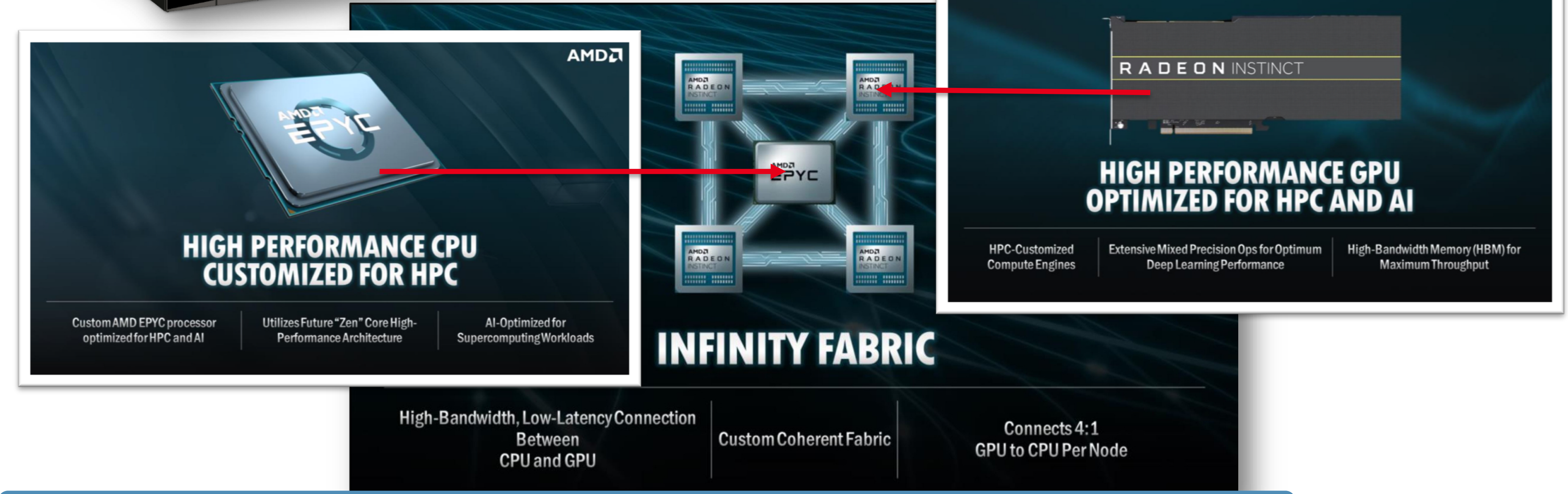
May, 2019

AMD'S EPYC AND RADEON TO POWER EXASCALE SUPERCOMPUTER

https://www.amd.com/es/products/frontier



- Performance target: 1.5 exaflops; 40 MW; 37 GFLOPS/W
- Compute node: x1 CPU + x4 GPU + coherent fabric



HIGH PERFORMANCE CPU CUSTOMIZED FOR HPC

Custom AMD EPYC processor optimized for HPC and AI | Utilizes Future "Zen" Core High-Performance Architecture | AI-Optimized for Supercomputing Workloads

INFINITY FABRIC

High-Bandwidth, Low-Latency Connection Between CPU and GPU | Custom Coherent Fabric | Connects 4:1 GPU to CPU Per Node

HIGH PERFORMANCE GPU OPTIMIZED FOR HPC AND AI

HPC-Customized Compute Engines | Extensive Mixed Precision Ops for Optimum Deep Learning Performance | High-Bandwidth Memory (HBM) for Maximum Throughput

➡ Heterogeneous integration requires high-bandwidth, low-latency connection

From Denis Dutoit

52ND EDITION OF THE TOP500 LIST (NOVEMBER 11TH, 2018)

Heterogeneous integration ?

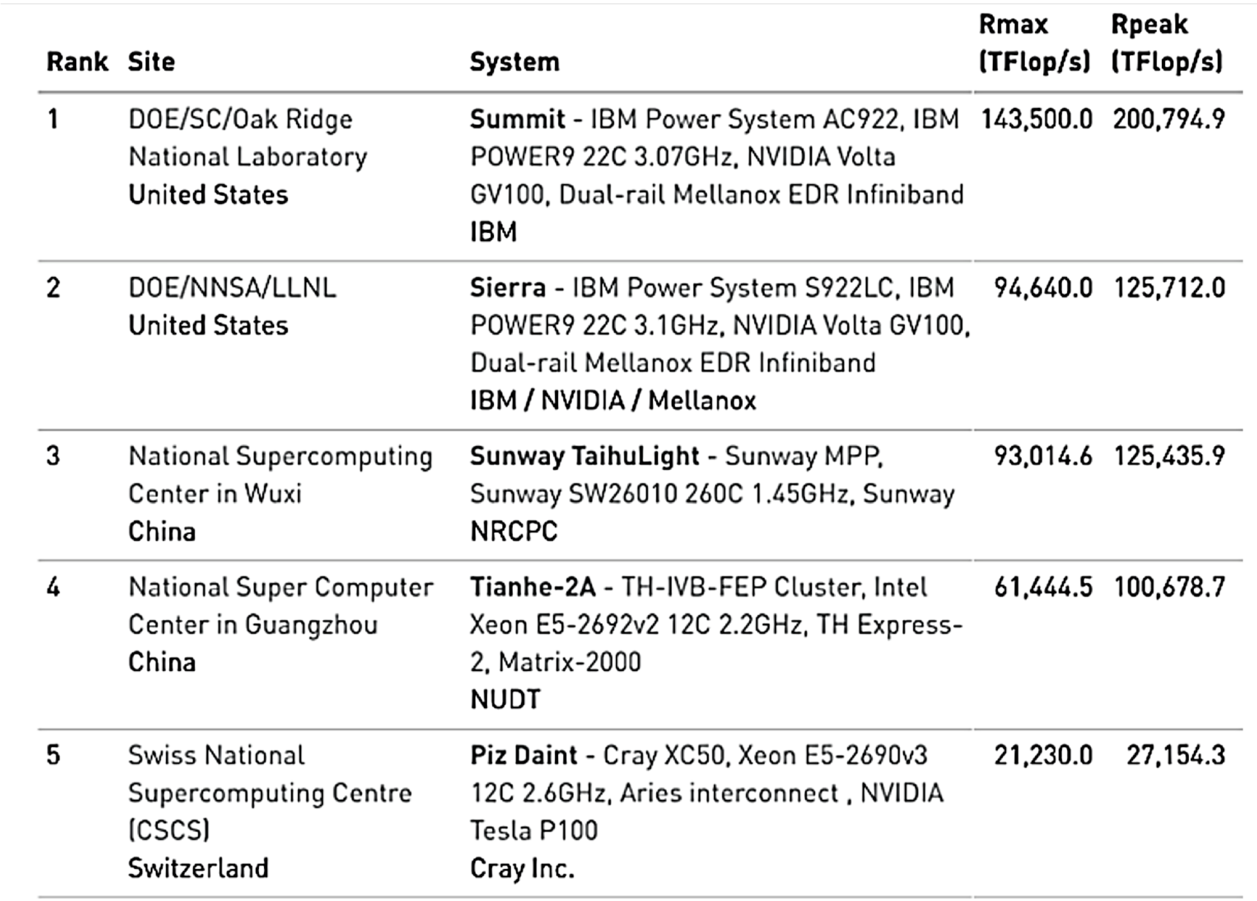
Yes

Yes

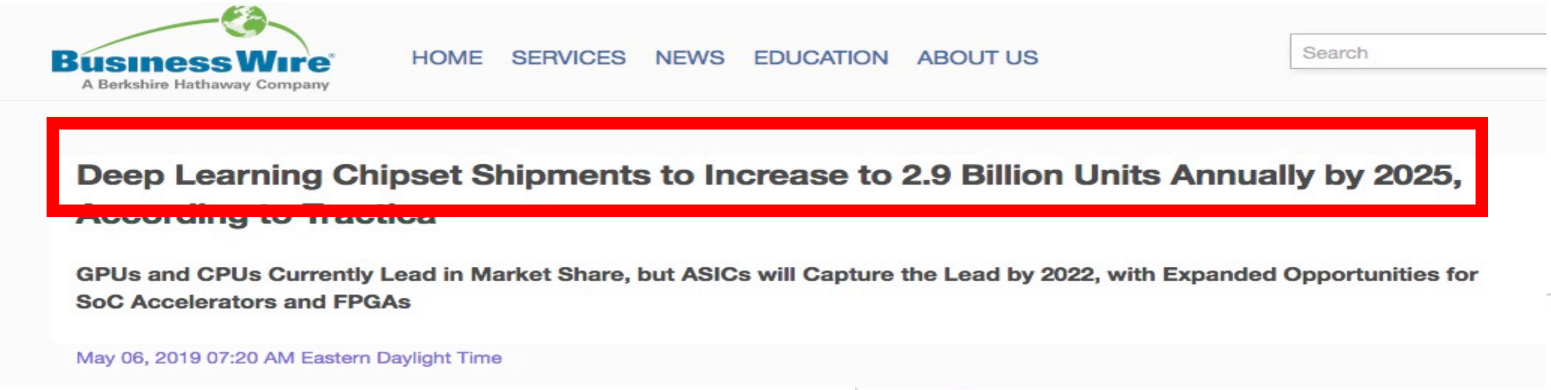
No

Yes

Yes



Rank	Site	System	Rmax (TFlop/s)	Rpeak (TFlop/s)
1	DOE/SC/Oak Ridge National Laboratory United States	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM	143,500.0	200,794.9
2	DOE/NNSA/LLNL United States	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM / NVIDIA / Mellanox	94,640.0	125,712.0
3	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	93,014.6	125,435.9
4	National Super Computer Center in Guangzhou China	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 NUDT	61,444.5	100,678.7
5	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Cray Inc.	21,230.0	27,154.3

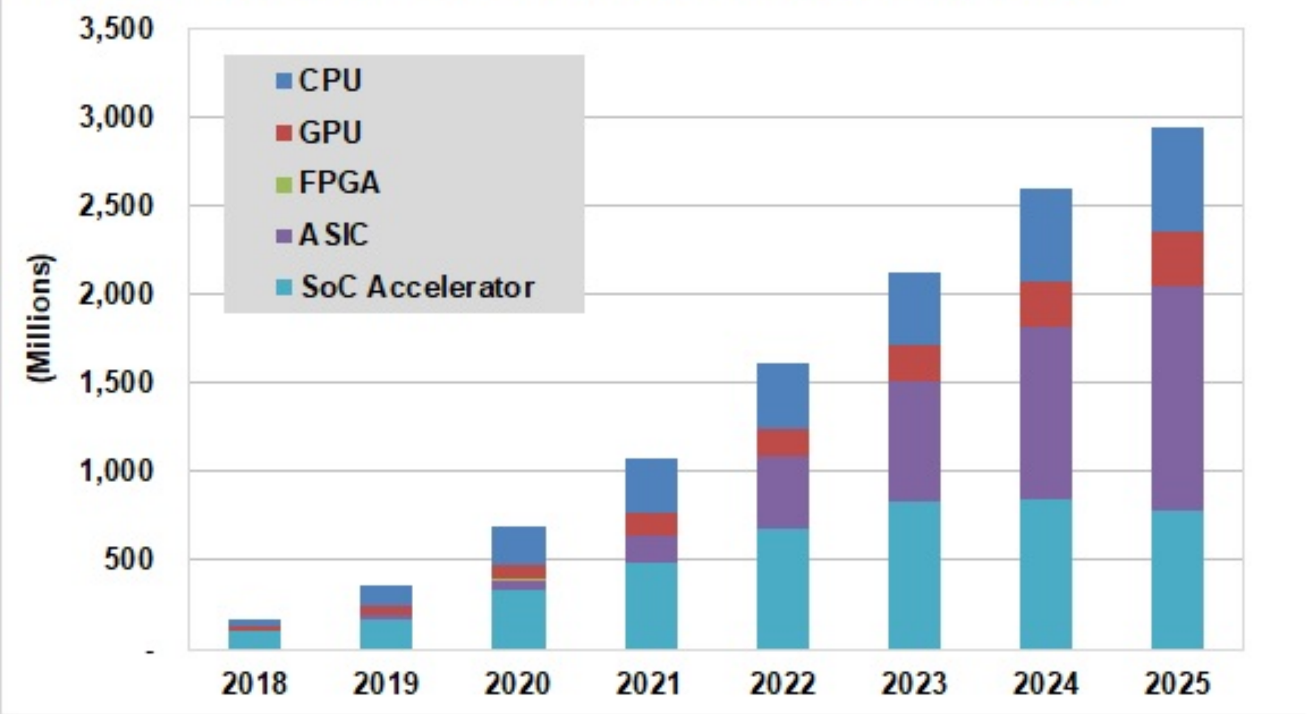


Deep Learning Chipset Shipments to Increase to 2.9 Billion Units Annually by 2025, According to Tractica

GPUs and CPUs Currently Lead in Market Share, but ASICs will Capture the Lead by 2022, with Expanded Opportunities for SoC Accelerators and FPGAs

May 06, 2019 07:20 AM Eastern Daylight Time

Deep Learning Chipset Unit Shipments by Type, World Markets: 2018-2025

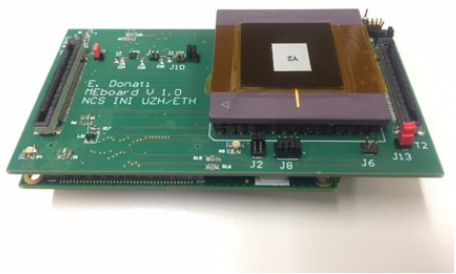


Source: Tractica

GOING NEURO-INSPIRED: “SPIKING” NEURAL NETWORKS

Using another way of coding information…not using bits

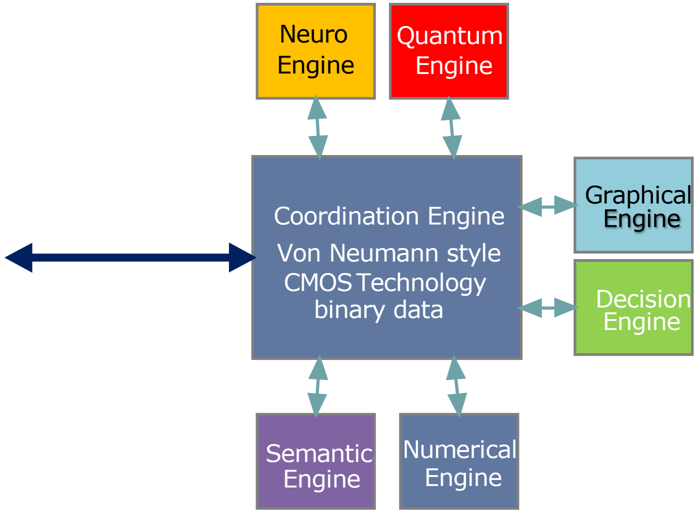
	IBM TrueNorth	Intel Loihi	DynapSEL
Technology	28nm CMOS	14 nm CMOS	28 nm FDSOI
Supply Voltage	0.7-1.05 V	0.5-1.25 V	0.73-1 V
Design Type	Digital	Digital	Mixed-signal
Neurons per core	256	Max 1k	256
Core Area	0.094 mm^2	0.4 mm^2	0.36 mm^2
Computation	Time multiplexing	Time multiplexing	**Parallel processing**
Fan In/Out	256/256	16/4k	**2k/8k**
On-line Learning	No	Programmable	**STDP**
Synaptic Operation / Second / Watt	46 GSOPS/W		**300 GSOPS/W**
Energy per synaptic operation	26 pJ	23.6 pJ	**<2 pJ**



FUSING PARADIGMS AT HARDWARE LEVEL

At the hardware level, the good old Von Neumann/ CMOS partnership can act as a computing substrate, or orchestrator of various accelerators/technologies

- Acting as coordination / communication node
- Allowing Hardware / Software integration



Neuro Engine

Quantum Engine

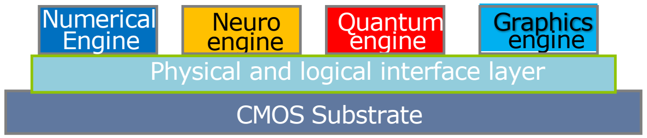
Coordination Engine
Von Neumann style
CMOS Technology
binary data

Graphical Engine

Decision Engine

Semantic Engine

Numerical Engine



Numerical Engine

Neuro engine

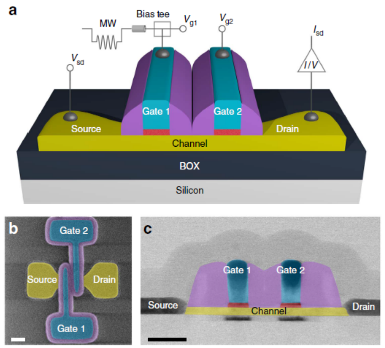
Quantum engine

Graphics engine

Physical and logical interface layer

CMOS Substrate

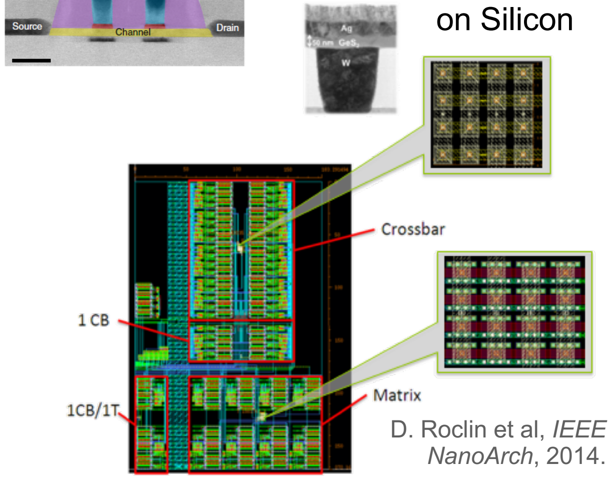
Slide from Christian Gamrat



Qubits on Silicon

Maurand et al,
Nature Com.,
Jul. 2016.

NVM Synapses
on Silicon



D. Roclin et al, *IEEE NanoArch*, 2014.

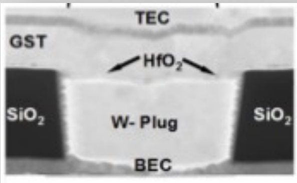
NON VOLATILE MEMORIES

PCM

GST

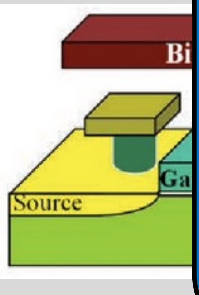
GeTe

GST + HfO_2



Thermal effect

MR



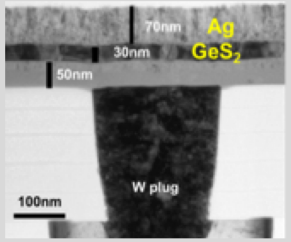
- Can change the structure of memory hierarchy?

+ 64/128 addressing scheme

⇒ Do we still need files?

⇒ Direct access of objects

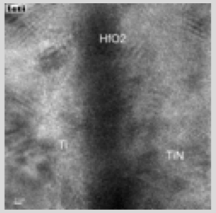
Ag / GeS_2



Electrochemical effect

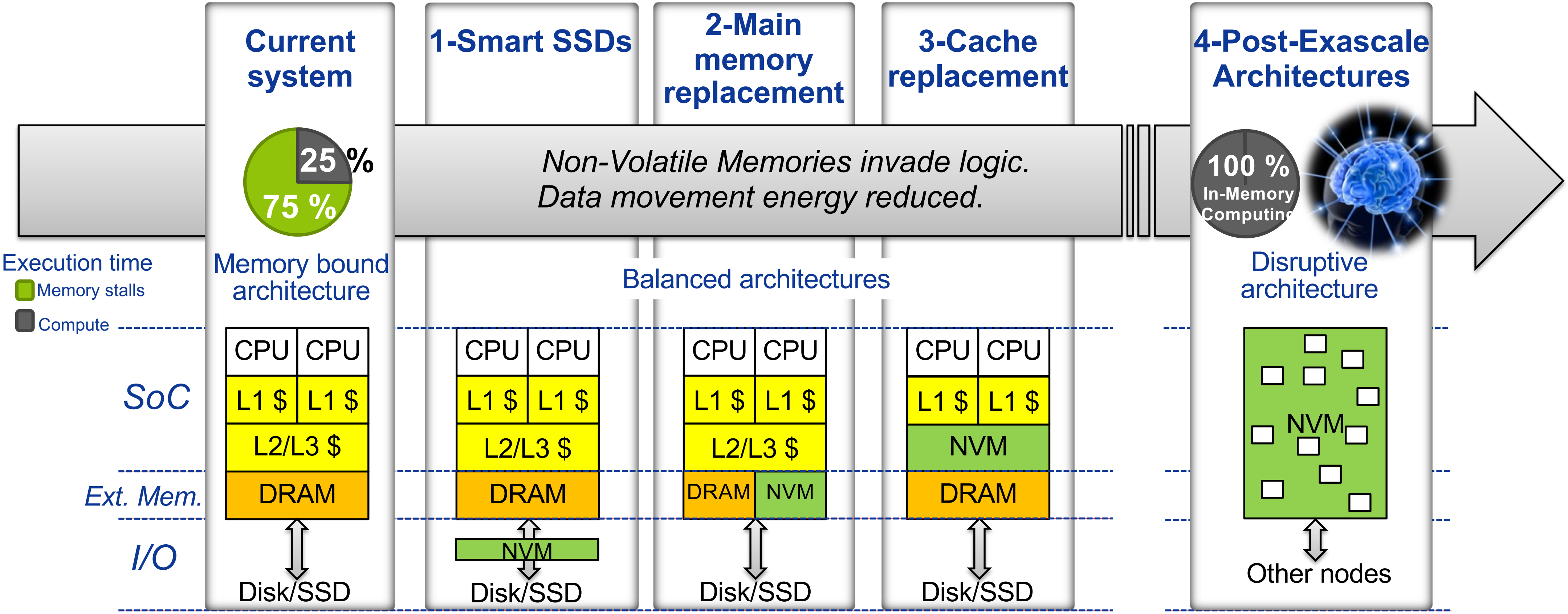
OXRAM

$TiN/HfO_2/Ti/TiN$



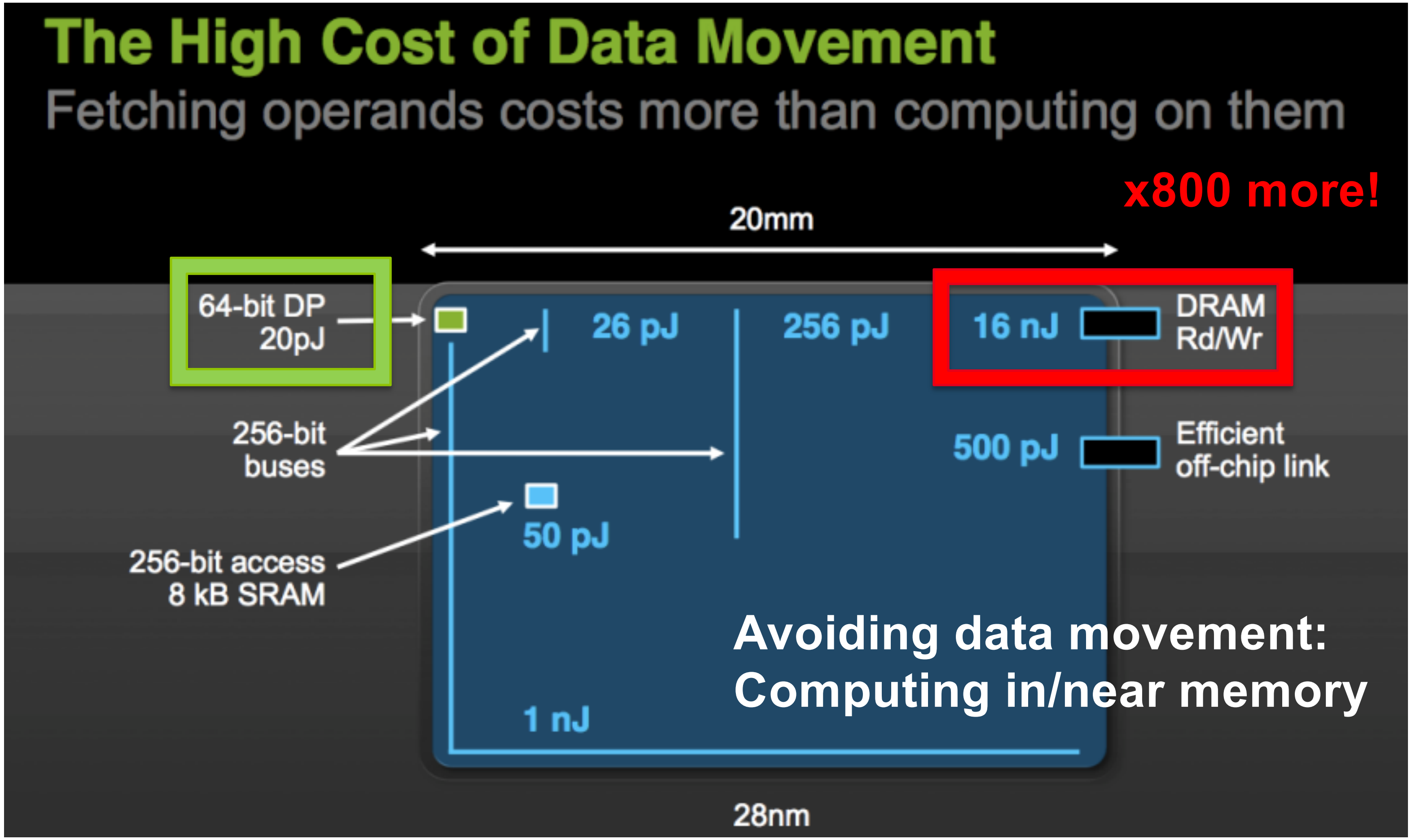
Electronic effect oxygen vacancies

NEW ARCHITECTURE PARADIGMS WITH NVM



From Denis Dutoit

SOLVING THE ENERGY CHALLENGE: COST OF MOVING DATA



The High Cost of Data Movement

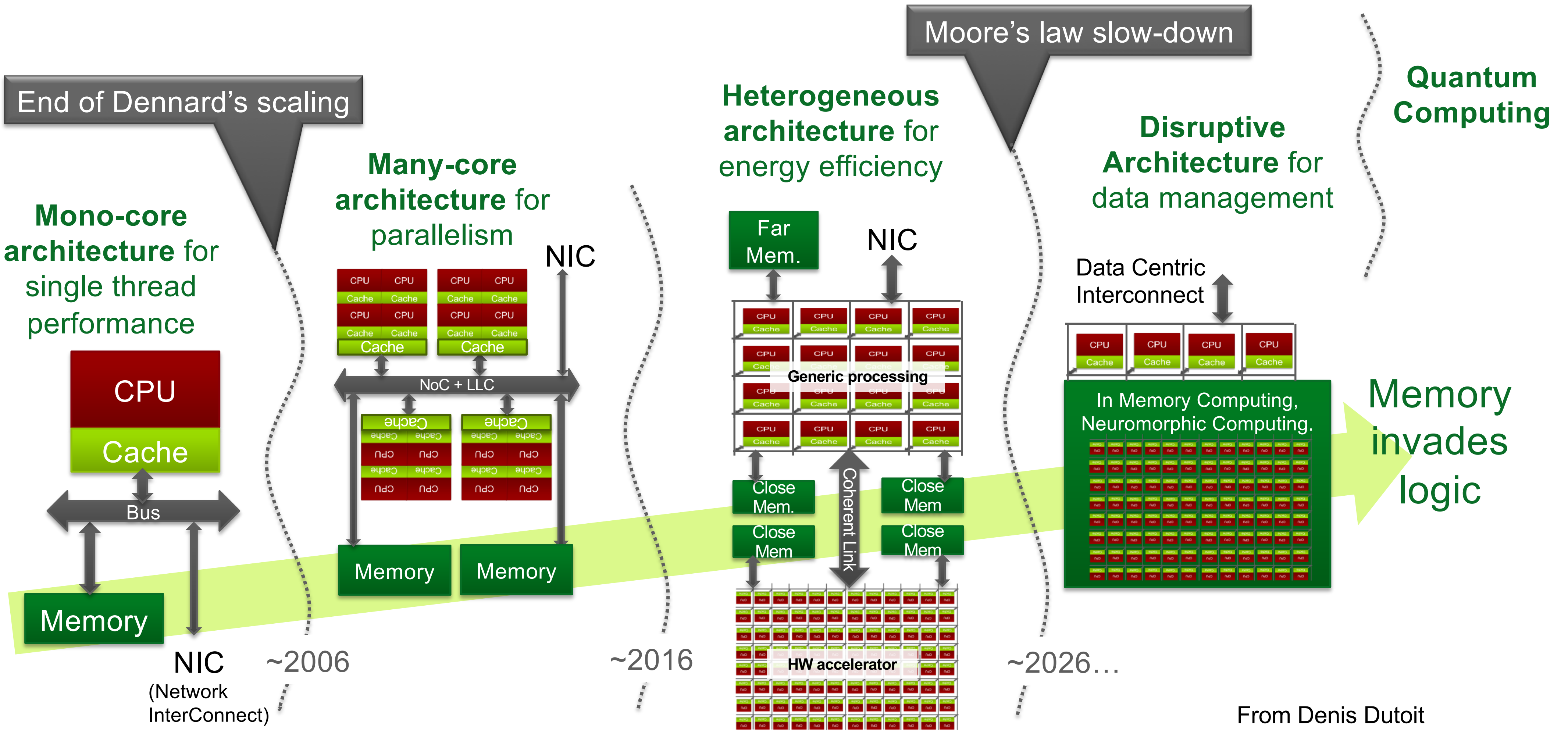
Fetching operands costs more than computing on them

x800 more!

Avoiding data movement:
Computing in/near memory

Source: Bill Dally, « To ExaScale and Beyond »
www.nvidia.com/content/PDF/sc_2010/theater/Dally_SC10.pdf

PROCESSOR ARCHITECTURE EVOLUTION



End of Dennard's scaling

Moore's law slow-down

Mono-core architecture for single thread performance

CPU

Cache

Bus

Memory

NIC
(Network
InterConnect)

~2006

Many-core architecture for parallelism

CPU CPU
Cache Cache
CPU CPU
Cache Cache
Cache

NIC

NoC + LLC

Memory

Memory

~2016

Heterogeneous architecture for energy efficiency

Far Mem.

NIC

Generic processing

Close Mem.

Close Mem

Coherent Link

HW accelerator

~2026…

Disruptive Architecture for data management

Data Centric Interconnect

In Memory Computing, Neuromorphic Computing.

Memory invades logic

Quantum Computing

From Denis Dutoit

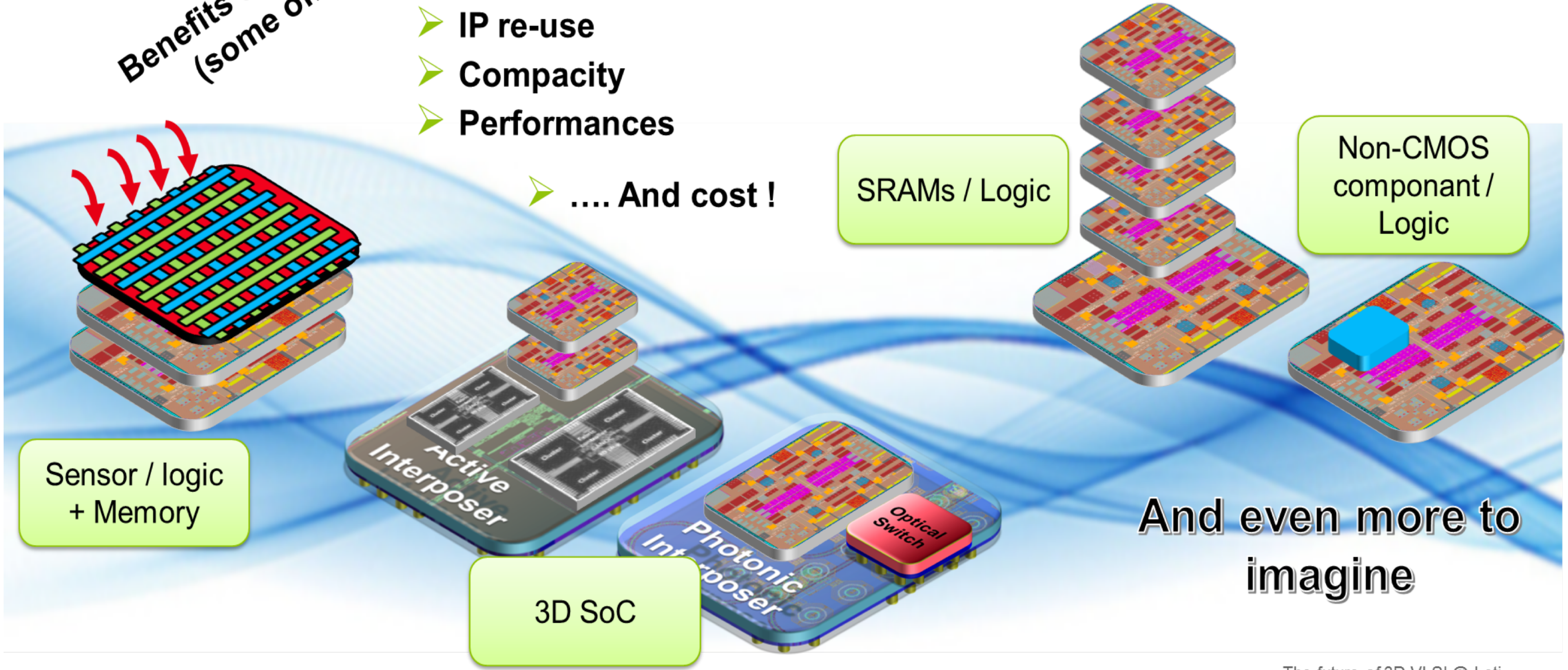
HETEROGENEOUS INTEGRATION

2.5 stacking with chiplets and interposers for heterogeneous integration

Benefits of 3DVLSI
(some of..)

- Scalability
- Dedicated die function
- IP re-use
- Compactity
- Performances

- …. And cost !



Sensor / logic
+ Memory

Active Interposer

3D SoC

Photonic Interposer

Optical Switch

SRAMs / Logic

Non-CMOS componant / Logic

And even more to imagine

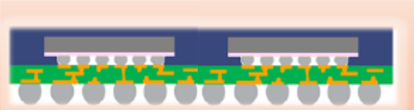
FROM ADVANCED PACKAGING TECHNOLOGIES ….

… TO CHIPLETS

Advanced Integration

SiP

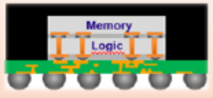
Multi-Chip-Module



Interconnect density: 100µm x 100µm

3D

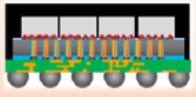
Die stacking



Interconnect density: 10µm x 10µm

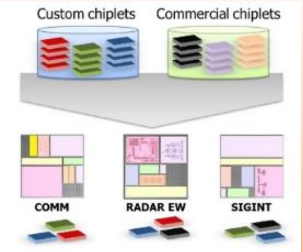
2.5D

Interposer based

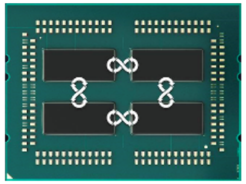


Interconnect density: 10µm x 10µm

CHIPLET partitioning

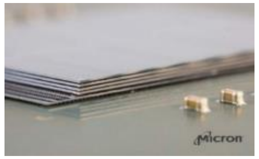


Source: DARPA



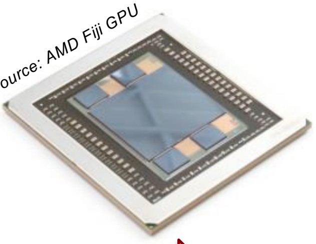
Source: AMD EPYC 7260, 4-chiplet chip

System-in-Package

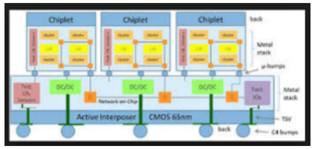


Source: Micron High-Bandwidth-Memory

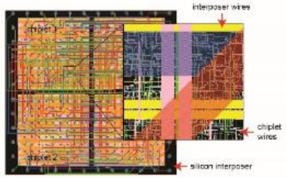
Source: AMD Fiji GPU



3D Integrated-Circuit (3D IC)



Source: LETI

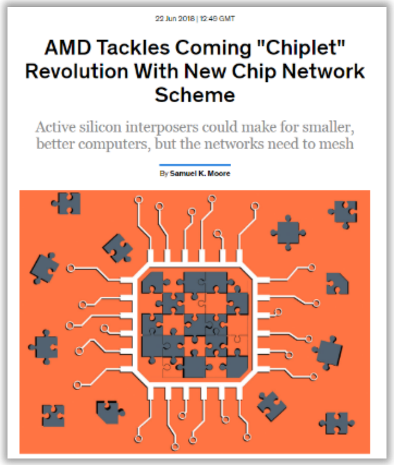


Source: GeorgiaTech

From Denis Dutoit

SOME RECENT ANNOUCEMENTS
… ON CHIPLETS & ACTIVE INTERPOSERS

From AMD ….



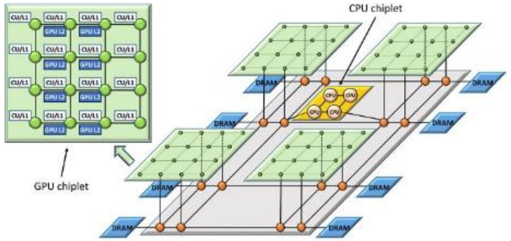
22 Jun 2018 | 12:48 GMT

AMD Tackles Coming "Chiplet" Revolution With New Chip Network Scheme

Active silicon interposers could make for smaller, better computers, but the networks need to mesh

By Samuel K. Moore

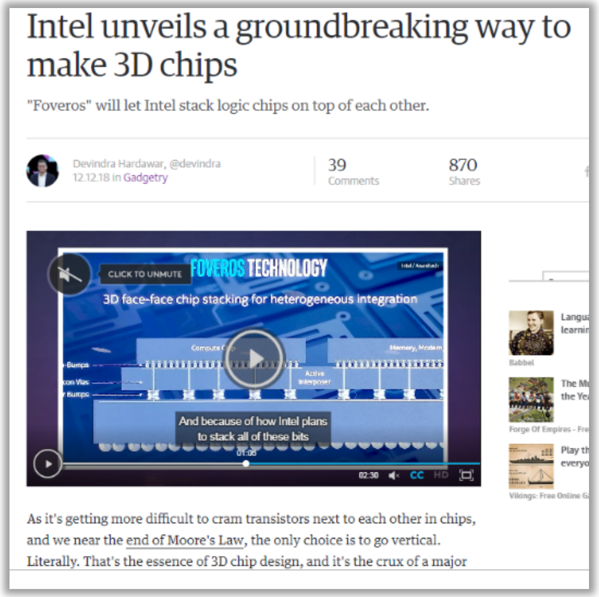
https://spectrum.ieee.org/tech-talk/semiconductors/design/amd-tackles-coming-chiplet-revolution-with-new-chip-network-scheme



[J. Yin et al., "Modular Routing Design for Chiplet-based Systems", ISCA'2018]

… and INTEL

FOVEROS Technology



Intel unveils a groundbreaking way to make 3D chips

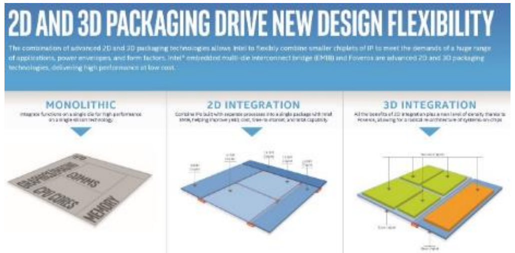
"Foveros" will let Intel stack logic chips on top of each other.

Devindra Hardawar, @devindra
12.12.18 in Gadgetry

39 Comments

870 Shares

As it's getting more difficult to cram transistors next to each other in chips, and we near the end of Moore's Law, the only choice is to go vertical. Literally. That's the essence of 3D chip design, and it's the crux of a major



2D AND 3D PACKAGING DRIVE NEW DESIGN FLEXIBILITY

The combination of advanced 2D and 3D packaging technologies allows Intel to flexibly combine smaller chiplets of IP to meet the demands of a huge range of applications, power envelopes, and form factors. Intel® embedded multi-die interconnect bridge (EMIB) and Foveros are advanced 2D and 3D packaging technologies, delivering high performance at low cost.

MONOLITHIC
Integrate functions on a single die for high performance on a single silicon technology

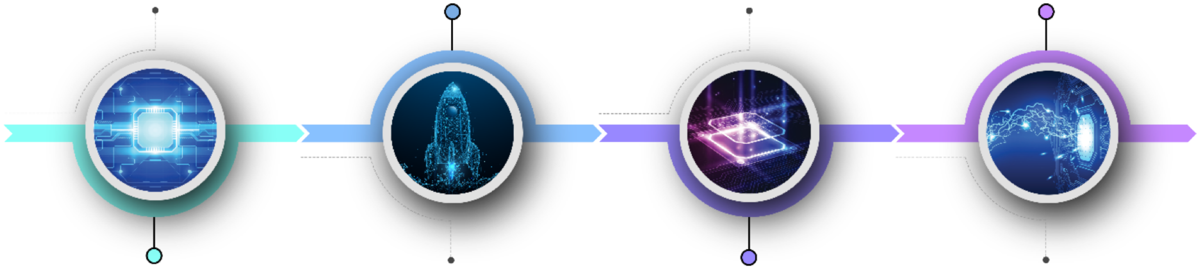
2D INTEGRATION
Combine IPs built with separate processes into a single package with Intel EMIB, helping improve yield, cost, time-to-market, and total capability

3D INTEGRATION
All the benefits of 2D integration plus a new level of density thanks to Foveros, allowing for a radical re-architecture of systems-on-chip

https://www.engadget.com/2018/12/12/intel-foverus-3d-chip/?yptr=yahoo&guccounter=2

From Denis Dutoit

EUROPEAN PROCESSOR INITIATIVE



2021

Rhea Family - Gen1 GPP

EPI Common Platform
ARM & RISC-V
External IPs
•
HPC System PreExascale
Automotive PoC

EPI IP's Launch Pad
&
Pan European Research
Platform for HPC and AI

2021-2022

2022-2023

Cronos Family - Gen2 GPP

EPI Common Platform
ARM & RISC-V
•
HPC System Exascale
Automotive CPU

Gen3 GPP Family

2024-...

PROJECT PILLARS

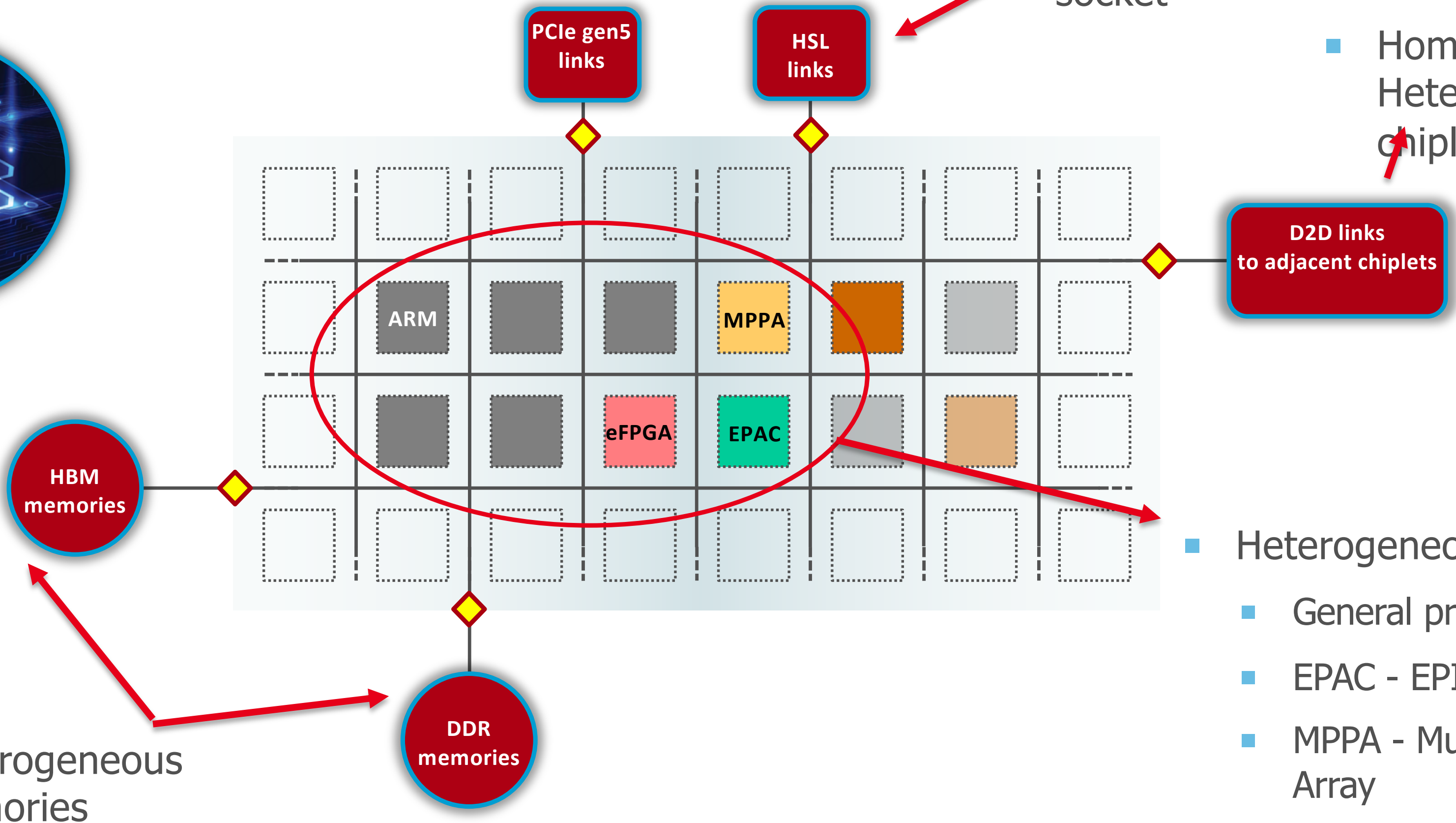
- Common platform and global architecture stream
- HPC general purpose processor stream
- Accelerator stream
- Automotive platform stream

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826647

www.european-processor-initiative.eu

BMW GROUP, MINI, Rolls-Royce Motor Cars Limited
Atos, infineon, JÜLICH Forschungszentrum
Barcelona Supercomputing Center – Centro Nacional de Supercomputación, FORTH Institute of Computer Science
KALRAY, semidynamics
Alma Mater Studiorum Università di Bologna, CHALMERS, FER, Università di Pisa
Fraunhofer ITWM, EXTOLL latency matters.
E4 Computer Engineering, CINECA, cea, ST
SURF SARA, TÉCNICO LISBOA, Elektrobit
ETH zürich, GENCI
Prove & Run, KIT Karlsruher Institut für Technologie, menta

COMMON PLATFORM FOR MULTI-LEVEL HETEROGENEOUS INTEGRATION



- Heterogeneous socket

- Homogenegous/ Heterogeneous chiplet

- Heterogeneous SoC:
  - General processing core
  - EPAC - EPI Accelerator
  - MPPA - Multi-Purpose Processing Array
  - eFPGA - embedded FPGA

- Heterogeneous memories

From Denis Dutoit

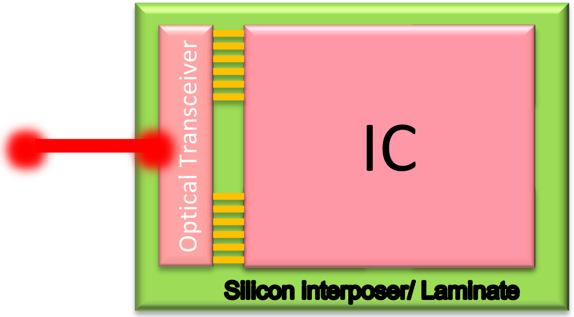
ELECTRONS VERSUS PHOTONS

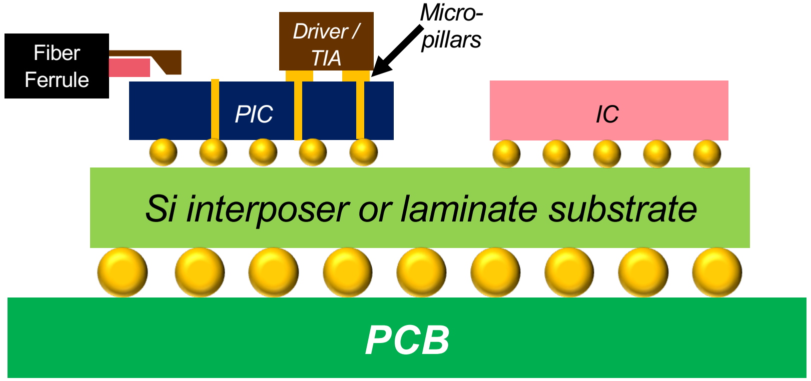
Electrons: Easy to create and interface
Attenuation with the distance (Ohm’s law)

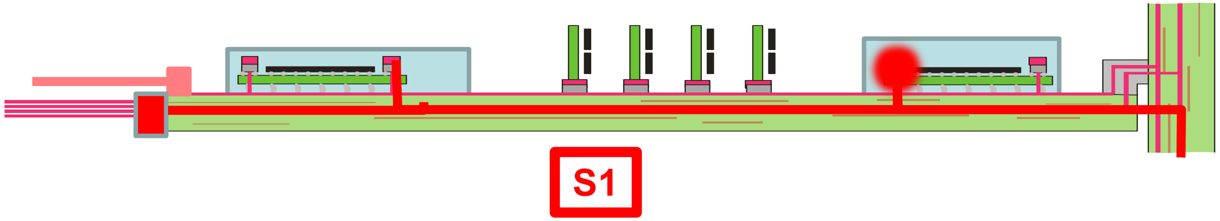
Photons: Energy demanding for creation and interfacing
Low attenuation with the distance

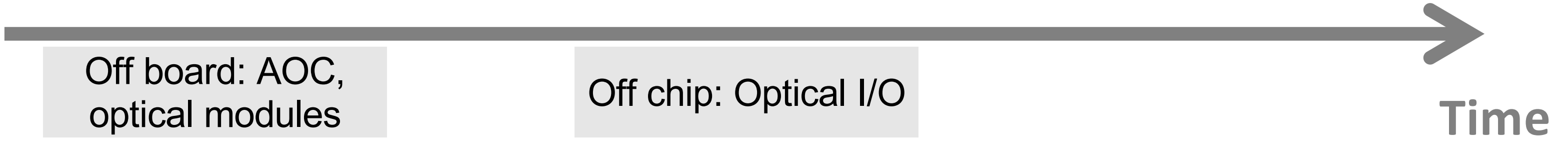
OFF-CHIP PHOTONICS

Photonics: cost in sending information, nearly ***nothing in transmission***









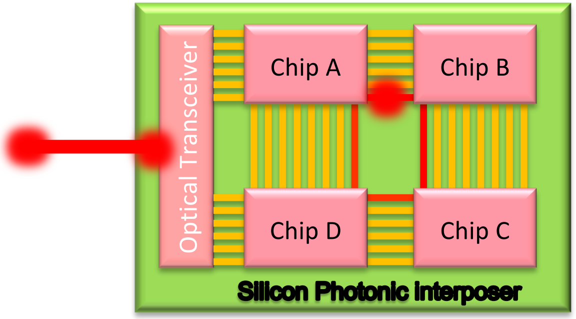
Off board: AOC, optical modules

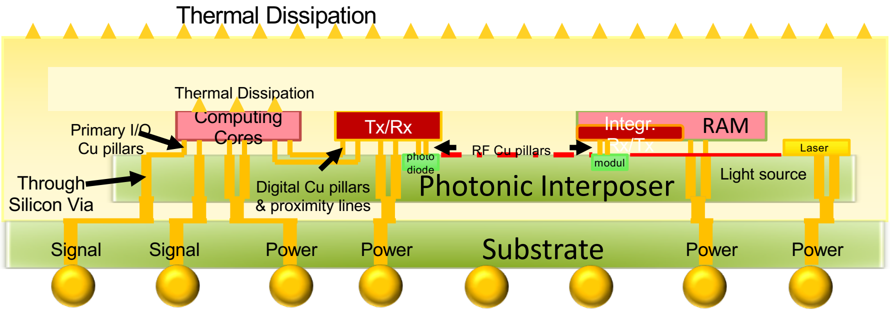
Off chip: Optical I/O

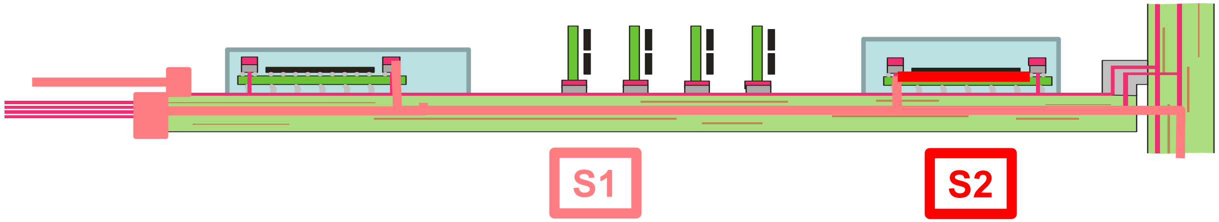
Now

Step 1

IN-PACKAGE PHOTONICS







Off board: AOC, optical modules

Now

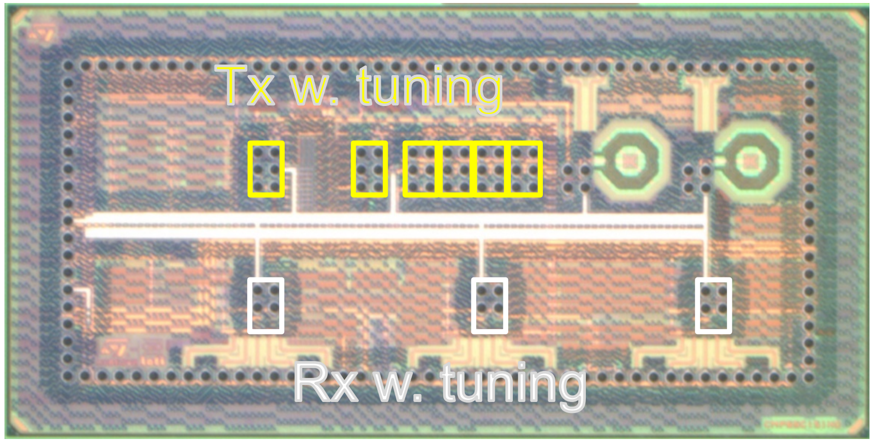
Off chip: Optical I/O

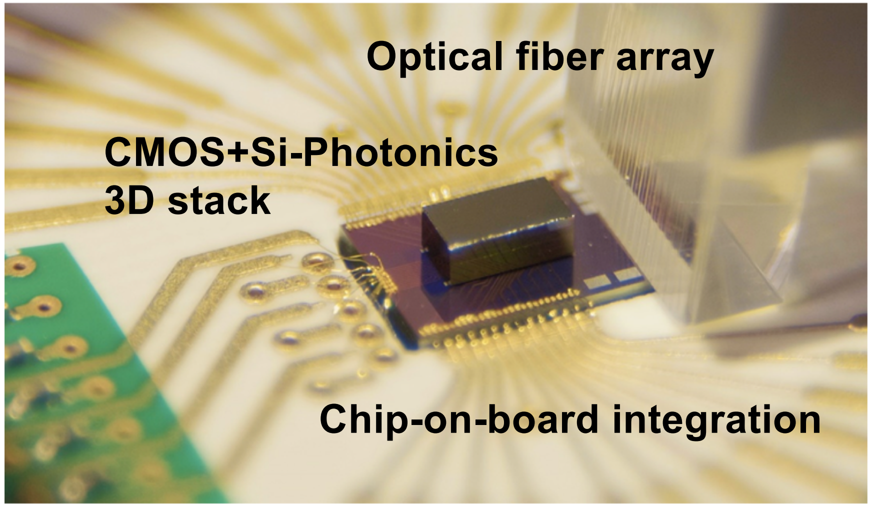
Step 1

Optical network in package

Step 2

DEMONSTRATION OF A THERMALLY TUNED WDM ELECTRO-OPTICAL LINK

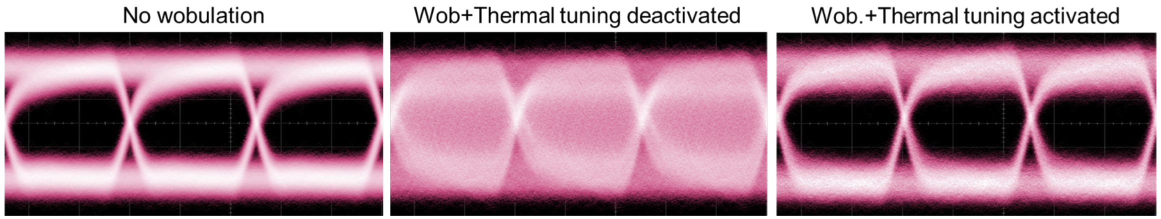




Y. Thonnart & al. ISSCC'2018

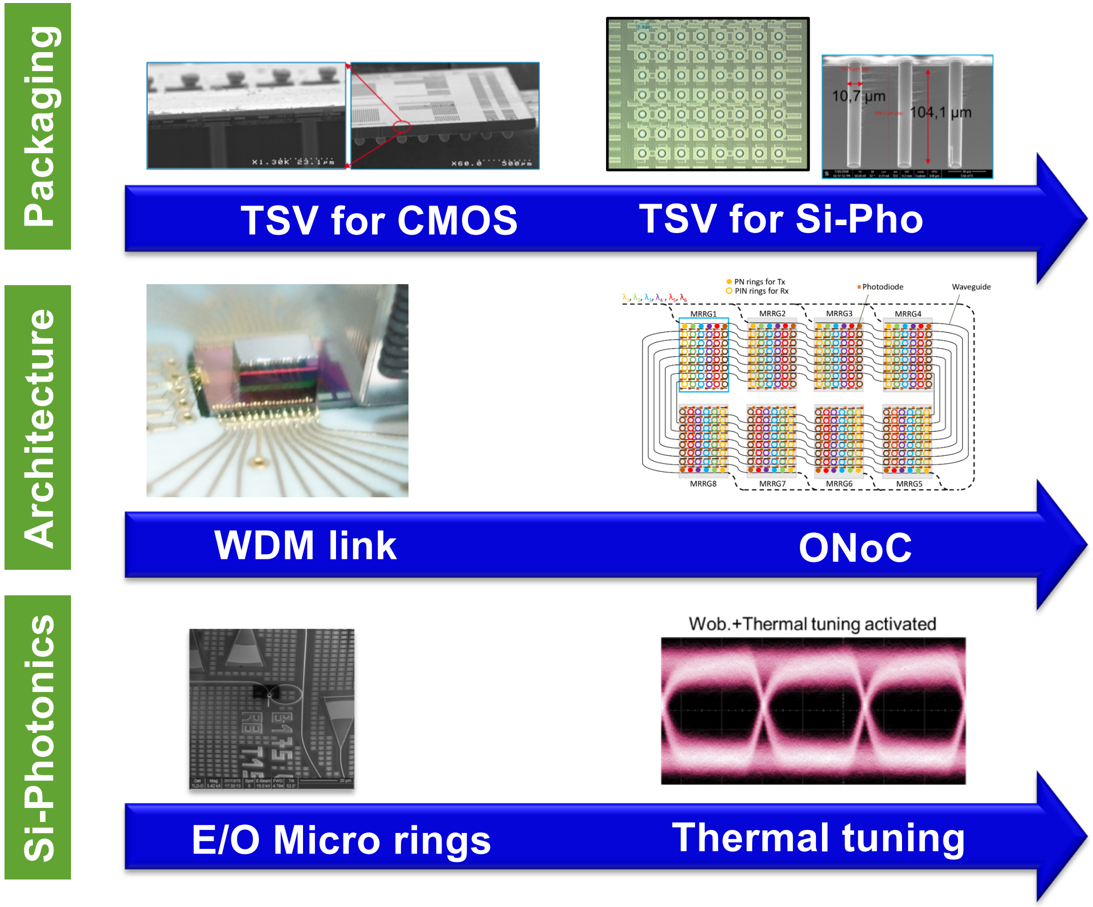
- **1Tbps/mm²** bandwidth density
- Tight technology integration of E/O ring modulators within a 3D stack
- Integrated thermal tuning robust to compute fabric heating

[LETI: Y. Thonnart, ISSCC2018]



From Denis Dutoit

LETI'S SI-PHOTONICS ROADMAP FOR POST-EXASCALE COMPUTING



Packaging

TSV for CMOS

TSV for Si-Pho

Architecture

WDM link

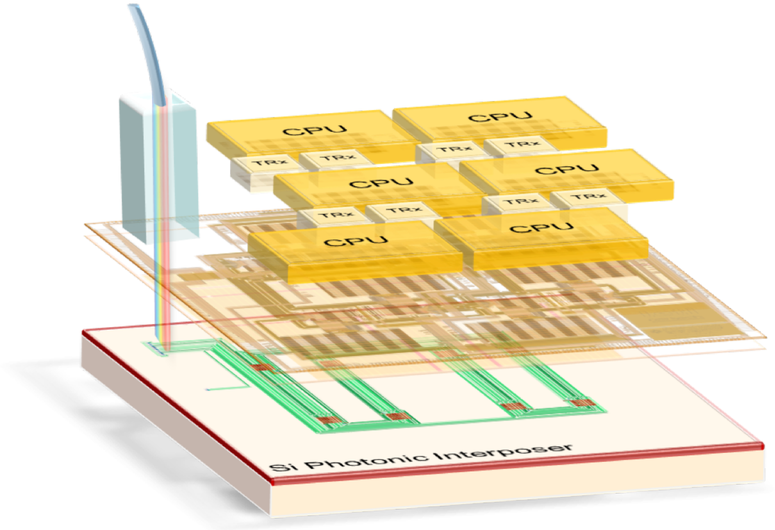
ONoC

Si-Photonics

E/O Micro rings

Wob.+Thermal tuning activated

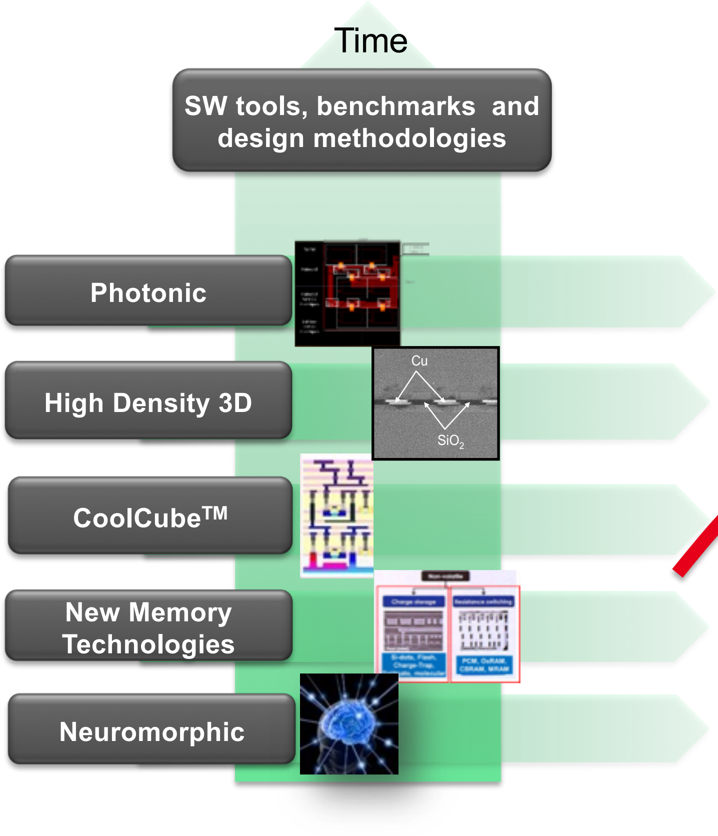
Thermal tuning



- Target demonstrator ➔ 2021
- 96-core cache-coherent processor
- Generic E/O chiplets
- 8-node optical NoC
  - 576 Gbit/s aggregated bandwidth
  - 384 microring resonators
  - ~10 ns electro-optical latency

From Denis Dutoit

POTENTIAL SOLUTION FOR POST EXASCALE BOARD



Time

SW tools, benchmarks and design methodologies

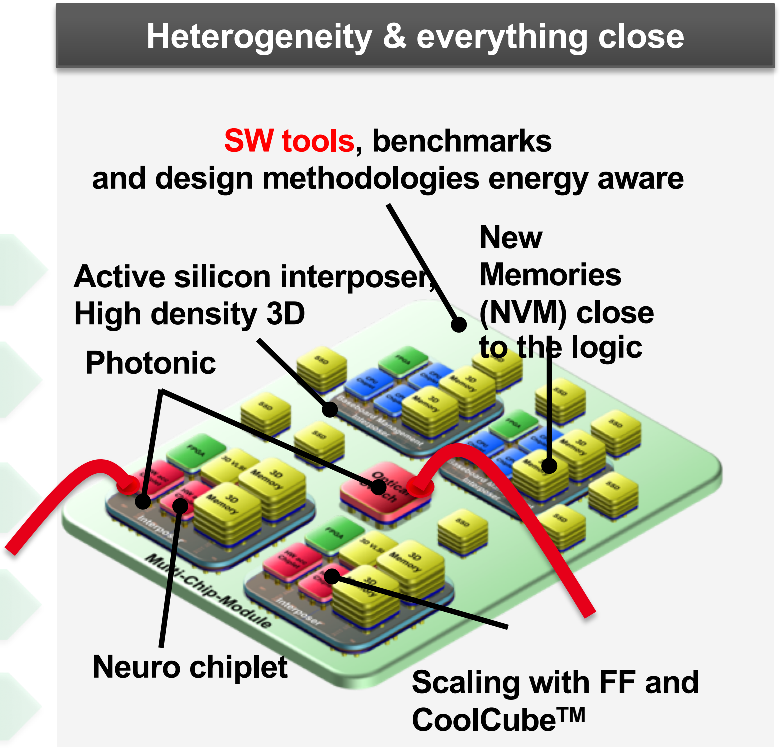
Photonic

High Density 3D

CoolCube™

New Memory Technologies

Neuromorphic



Heterogeneity & everything close

SW tools, benchmarks and design methodologies energy aware

Active silicon interposer, High density 3D

New Memories (NVM) close to the logic

Photonic

Neuro chiplet

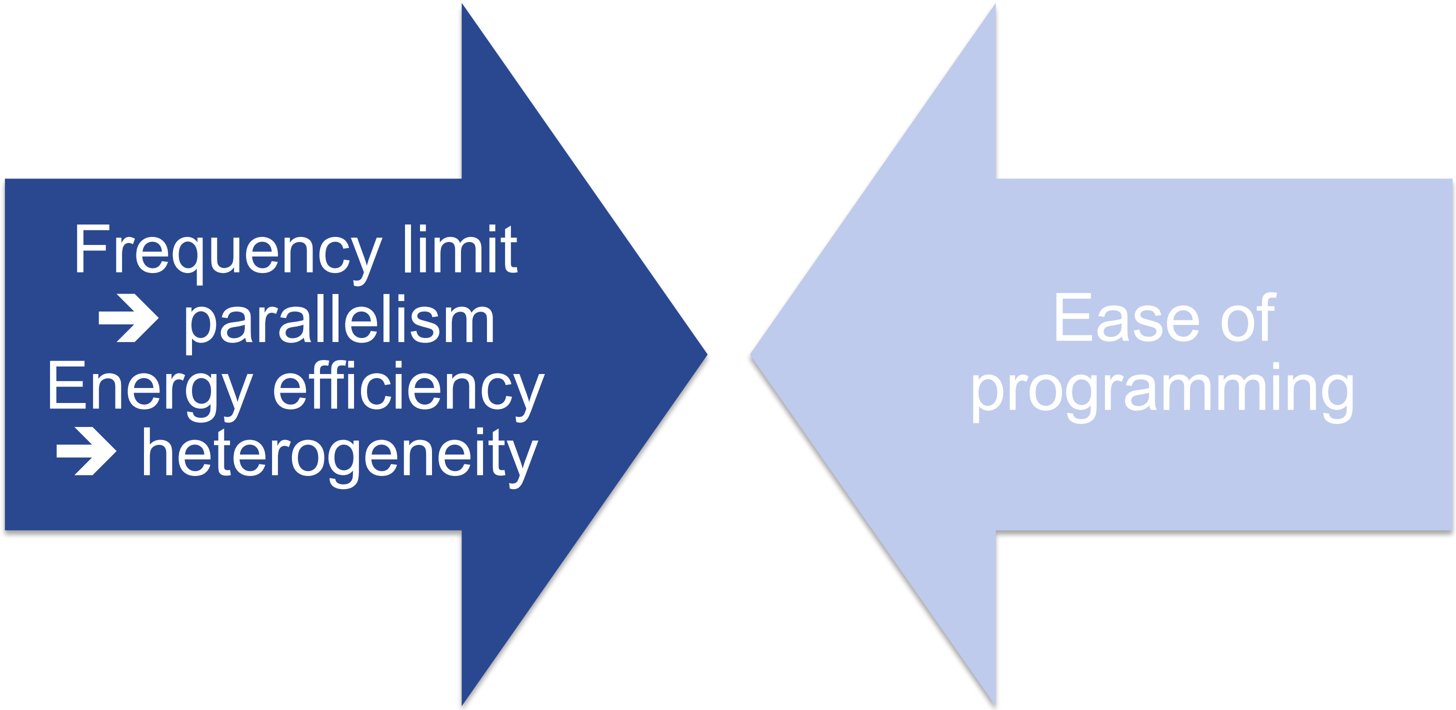
Scaling with FF and CoolCube™

From Denis Dutoit

Outline

1) Evolution of application scope: the continuum
2) Hardware heterogeneity and orchestration
3) Software?

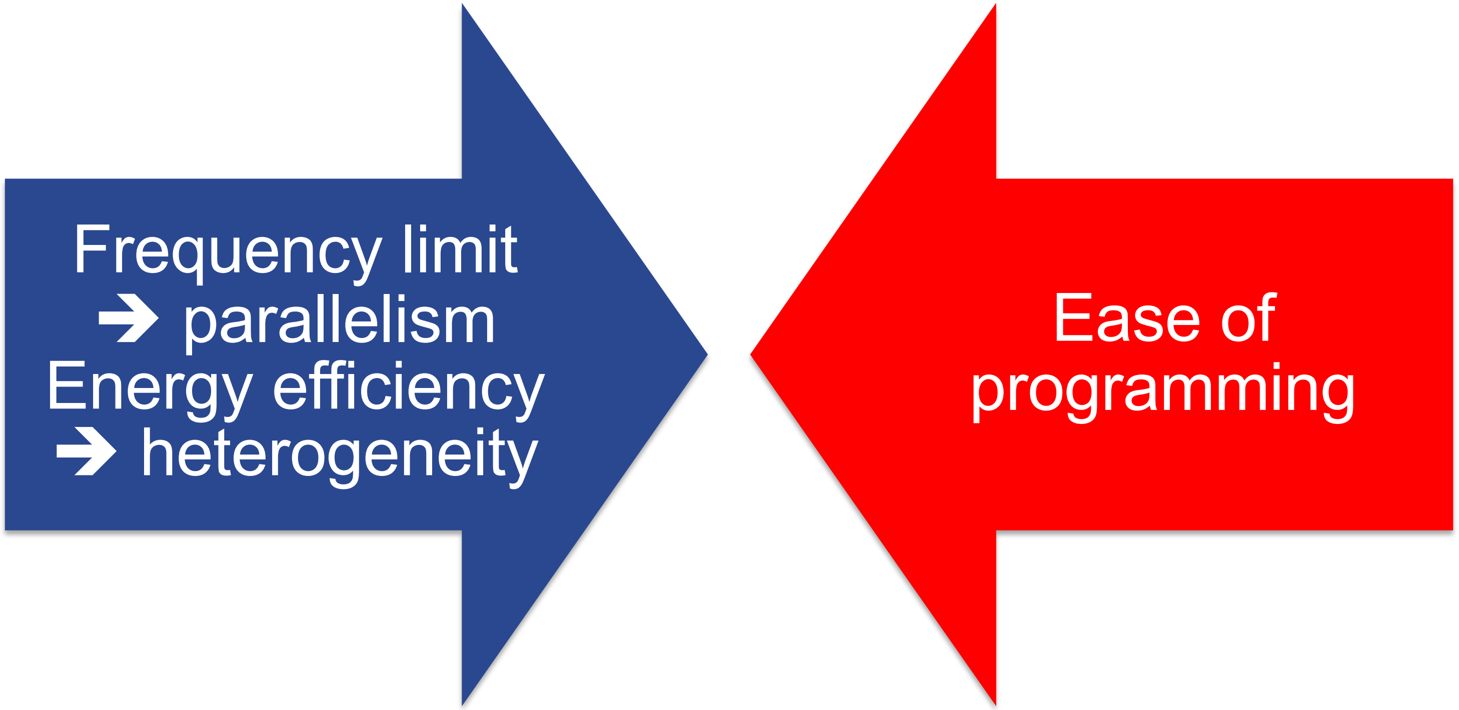
PARALLELISM AND SPECIALIZATION ARE NOT FOR FREE…



Frequency limit
→ parallelism
Energy efficiency
→ heterogeneity

Ease of programming

PARALLELISM AND SPECIALIZATION ARE NOT FOR FREE…



Frequency limit
→ parallelism
Energy efficiency
→ heterogeneity

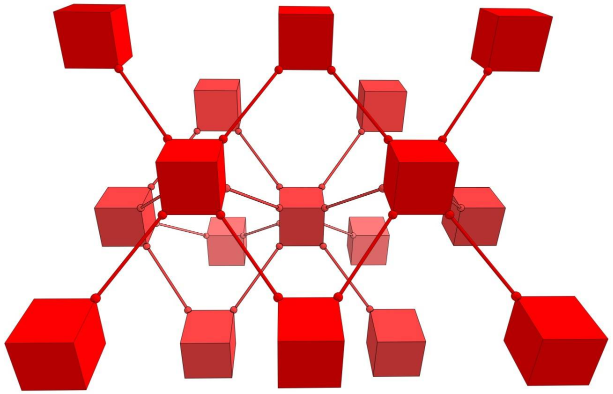
Ease of programming

MANAGING COMPLEXITY….

"Nontrivial software written with threads, semaphore, and mutexes is incomprehensible by humans"

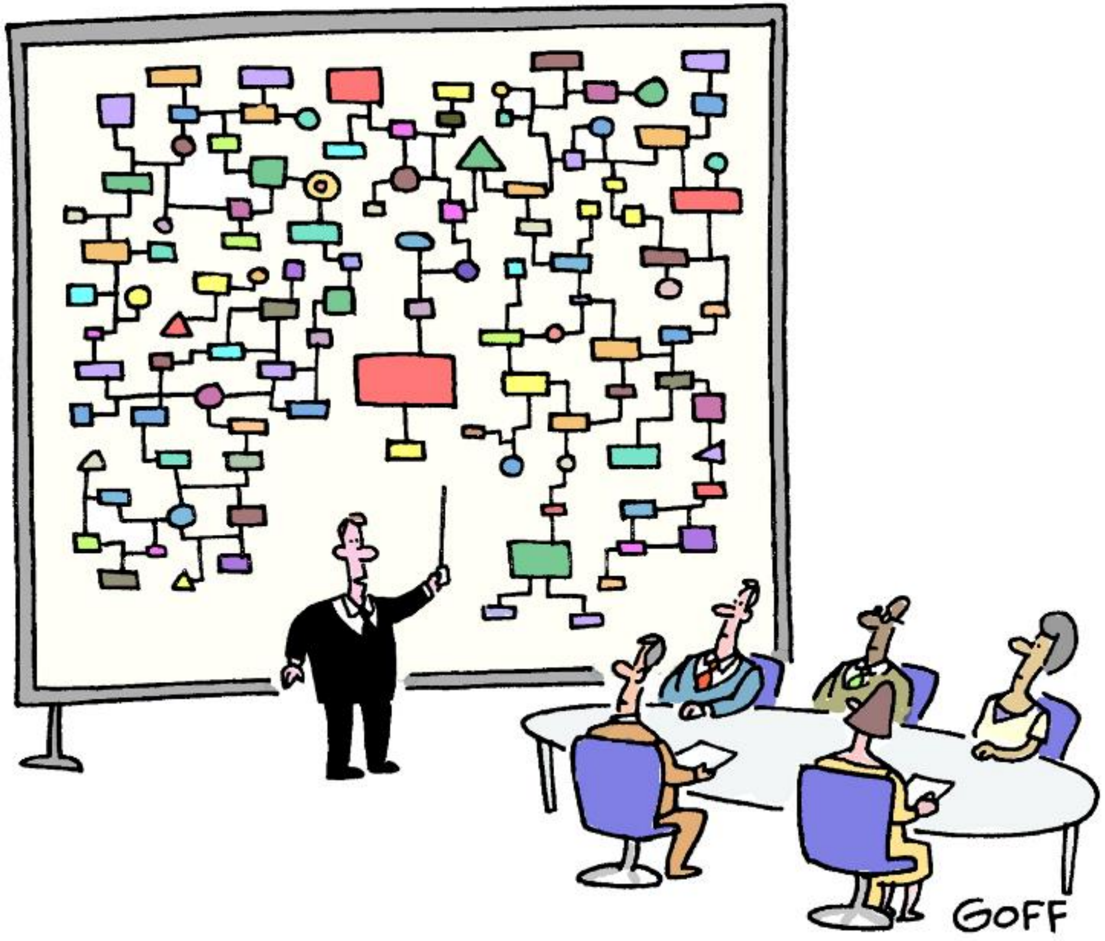
Edward A. Lee

The future of embedded software
ARTEMIS 2006



Parallelism, multi-cores, heterogeneity, distributed computing, seems to be too complex for humans ?

MANAGING COMPLEXITY



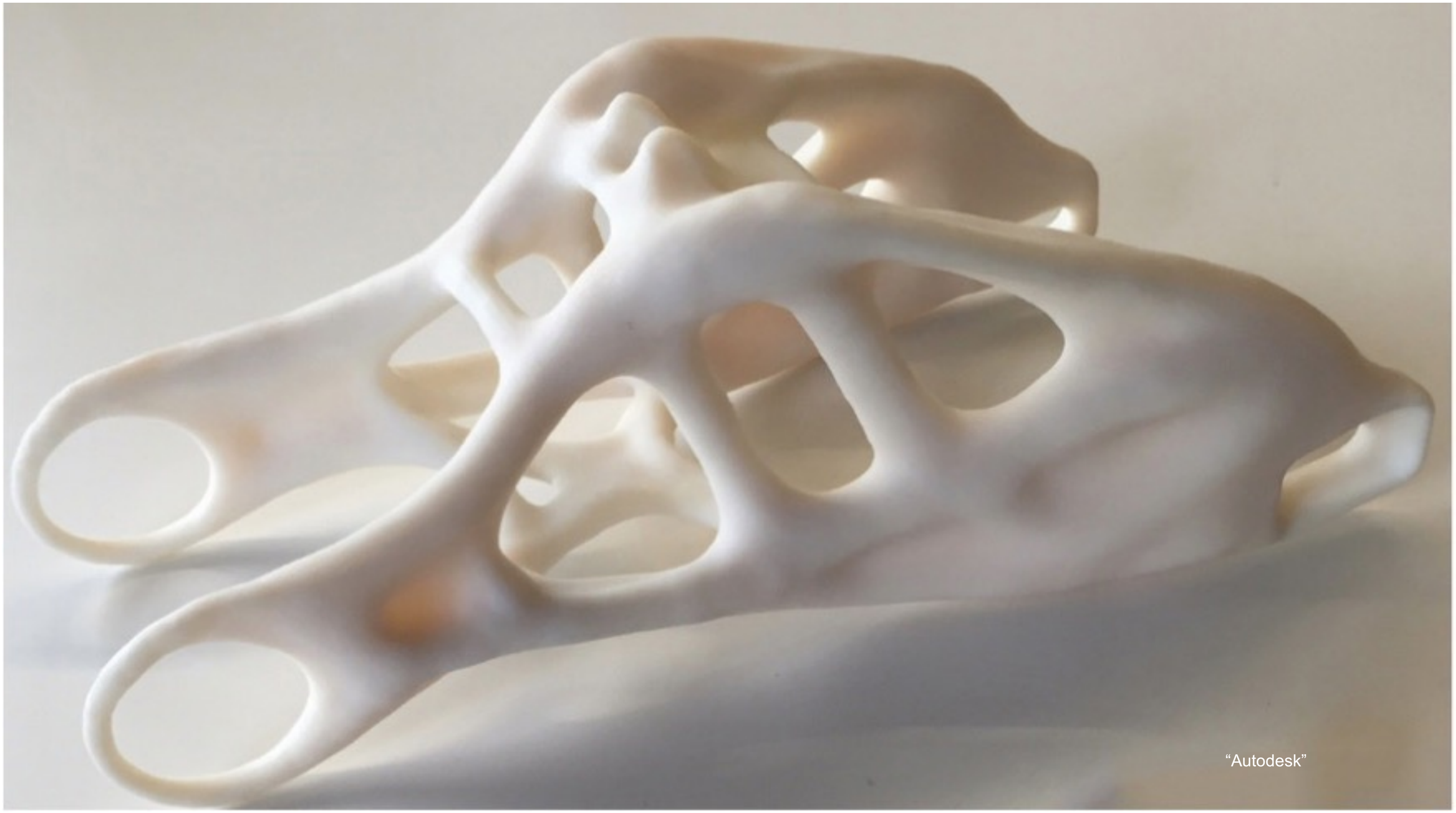
"And that's why we need a computer."

Cognitive solutions for complex computing systems:

- Using **AI and optimization techniques for computing systems**
  - Creating new hardware
  - Generating code
  - Optimizing systems
- Similar to ***Generative design*** for mechanical engineering

USING AI FOR MAKING COMPUTING SYSTEMS: “GENERATIVE DESIGN” APPROACH

The user ***only states desired goals and constraints***
-> The ***complexity wall*** might ***prevent explaining*** the solution



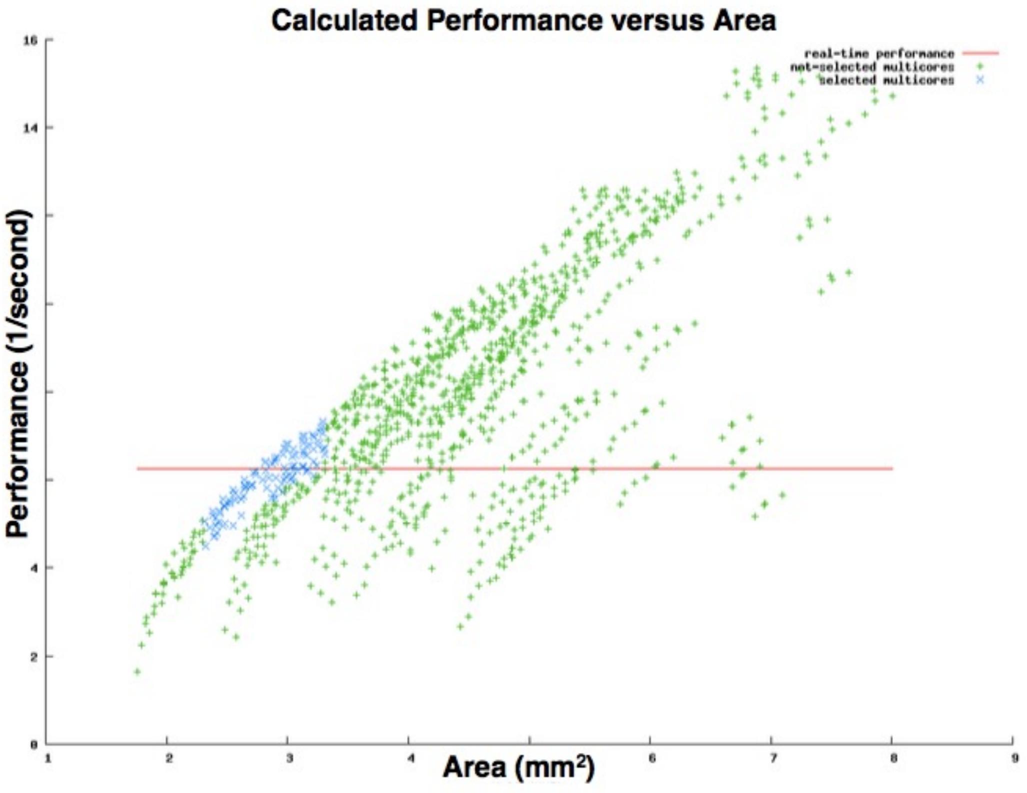
“Autodesk”

Motorcycle swingarm: the piece that hinges the rear wheel to the bike’s frame

EXAMPLE: DESIGN SPACE EXPLORATION FOR DESIGN MULTI-CORE PROCESSORS[1] (2010)

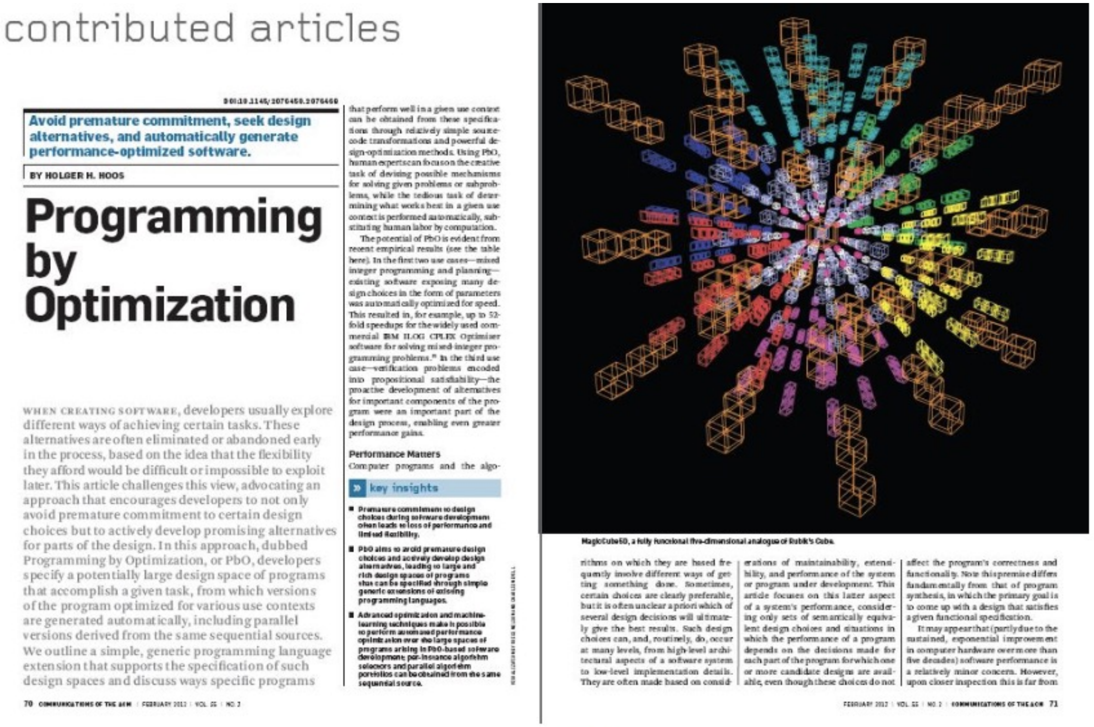
- **Ne-XVP project – Follow-up of the TriMedia VLIW (https://en.wikipedia.org/wiki/Ne-XVP)**
- **1,105,747,200 heterogeneous multicores in the design space**
- **2 millions years to evaluate all design points**
- **“AI inspired” techniques allowed to reduce the induction time to only few days**

=> *x16 performance increase*



[1] M. Duranton et all., “Rapid Technology-Aware Design Space Exploration for Embedded HeterogeneousMultiprocessors” in Processor and System-on-Chip Simulation, Ed. R. Leupers, 2010

THIS IS ALSO VALID FOR SOFTWARE: AUTOML AND OTHER PROGRAM GENERATORS



contributed articles

DOI:10.1145/2076450.2076469

Avoid premature commitment, seek design alternatives, and automatically generate performance-optimized software.

BY HOLGER H. HOOS

Programming by Optimization

WHEN CREATING SOFTWARE, developers usually explore different ways of achieving certain tasks. These alternatives are often eliminated or abandoned early in the process, based on the idea that the flexibility they afford would be difficult or impossible to exploit later. This article challenges this view, advocating an approach that encourages developers to not only avoid premature commitment to certain design choices but to actively develop promising alternatives for parts of the design. In this approach, dubbed Programming by Optimization, or PbO, developers specify a potentially large design space of programs that accomplish a given task, from which versions of the program optimized for various use contexts are generated automatically, including parallel versions derived from the same sequential sources. We outline a simple, generic programming language extension that supports the specification of such design spaces and discuss ways specific programs that perform well in a given use context can be obtained from these specifications through relatively simple source-code transformations and powerful design-optimization methods. Using PbO, human experts can focus on the creative task of devising possible mechanisms for solving given problems or subproblems, while the tedious task of determining what works best in a given use context is performed automatically, substituting human labor by computation.

The potential of PbO is evident from recent empirical results (see the table here). In the first two use cases—mixed integer programming and planning—existing software exposing many design choices in the form of parameters was automatically optimized for speed. This resulted in, for example, up to 52-fold speedups for the widely used commercial IBM ILOG CPLEX Optimiser software for solving mixed-integer programming problems.[a] In the third use case—verification problems encoded into propositional satisfiability—the proactive development of alternatives for important components of the program were an important part of the design process, enabling even greater performance gains.

Performance Matters

Computer programs and the algorithms on which they are based frequently involve different ways of getting something done. Sometimes, certain choices are clearly preferable, but it is often unclear a priori which of several design decisions will ultimately give the best results. Such design choices can, and, routinely, do, occur at many levels, from high-level architectural aspects of a software system to low-level implementation details. They are often made based on considerations of maintainability, extensibility, and performance of the system or program under development. This article focuses on this latter aspect of a system's performance, considering only sets of semantically equivalent design choices and situations in which the performance of a program depends on the decisions made for each part of the program for which one or more candidate designs are available, even though these choices do not affect the program's correctness and functionality. Note this premise differs fundamentally from that of program synthesis, in which the primary goal is to come up with a design that satisfies a given functional specification.

It may appear that (partly due to the sustained, exponential improvement in computer hardware over more than five decades) software performance is a relatively minor concern. However, upon closer inspection this is far from

» key insights

- Premature commitment to design choices during software development often leads to loss of performance and limited flexibility.
- PbO aims to avoid premature design choices and actively develop design alternatives, leading to large and rich design spaces of programs that can be specified through simple generic extensions of existing programming languages.
- Advanced optimization and machine-learning techniques make it possible to perform automated performance optimization over the large spaces of programs arising in PbO-based software development; per-instance algorithm selectors and parallel algorithm portfolios can be obtained from the same sequential source.

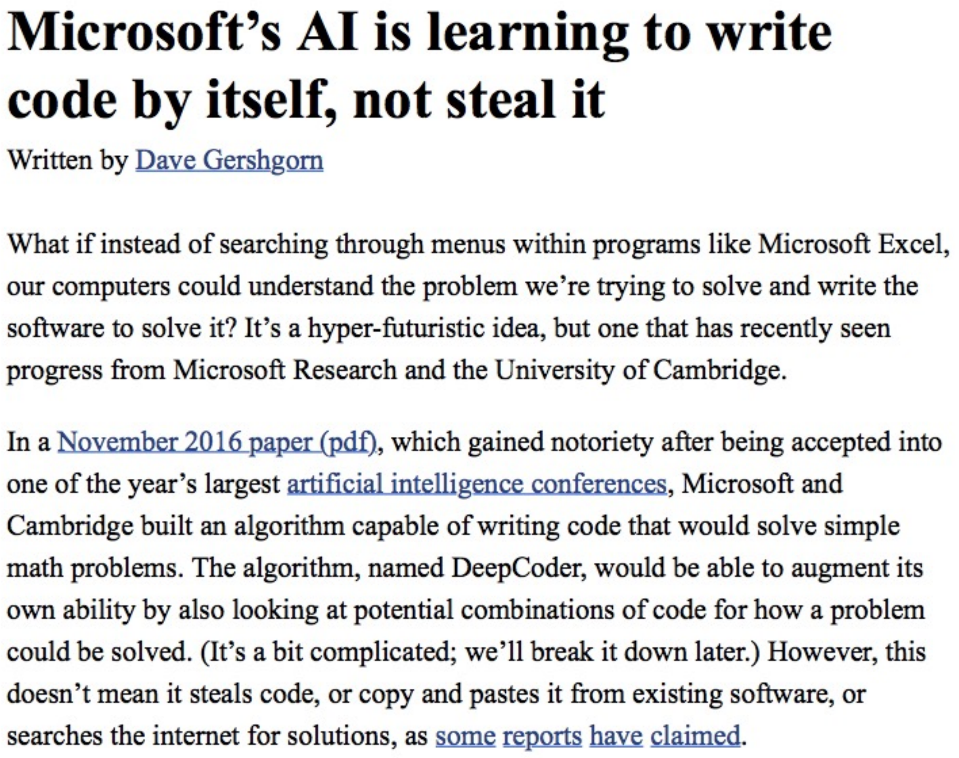
MagicCube5D, a fully functional five-dimensional analogue of Rubik's Cube.

70 COMMUNICATIONS OF THE ACM | FEBRUARY 2012 | VOL. 55 | NO. 2

FEBRUARY 2012 | VOL. 55 | NO. 2 | COMMUNICATIONS OF THE ACM 71

Communications of the ACM, 55(2), pp. 70–80, February 2012

www.prog-by-opt.net



Microsoft's AI is learning to write code by itself, not steal it

Written by Dave Gershgorn

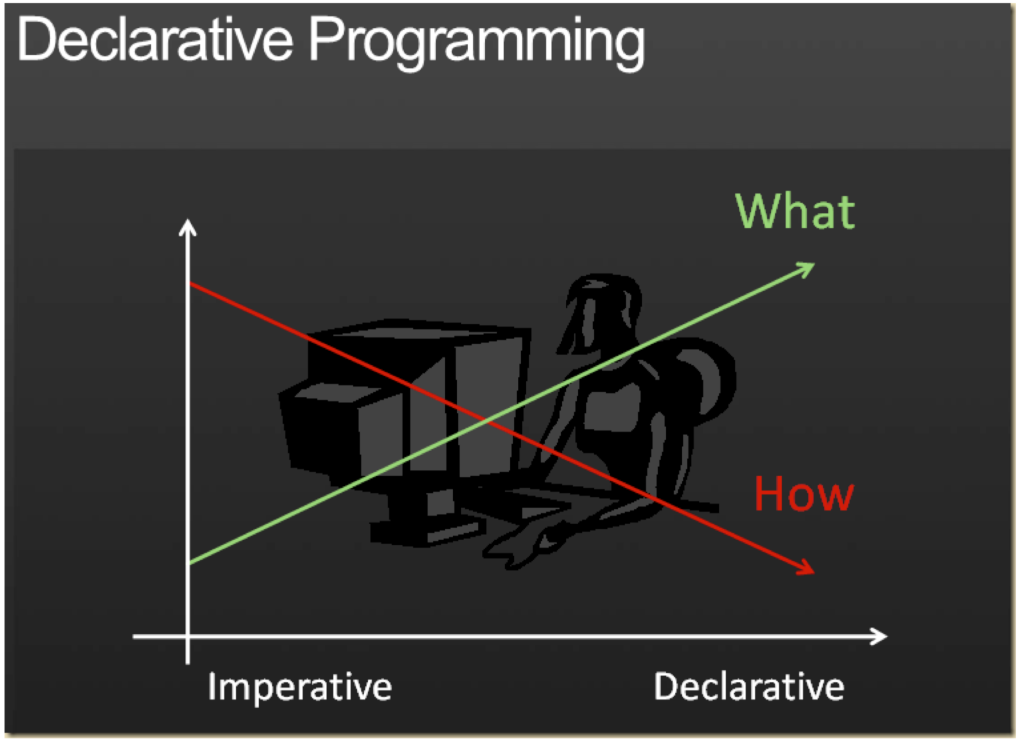
What if instead of searching through menus within programs like Microsoft Excel, our computers could understand the problem we're trying to solve and write the software to solve it? It's a hyper-futuristic idea, but one that has recently seen progress from Microsoft Research and the University of Cambridge.

In a November 2016 paper (pdf), which gained notoriety after being accepted into one of the year's largest artificial intelligence conferences, Microsoft and Cambridge built an algorithm capable of writing code that would solve simple math problems. The algorithm, named DeepCoder, would be able to augment its own ability by also looking at potential combinations of code for how a problem could be solved. (It's a bit complicated; we'll break it down later.) However, this doesn't mean it steals code, or copy and pastes it from existing software, or searches the internet for solutions, as some reports have claimed.

PROGRAMMING 2.0: LET THE COMPUTER DO THE JOB

Describing *what* the program should accomplish, rather than describing *how* to accomplish it

- For example, describe the *concurrency* of an application, not how to parallelize the code for it.
- (Good) compilers know better about architecture than humans, they are better at optimizing code…



Declarative Programming

CONCLUSION

CONCLUSION: WE LIVE AN EXCITING TIME!

> ***“The best way to predict the future is to invent it.”***
>
> *Alan Kay*

Thank you for your attention

Special thank you to Denis Dutoit, Christian Gamrat, Carlo Reita for their slides I borrowed.

Commissariat à l'énergie atomique et aux énergies alternatives

Institut List | CEA SACLAY NANO-INNOV | BAT. 861 – PC142
91191 Gif-sur-Yvette Cedex - FRANCE
www-list.cea.fr

www.cea-tech.fr
Établissement public à caractère industriel et commercial | RCS Paris B 775 685 0

Institut Leti, technology research institute | Minatec Campus | 17 rue des Martyrs | 38054 Grenoble Cedex | FRANCE
www.leti.fr