2018 ECTC Plenary Session
"Artificial Intelligence and Its Impact on System Design"

Chair: Kemal Aygun – Intel

Panelists: Igor Arsovski – GLOBALFOUNDRIES

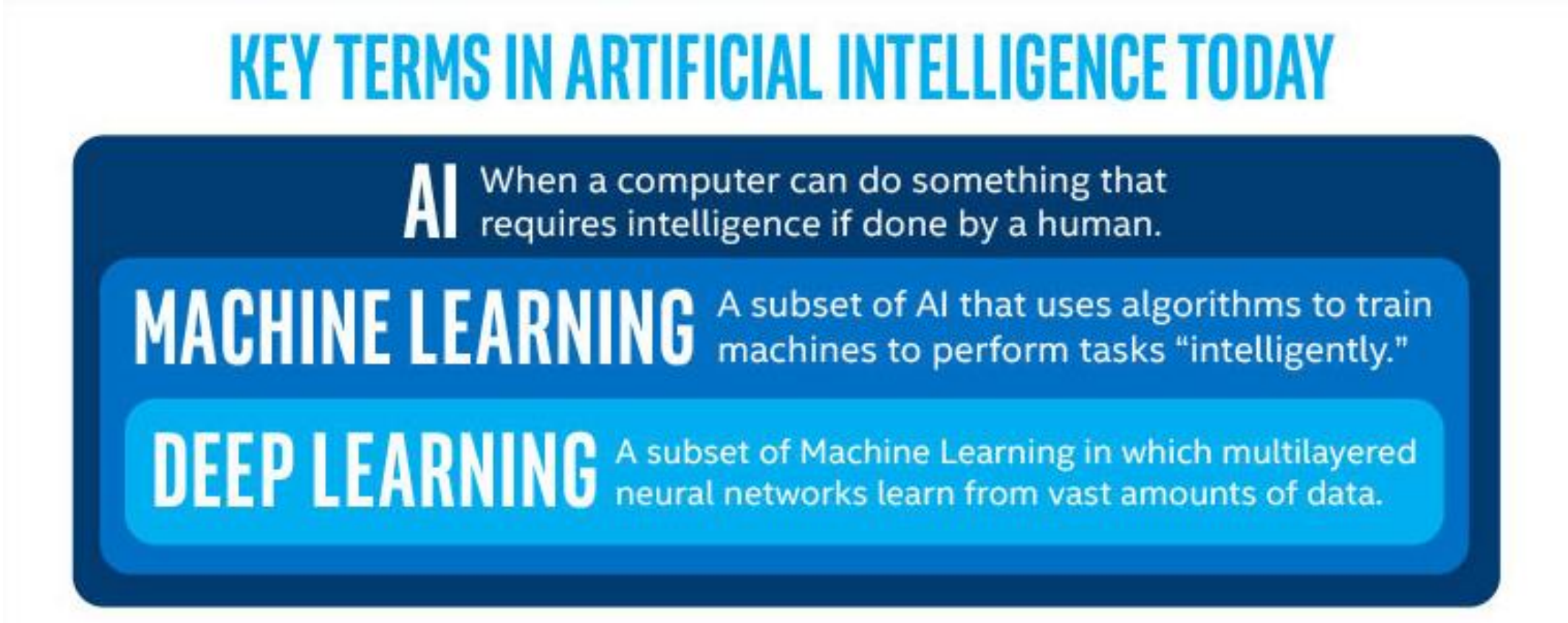
Kailash Gopalakrishnan – IBM

Andrew Putnam – Microsoft

Dan Oh – Samsung

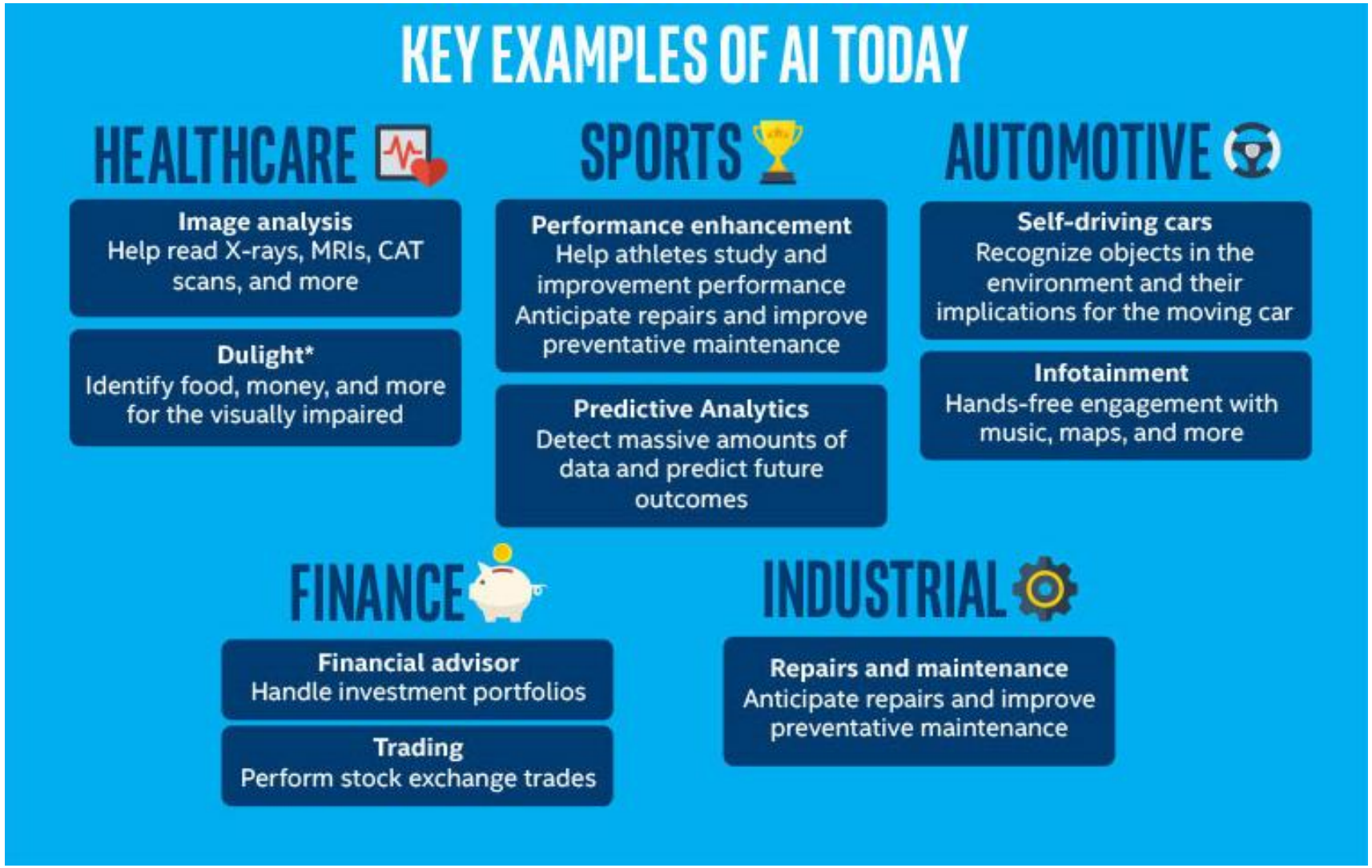
Madhavan Swaminathan – Georgia Tech

What is AI?



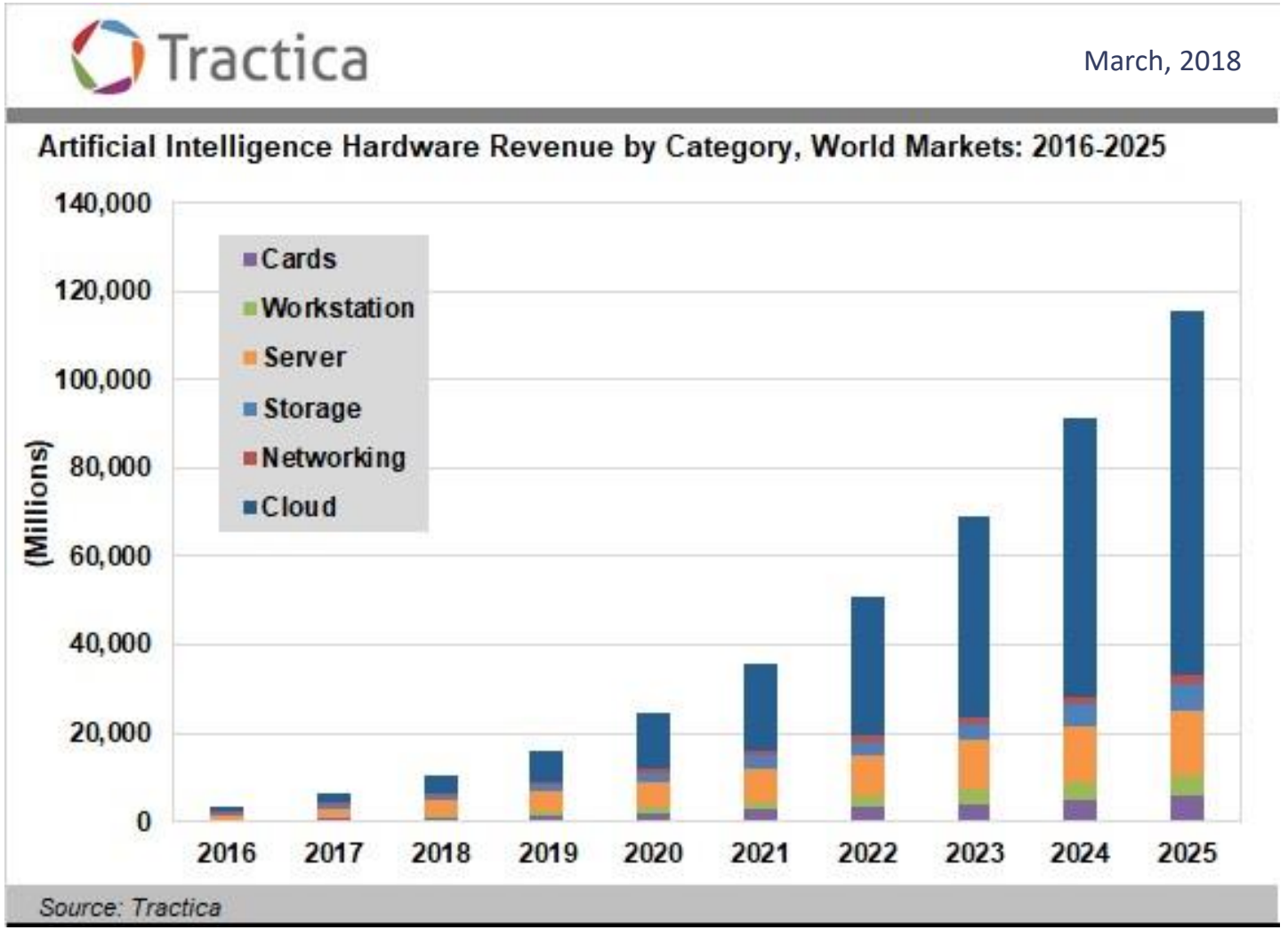
Source: Niven Singh (https://software.intel.com)

Where is AI used?



Source: Niven Singh (https://software.intel.com)

Why do we care?



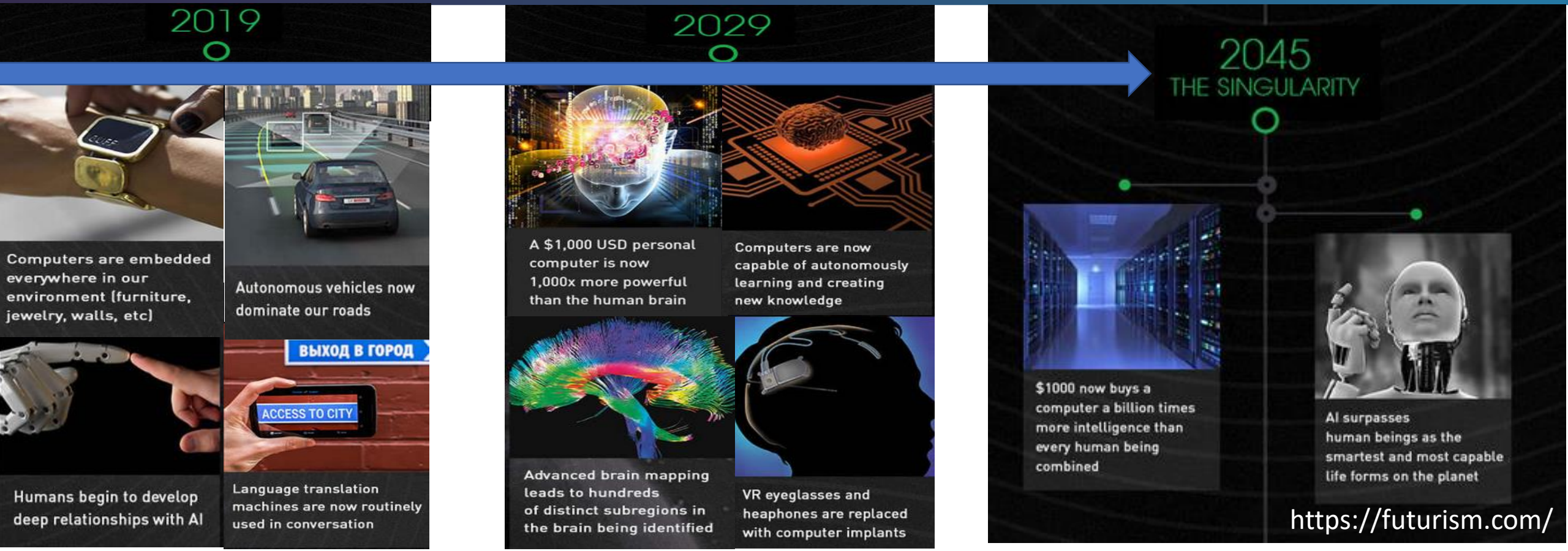
https://www.tractica.com/newsroom/press-releases/artificial-intelligence-driven-hardware-sales-will-reach-115-billion-worldwide-by-2025/

Reproduced by permission of Tractica

Speaker Bio: Igor Arsovski

- Igor Arsovski is the Chief Technical Officer of the GlobalFoundry's ASIC Business Unit. He is responsible for ASIC Artificial Intelligence Strategy including IP and Methodology.
- His narrow focus is in semiconductor memories. His extended focus is energy efficient building blocks for Machine Learning and Automotive Electronics including 3D memory integration.
- Igor has authored 15 IEEE papers, and filed over 80 US patents.

Predictions for the future of Artificial Intelligence: (some predict the emergence of the singularity by 2045)



- Most AI future predictions assume Moore's Law continues
- More than Moore architectures and packaging are going to be key to enable AI

Artificial Intelligence Devices Classification: Training & Inference (Automotive Example)

Training

Inference

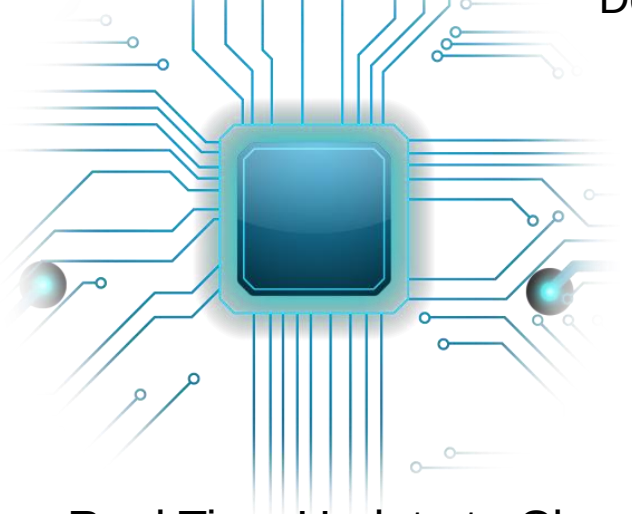
DNN Training

DNN

OTA Deployment

Data Center

Initial DNN Training and Continual Learning from Fleet

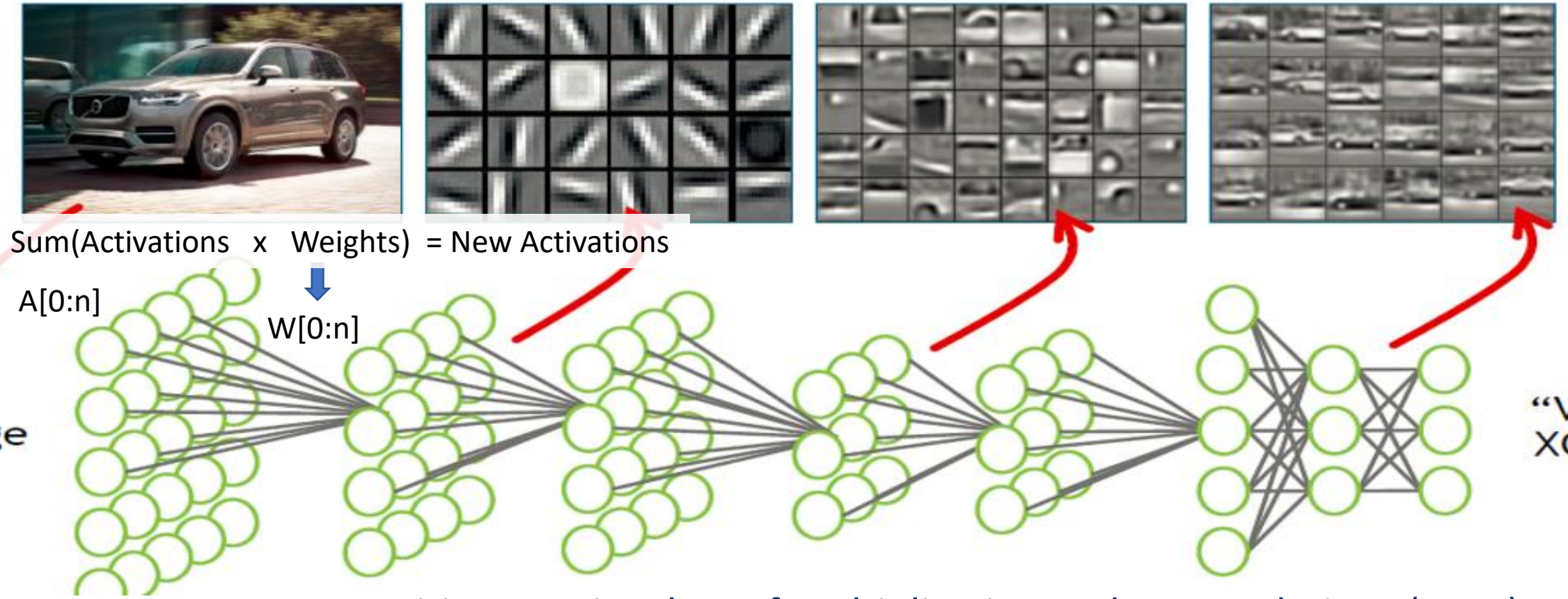


Connected Autonomous Computer

Local DNN Inferencing and Driving Decisions

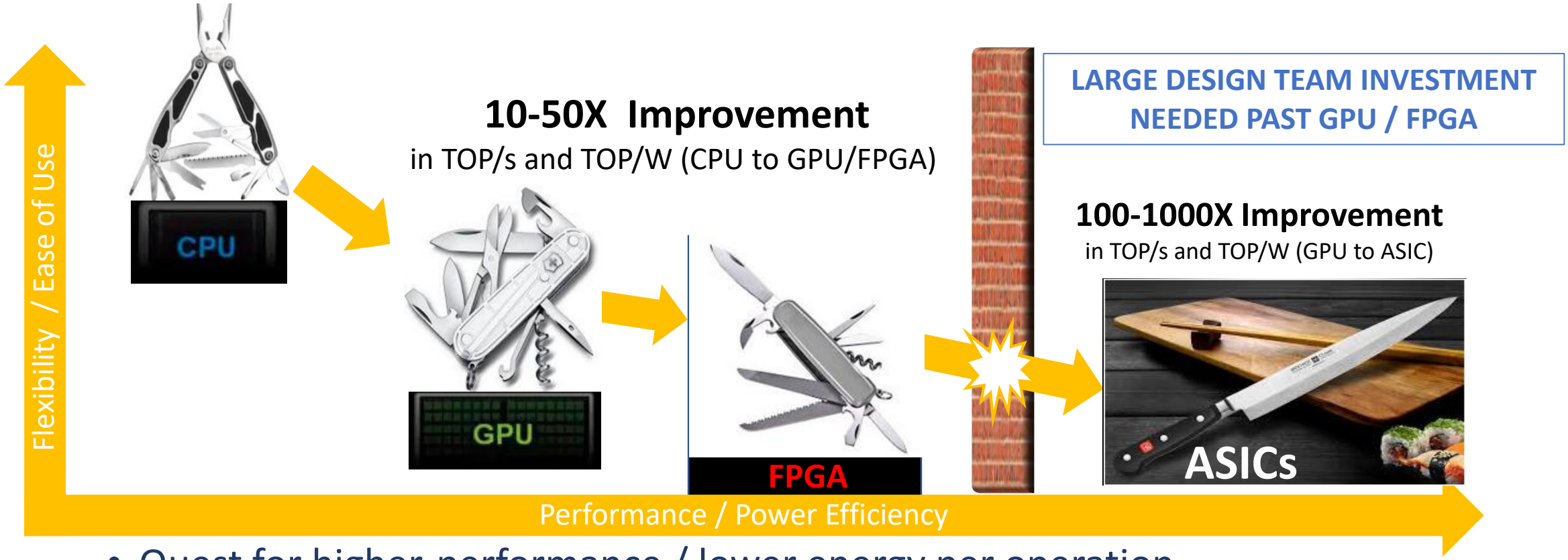
Real Time Update to Cloud

Artificial Intelligence Image Recognition Example



- Pattern recognition requires lots of Multiplication and Accumulations (MAC)
- Large data-sets requires large amount of Memory and MAC units
- Value of devices measured in TOPS/s and TOPS/W

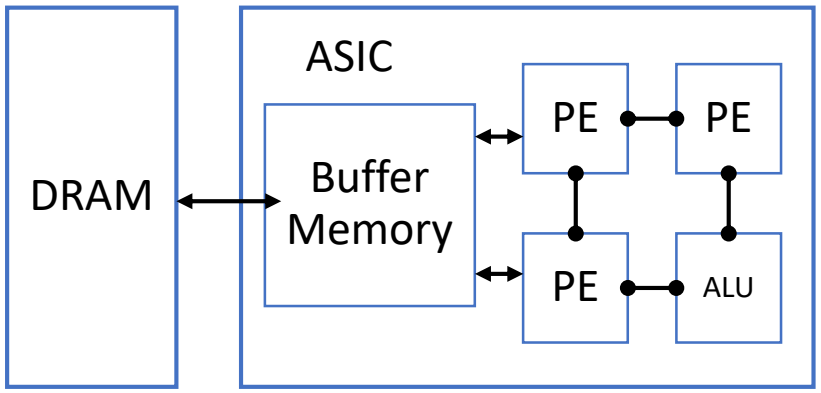
Artificial Intelligence: Industry Trends

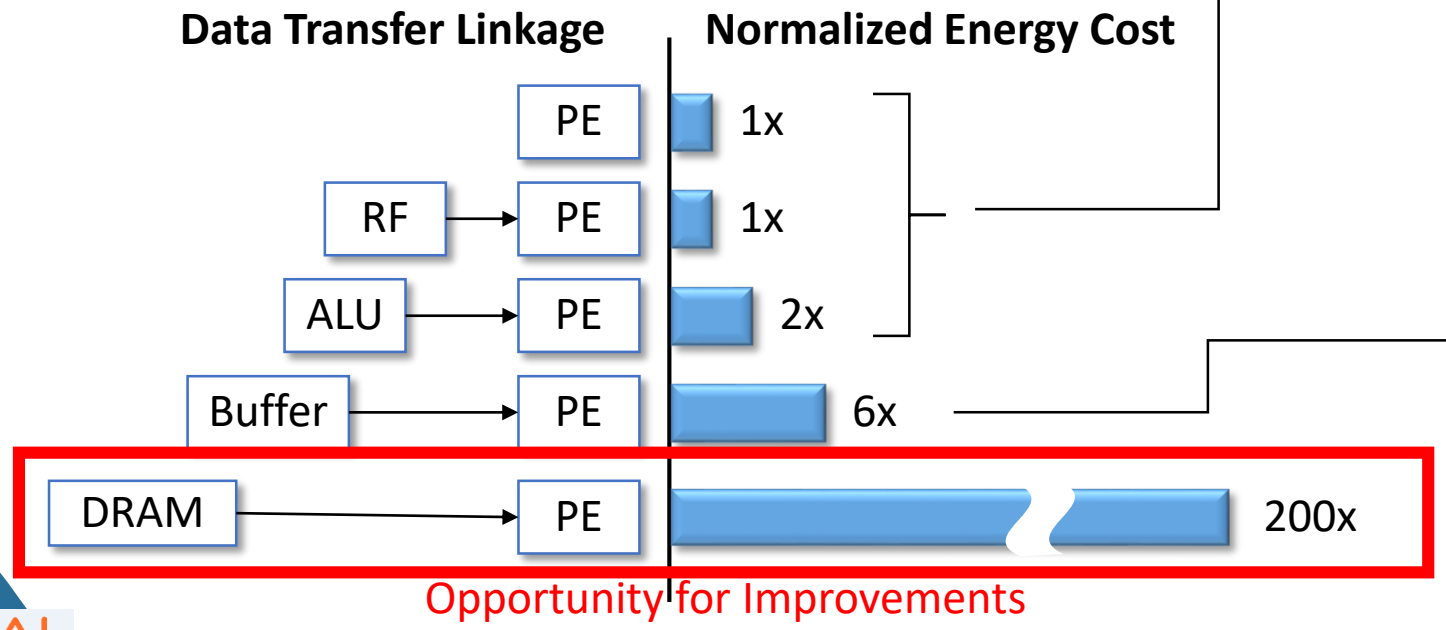


- Quest for higher-performance / lower energy per operation
- CPU to FPGA progression can be made without a chip-design team
- Move to ASIC requires a fully staffed design team

Energy Optimization in a AI designs

Source: MIT





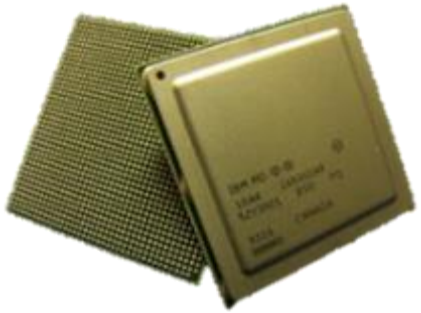
GLOBALFOUNDRIES Value Proposition

STDCELL Library
- Area/energy efficient ALU/MAC.
- DTCO support.

Local Memory
- High Speed Local SRAM
- Industry leading SRAM cell density.

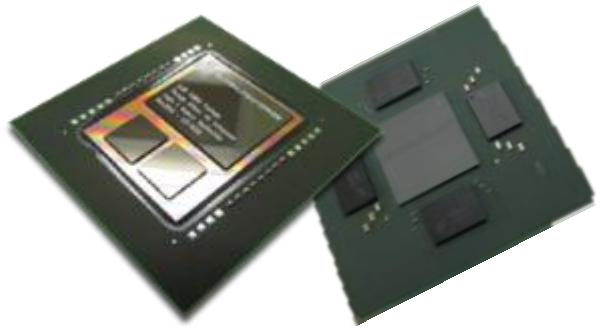
Packaging Options to Meet AI Needs
Power, Performance & Cost Needs

2D Packaging



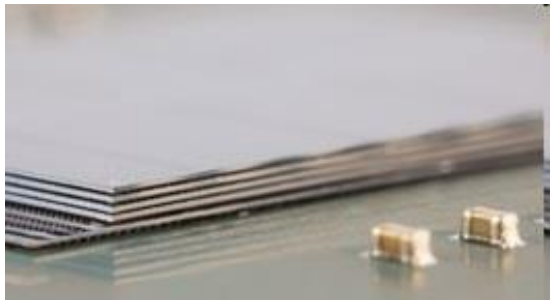
- Signaling speed increasing 30G to 112G

2.5D Integration



- 14nm HBM interface hardware verified
- Stitched interposer capability for large designs

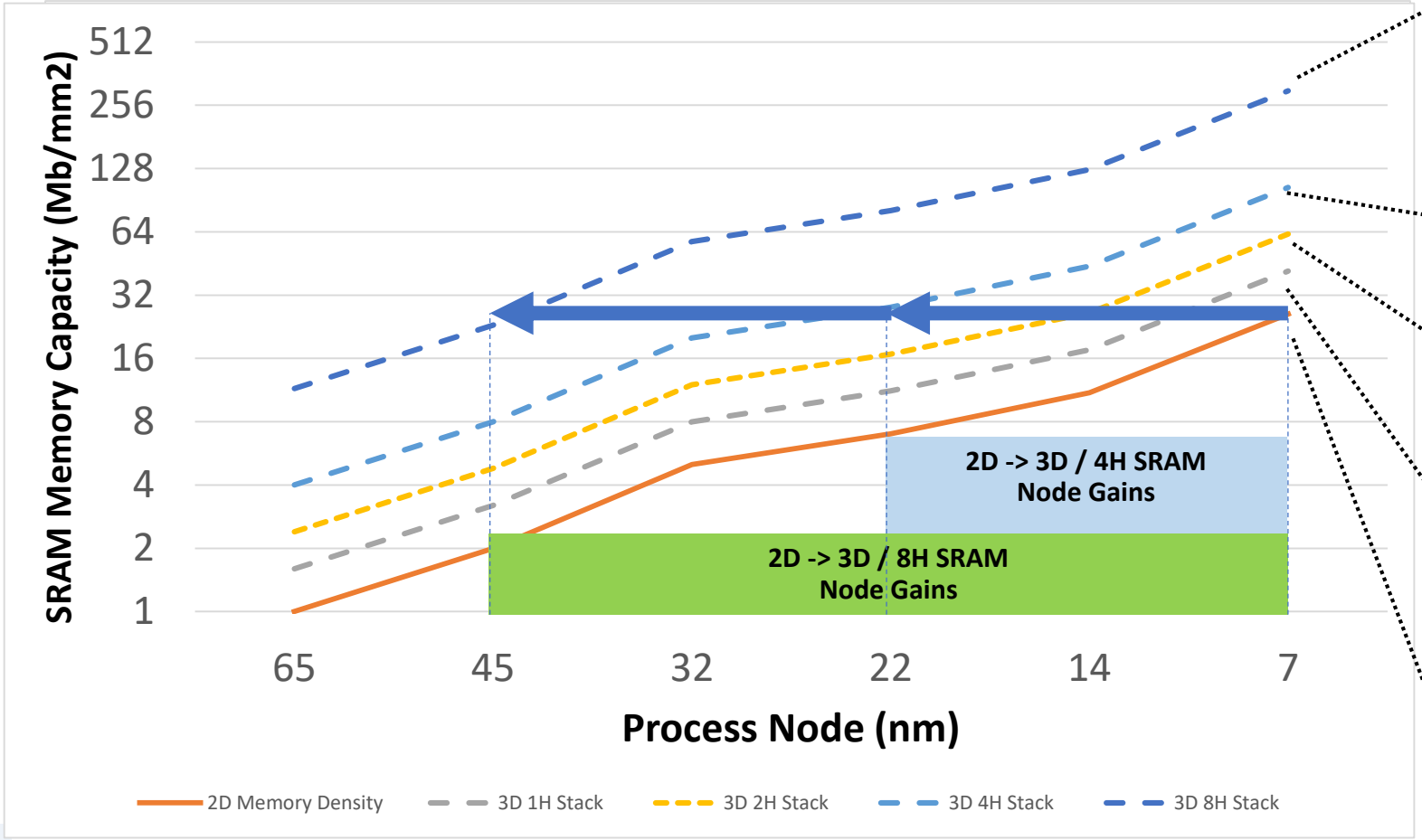
3D Integration

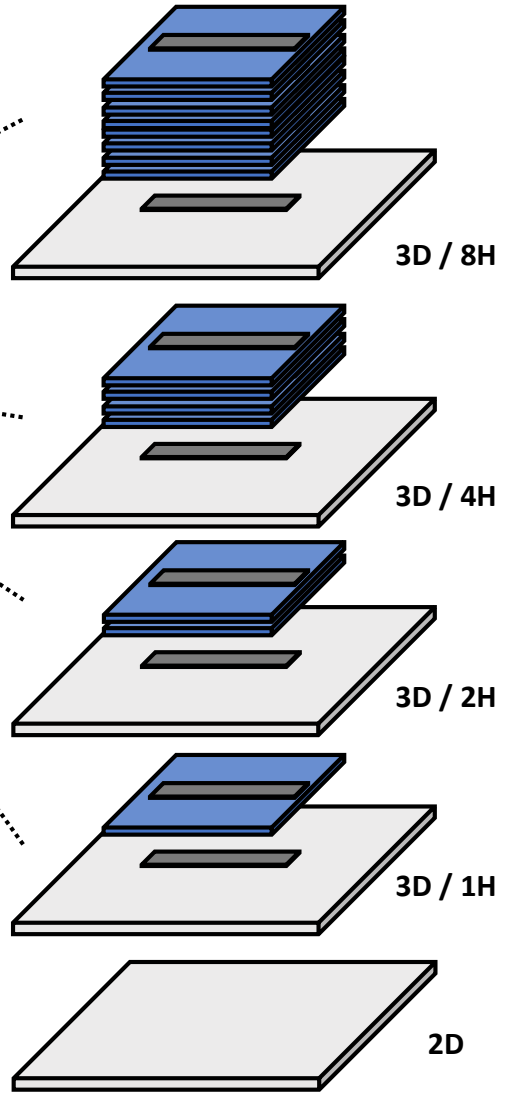


- 1st in volume production with 32nm
- Lowest interface power & smallest form factor

3D SRAM Memory Advantage

- Memory Capacity & Energy/Access critical for AI applications
- 3D stacking enables multiple node memory density scaling

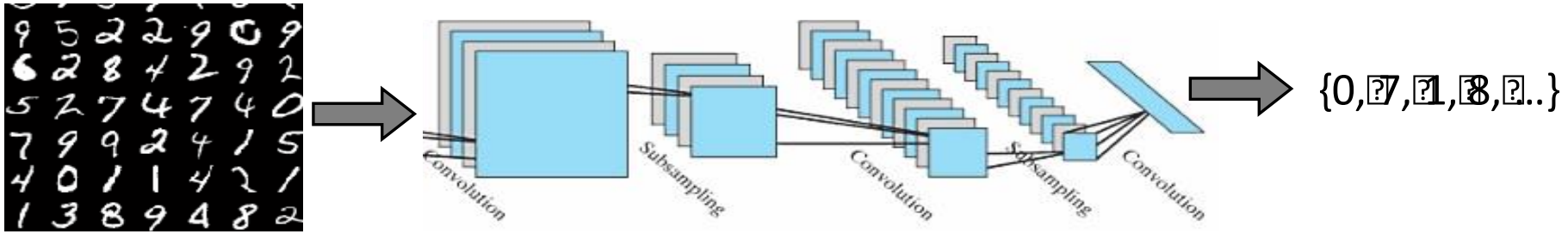




Speaker Bio: Kailash Gopalakrishnan

Kailash Gopalakrishnan is a Distinguished Research Staff member at IBM Research where he manages the Accelerator Architectures and Machine Learning group at the T. J. Watson Research Center, N.Y. Kailash has led work in the areas of semiconductor devices, emerging memory technologies, novel computer architectures, ASIC design and deep learning algorithms. His current passion is centered around hardware-software co-design of specialized architectures optimized for deep learning acceleration by pushing the boundaries of approximate computing techniques. He has a Ph.D. in Electrical Engineering from Stanford University and is a member of the IEEE.

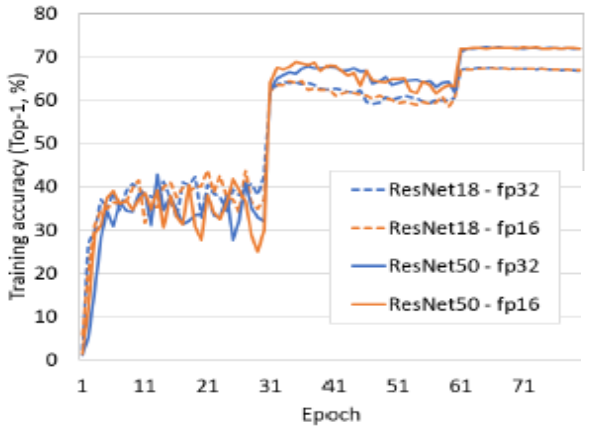
Overview



- Deep Learning Training & Inference today:
  - **Training:** Many big chips (300W) connected through proprietary links for inter-chip gradient reduction. Racks / pods in the data center – largely accelerator-centric (> 2:1 / 4:1 over CPUs) .
  - **Inference:** Huge push @ the edge + Standard PCIe attached < 75W cards in the data center.
- Strategic Thrusts:
  - Use of Approximate Computing techniques (scaled precision tuning, compression,...) to reduce computation and communication for Deep Learning – training and inference.
    1. **Scaled Precision** for **Training** (16/8?/4? bits) and **Hyper-scaled precision** for **Inference** (8/4/2?/1?bits). Impact on packaging and cooling for training.
    2. Use of **Compression techniques** to minimize bandwidth needs for **Training**. Impacts packaging.
  - Using these techniques to define **new cores for A.I.** & **Deep Learning SoCs**.

*Primarily a further out research perspective. This brief presentation reflects my research team's views largely – and not those of IBM Corp broadly.

Deep Learning Training : Computation vs. Communication



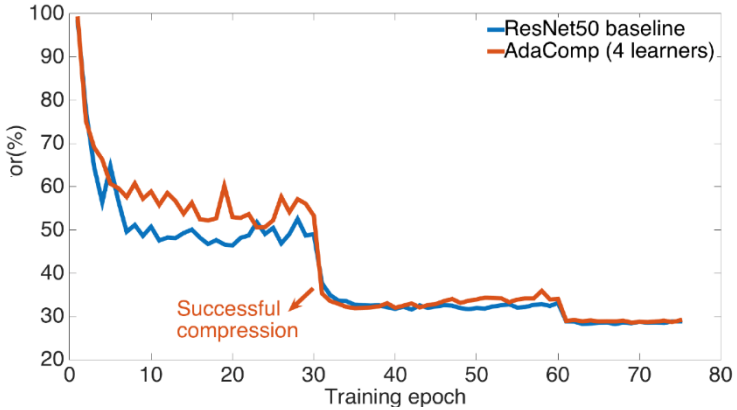
Training with 16-bits of precision

Deep Learning Processor Core for AI Training and Inference (2018 Symposium on VLSI Circuits): To be presented in June 2018

A Scalable Multi-TeraOPS Deep Learning Processor Core for AI Training and Inference

Bruce Fleischer, Sunil Shukla, Matthew Ziegler, Joel Silberman, Jinwook Oh, Vijayalakshmi Srinivasan, Jungwook Choi, Silvia Mueller[2], Ankur Agrawal, Tina Babinsky[2], Nianzheng Cao, Chia-Yu Chen, Pierce Chuang, Thomas Fox, George Gristede, Michael Guillorn, Howard Haynie[1], Michael Klaiber[2], Dongsoo Lee, Shih-Hsien Lo, Gary Maier[3], Michael Scheuermann, Swagath Venkataramani, Christos Vezyrtzis, Naigang Wang, Fanchieh Yee, Ching Zhou, Pong-Fei Lu, Brian Curran[1], Leland Chang, Kailash Gopalakrishnan

IBM TJ Watson Research Center, Yorktown Heights, NY; IBM Systems Group, [1]Poughkeepsie, NY; [2]Boeblingen Germany; [3]East Fishkill, NY



AdaComp : Lossy Compression (>50X)

- **Deep Learning Training** is a battle between raw computational throughput (Flops), memory bandwidth (**MBW**) and communication bandwidth (**CBW**).
  - Plenty of powerful 300W accelerators with lots of Flops trying to work together on 1 large problem.
- **Compute Precision** improves Flops significantly – but stresses CBW and MBW.
  - MBW & CBW are stressed since compute throughput grows ~ quadratically with reduction in precision.
- **(Lossy) Compression** techniques can dramatically improve CBW – but need to be low overhead and should not impact algorithmic convergence.
  - Will these techniques obviate the need for high bandwidth peer-to-peer connections?

Peek into the Future

- **DL Law of Precision Scaling** → expect continuous further reduction in precision.
  - **8-bit Training** on the horizon (end of the decade?) followed by **4-bit** a few years out?
  - Hyper-scaled Precision optimized DL core and system architectures – to improve computational efficiency.
- Expect severe **memory bandwidth** bottlenecks (i.e. beyond 2.5D and HBM)
  - Will drive the use of **3D** stacking – memory (cache/scratch-pads) on top of of the processor for compute efficiency improvements.
  - Thermal challenges - given the high (>300W) power envelope
- **Off-chip I/O** for peer-to-peer accelerator connections is a little less predictable
  - Past few generations have pushed more I/O links into the accelerator (e.g. NVLINK).
  - DL Compression schemes (if > 50X) may significantly reduce bandwidth needs.
  - This could simplify packaging & board design and facilitate the use of standard compliant links.

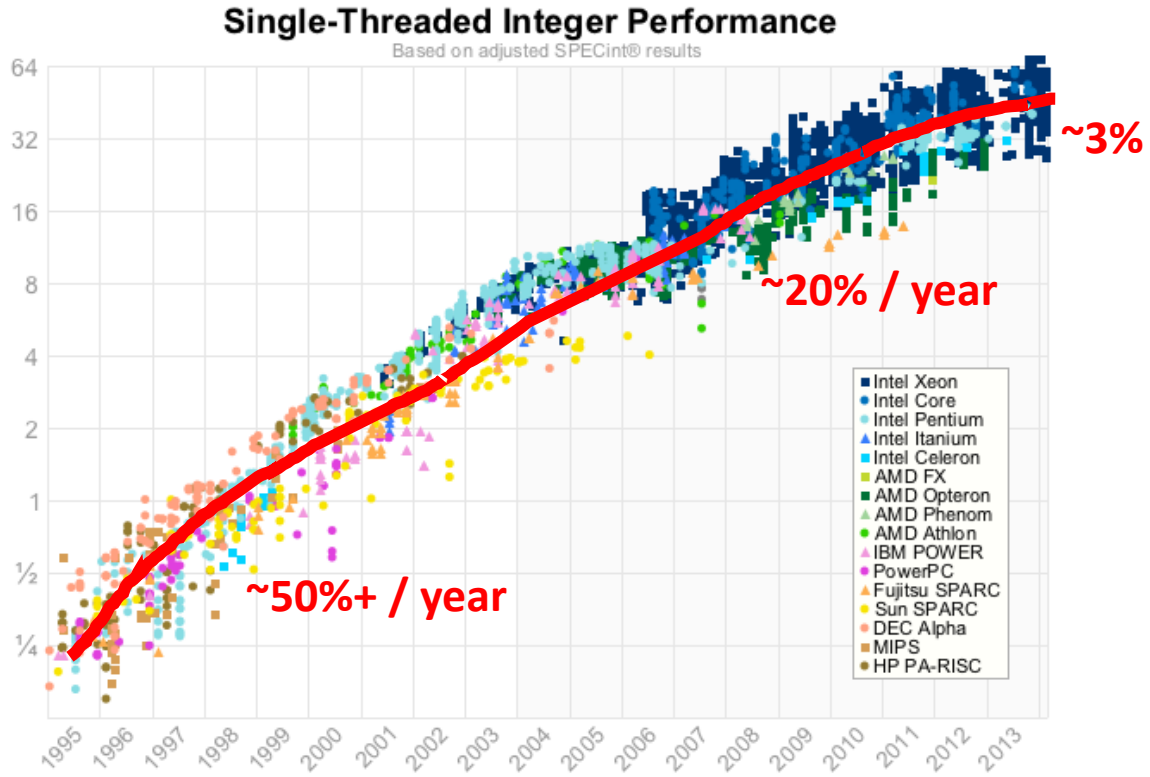
Speaker Bio: Andrew Putnam

- Principal Engineer in Microsoft Azure
- Joined Microsoft Research in 2009 after Ph.D. from U. of Washington CSE
- Co-Founder of the Microsoft Catapult project, the first to put FPGAs in every server in the datacenter
  - Bing web search acceleration
  - Azure SmartNIC for Accelerated Networking
  - Project BrainWave deep learning acceleration platform
- Currently leading the Azure SmartNIC FPGA team in Azure Networking

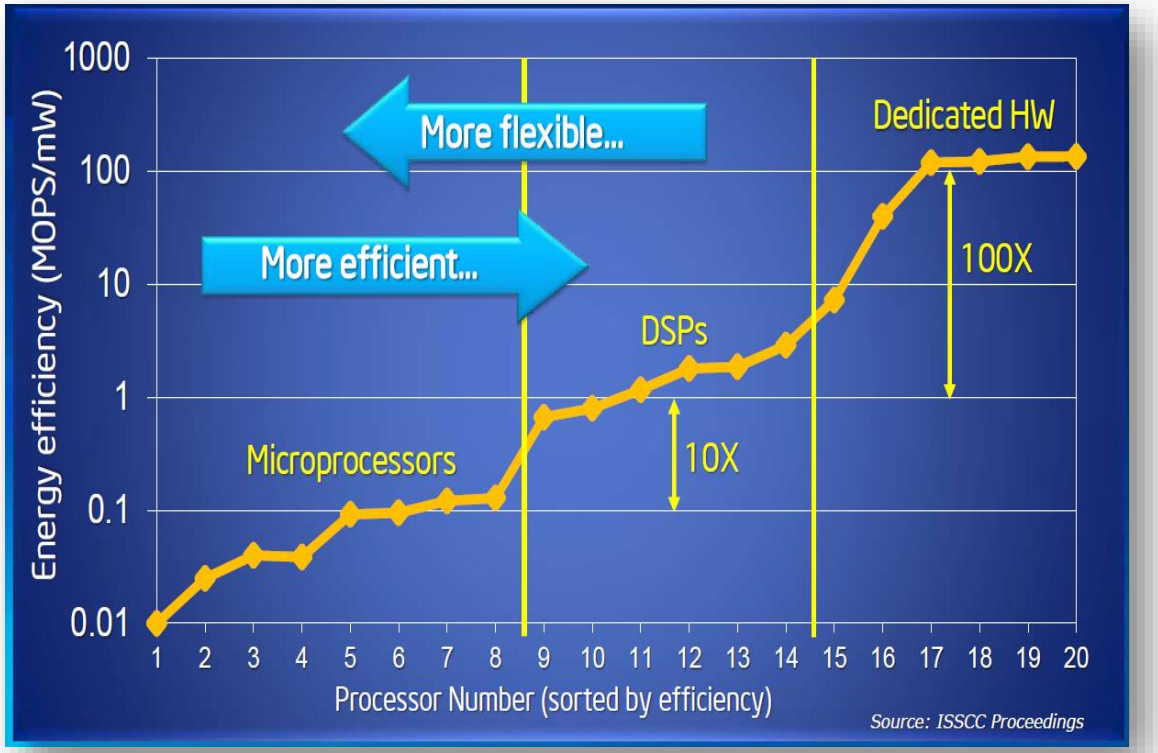
Cloud Growth is Exponential

~100%+ Growth for the past 5 years

Toward Specialization



Jeff Preshing, Henk Poley, http://preshing.com/20120208/a-look-back-at-single-threaded-cpu-performance/

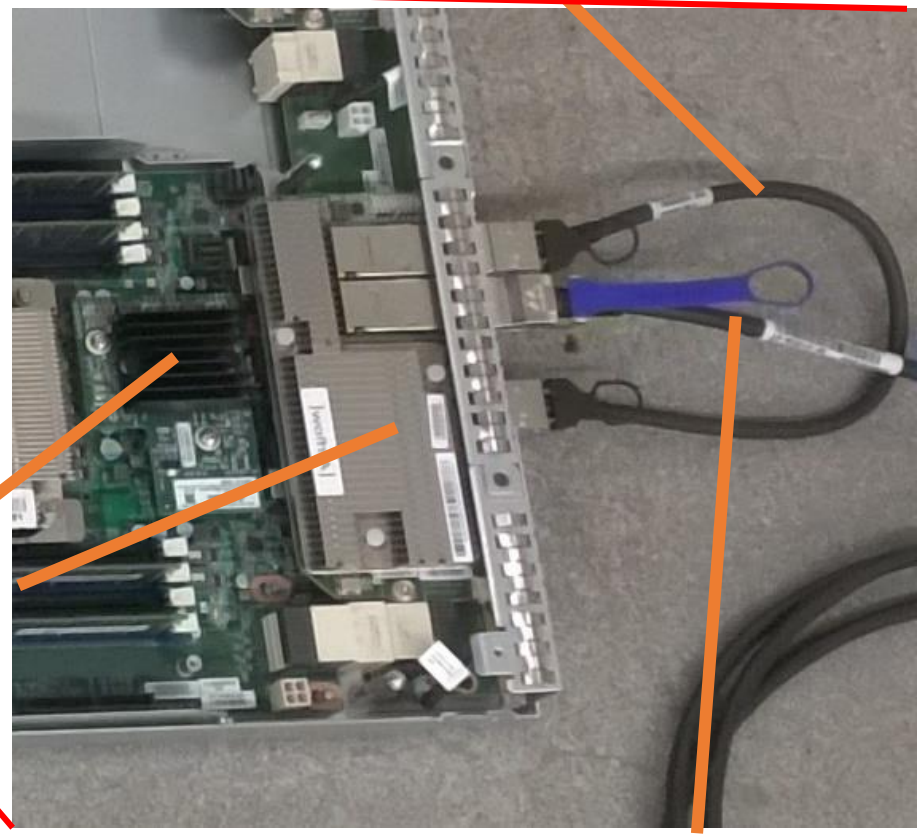


Source: Bob Broderson, Berkeley Wireless group

CPU performance isn't increasing

So now we need to specialize

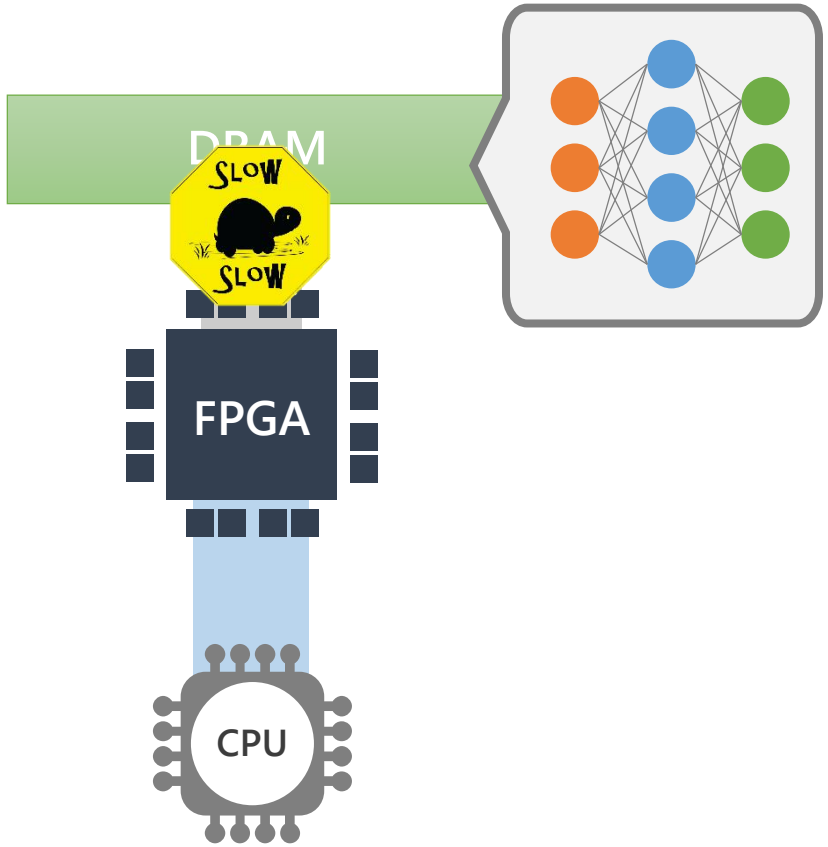
FPGAs in the Datacenter – Project Catapult



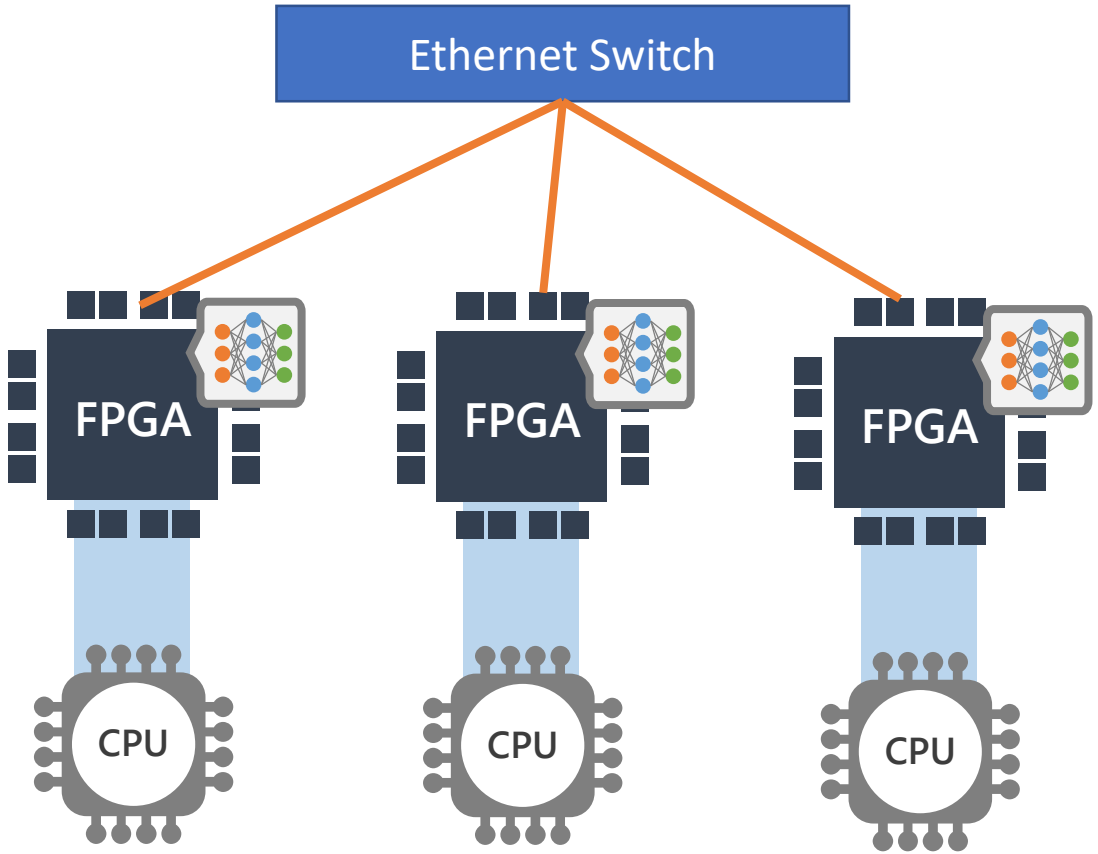
- Bump-in-the-wire architecture
- One FPGA in every server
  Microsoft has deployed since 2015

Microsoft now does RTL design!

Project Brainwave – Deep ML on FPGA



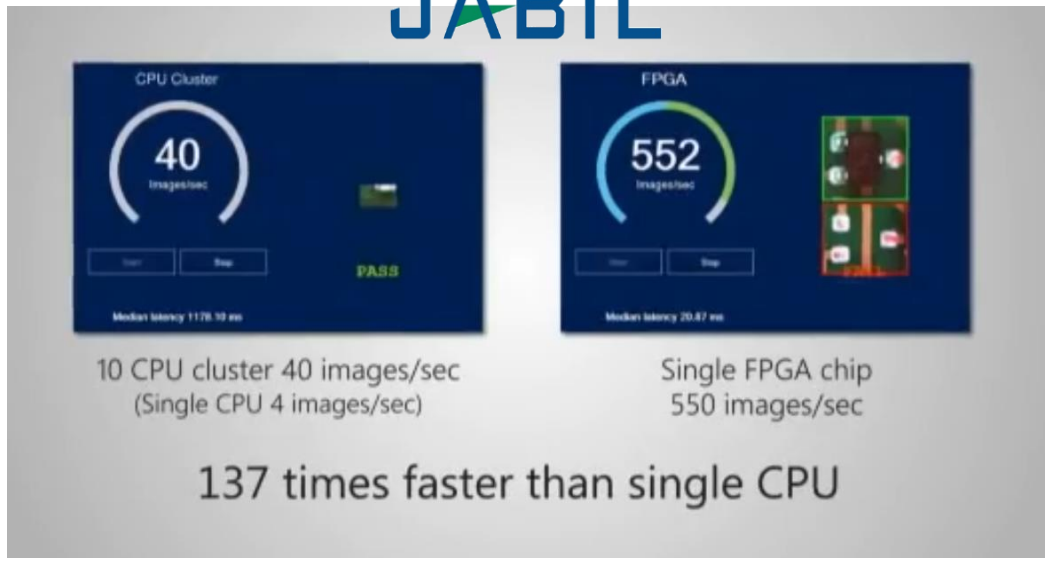
Traditional Approach



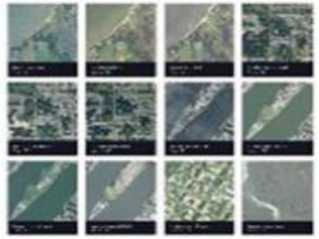
Project Brainwave

ResNet-50: 8 billion operations per image

JABIL



esri Satellite Images for the Entire US



NAIP Data

Stored on **Azure Premium Storage**

200M Images, 20TB

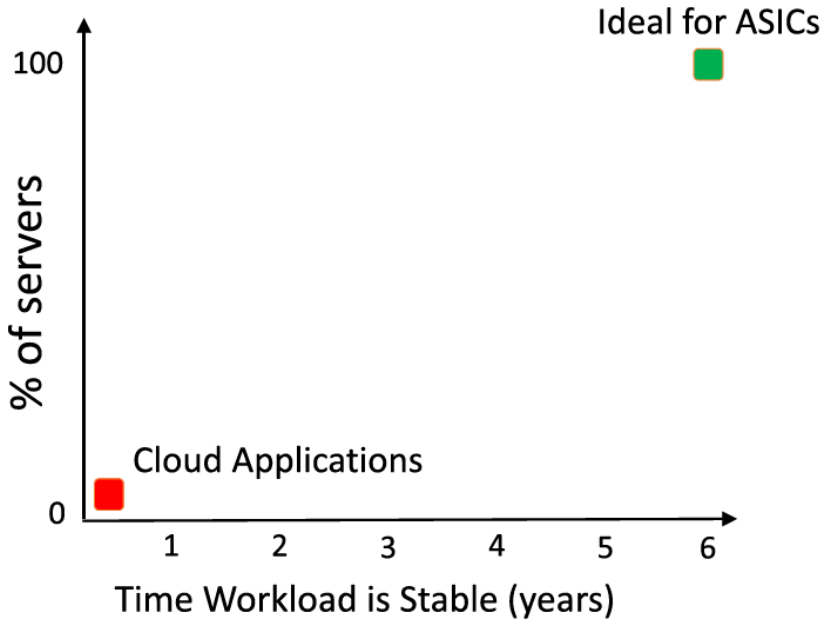
Land cover mapping for the whole of US in

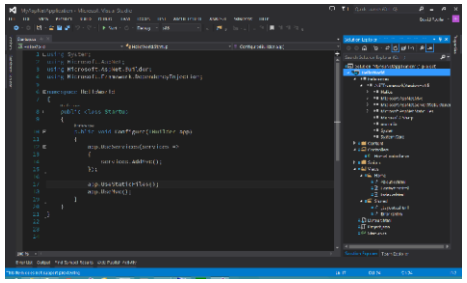
10+ minutes

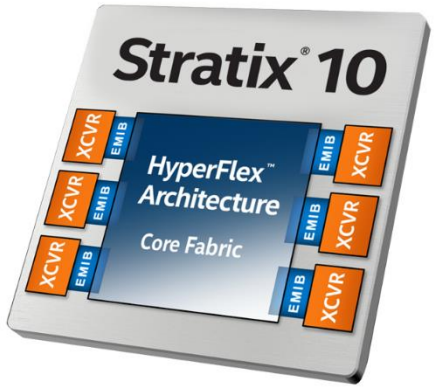
$42

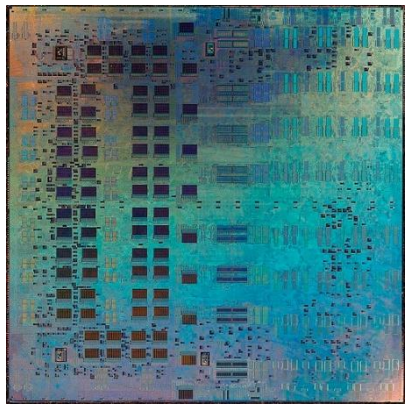
bing

Why not ASICs?









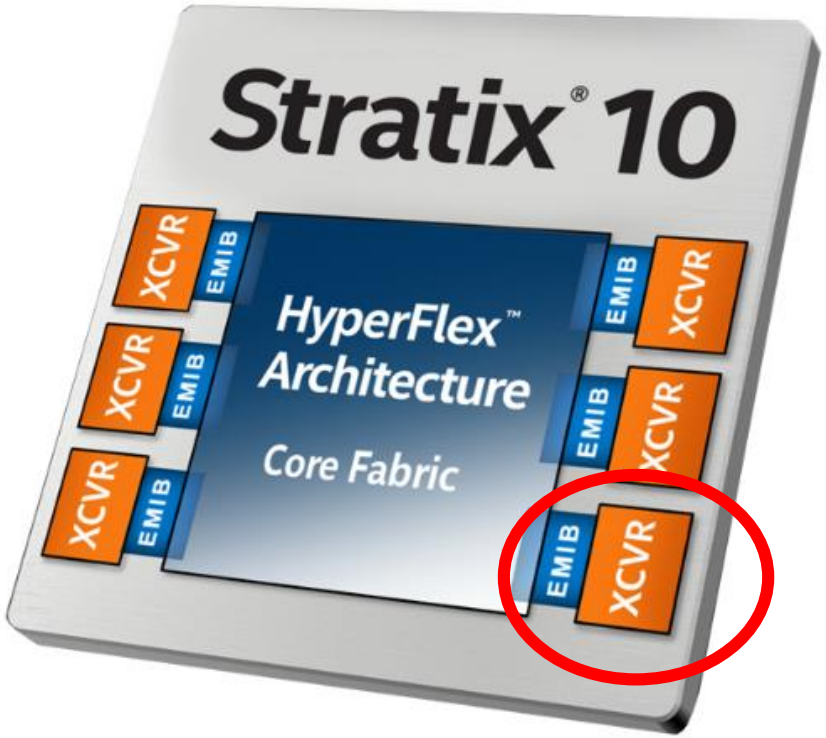
Software

FPGA

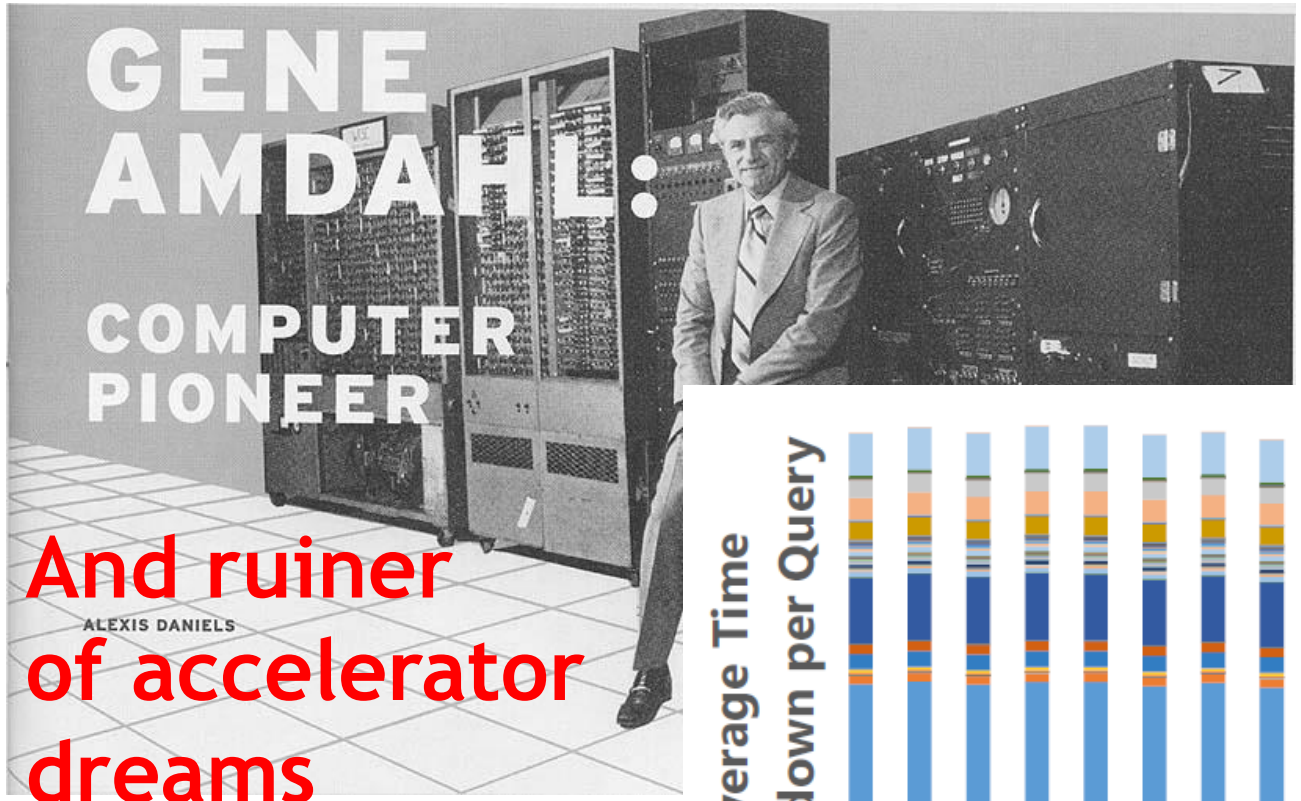
ASIC

Speeding ASIC Integration via Chiplets

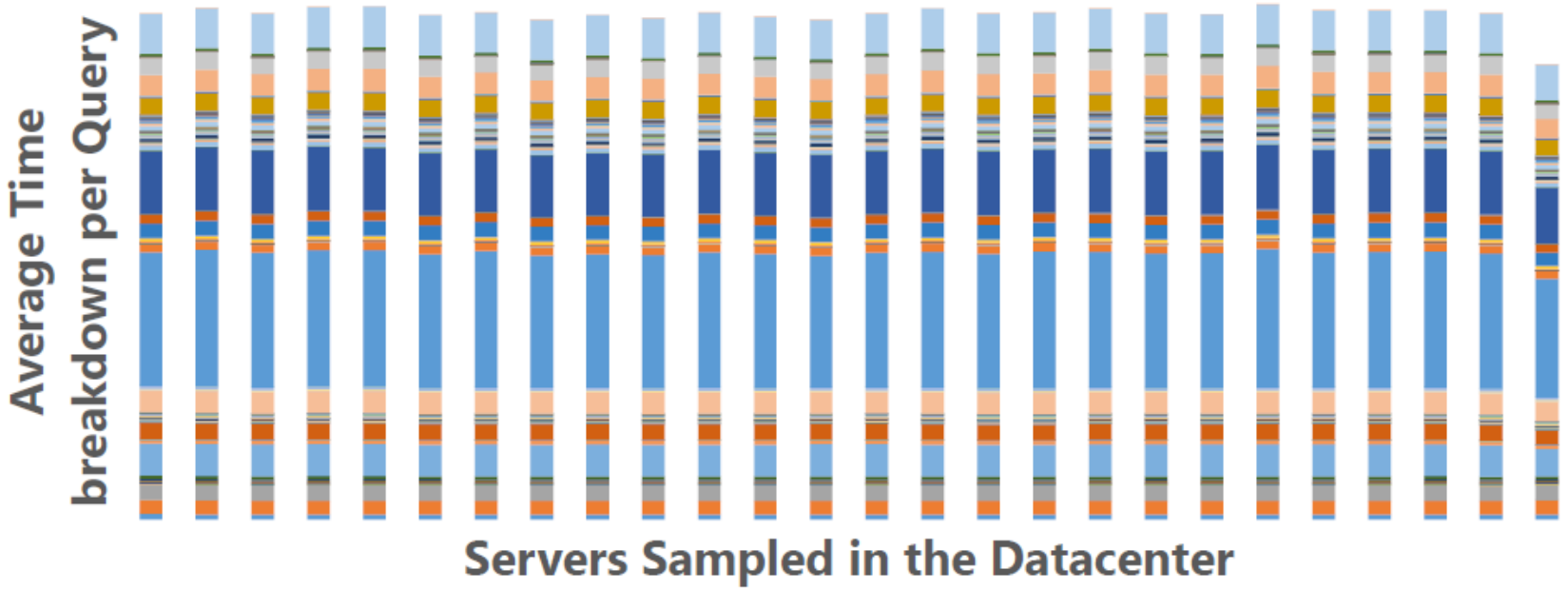
- FPGA provides common interfaces
  - DDR, PCIe, Ethernet, I2C can all be FPGA
- Focus on just the core value of your ASIC
- Use FPGA logic for common software API and “future proofing” interfaces
- Allows using separate process technology from FPGA
- Not necessarily specific to Intel



A Word of Caution – Amdahl's Law



- Deep Learning is generally only *part* of the full algorithm
- Still need general-purpose CPU platforms tightly integrated



Conclusions

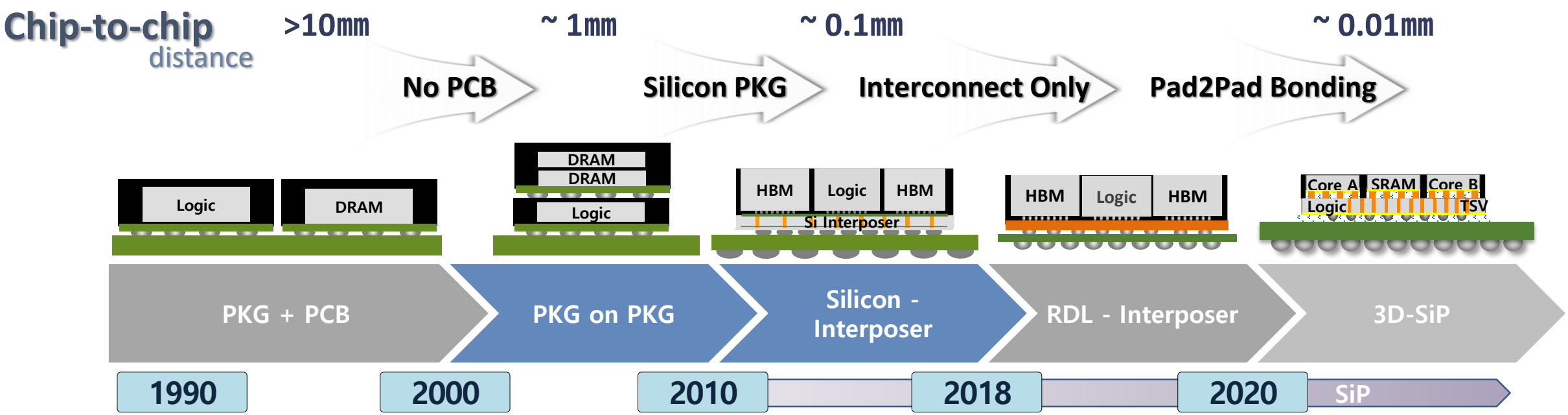
- Silicon customization is coming to the Cloud
- Deep Learning is pushing High-Performance Computing (HPC) from specialized clusters into the general-purpose fleet
- Network latency is critical… but so is cost
- Advanced packaging can greatly accelerate ASIC adoption in the cloud while still keeping pace with changes in AI/ML/Deep Learning

Speaker Bio: Dan Oh

- **EXPERIENCE**
  - Samsung Electronics, Package Development Team, Vice President (present)
  - Intel Corporation, Programmable Solution Division, SI/PI Architect (Aug. 2016)
  - Rambus Inc. Technical Director (June 2012)
- **EDUCATION**
  - Ph. D. Electrical Engineering, University of Illinois at Urbana-Champaign
- **Publication**
  - 66 patents and patent applications
  - Over 100 papers in IEEE journals and conferences
  - Book "High-speed Signaling: Jitter Modeling Analysis, and Budgeting."

Package Paradigm Shift

Integration solution leads technology

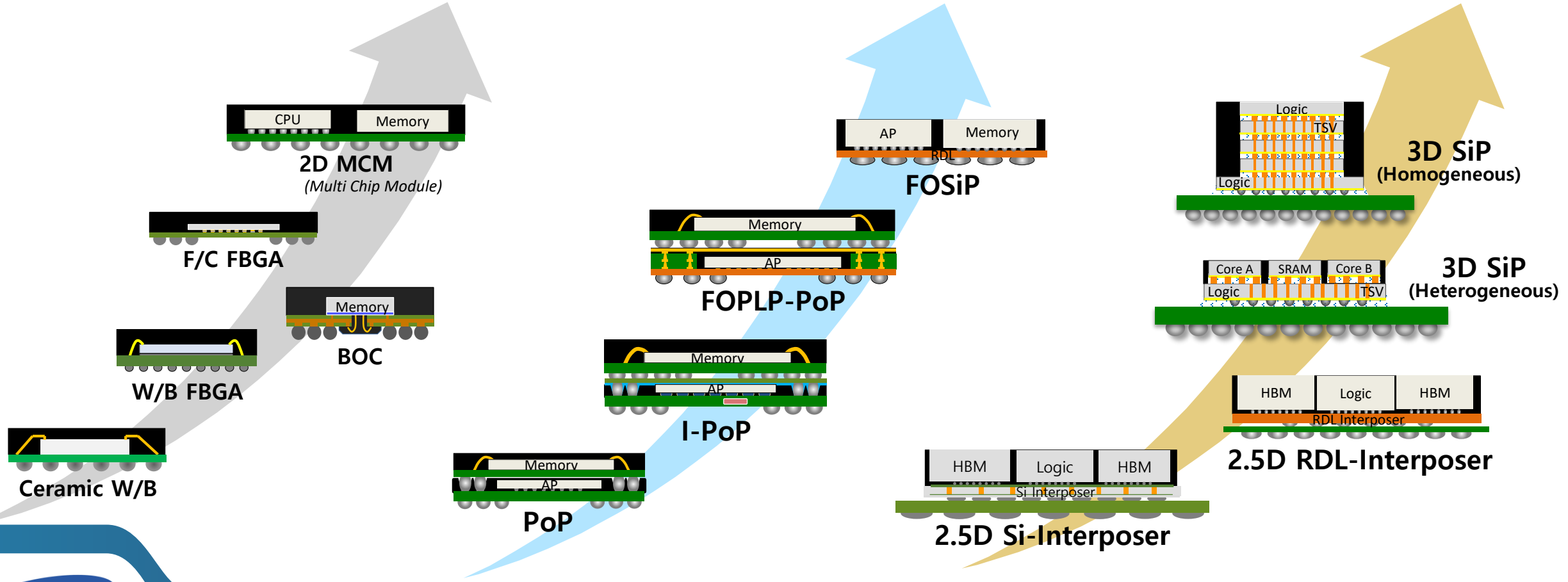


- Demands for **low-power** & **high-performance** accelerate chip-to-chip integration
- Integration technology continues to drive wider interconnects

Package Technology Evolution

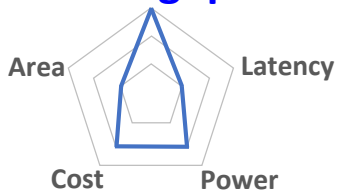
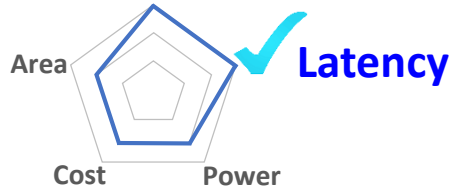
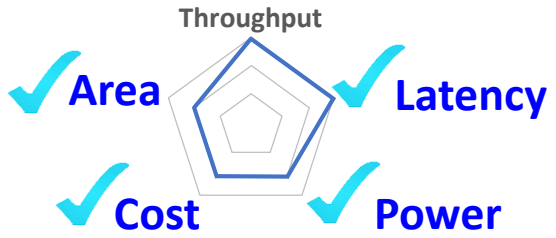
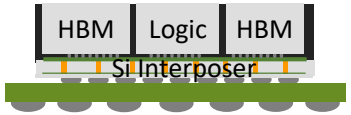
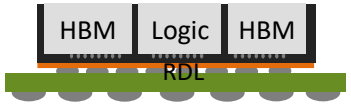
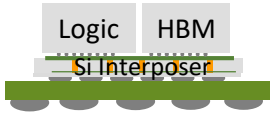
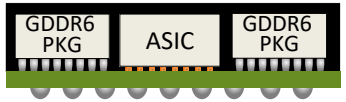
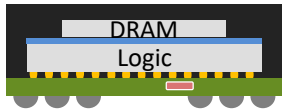
Multi-die integration plays a critical role in era of AI/HPC

Era of PC | **Era of Mobile** | **Era of AI/HPC**



Logic and Memory Integration for AI

Server Training (HBM), Inference (GDDR), Edge Inference (LPDDR, new DRAM?)

Application	Server		Edge
	Training	Inference	Inference
Key Index	✓ Throughput 	✓ Throughput 	
Demand	▪ High bandwidth ▪ High density	▪ High bandwidth ▪ Low latency	▪ Low latency, low power ▪ Small form-factor
Memory	HBM (4~6cube)	HBM (1cube), GDDR5/6	LPDDRx, new DRAM (?)
Package	**Large 2.5D interposer**  Si-interposer CoW  RDL interposer	**Small 2.5D interposer**  Si-interposer CoS  SiP Module	**Advanced SiP & Fan-out PKG**  Advanced SiP  Fan-out PKG

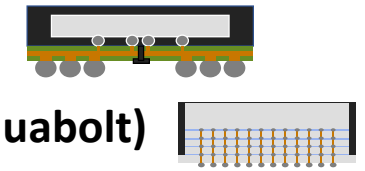
GDDR6 vs HBM

GDDR6 can be a good replacement of HBM for server inference, HPC, block chain, and automotive applications

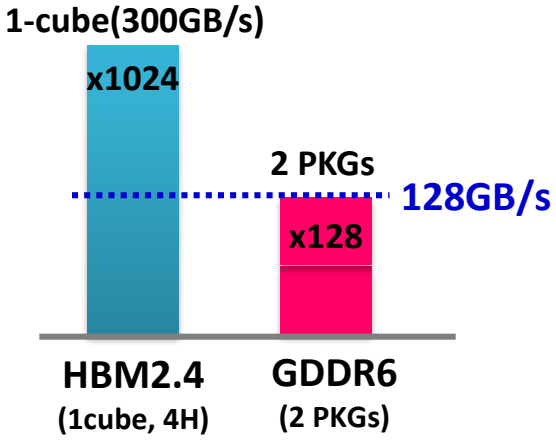
- ✓ Memory bandwidth assumption: **128GB/s**
  - HBM(Aquabolt, 4H): 1cube
  - GDDR6: 2ea

: GDDR6

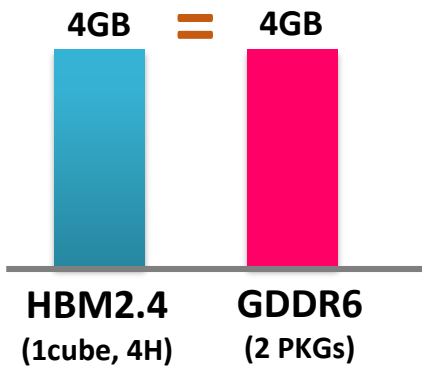
: HBM(Aquabolt)



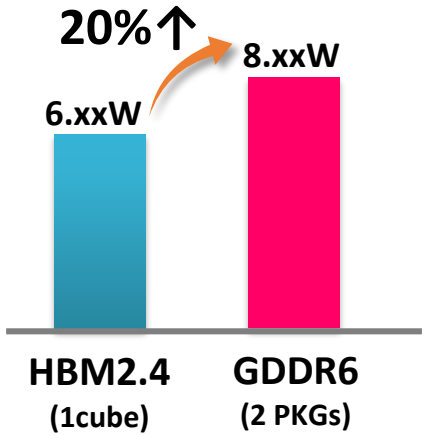
Bandwidth



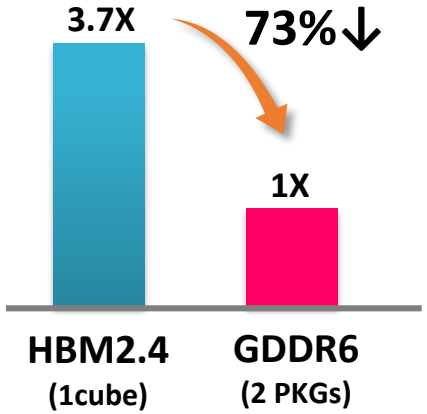
Density



Power



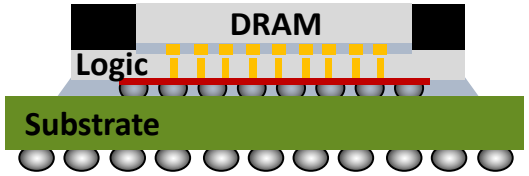
Cost



Potential Memory Architectures

3D memory integration addresses
Power, Throughput, Latency and area except Cost

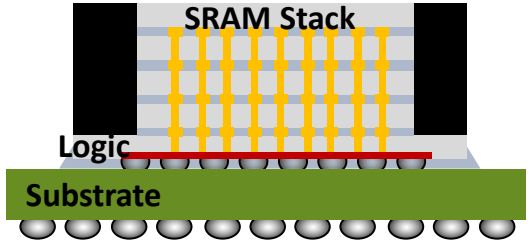
Mobile AI



High Bandwidth Memory (Non-HBM, x128~256)

- DRAM
- Mono die
- Bandwidth: 1x

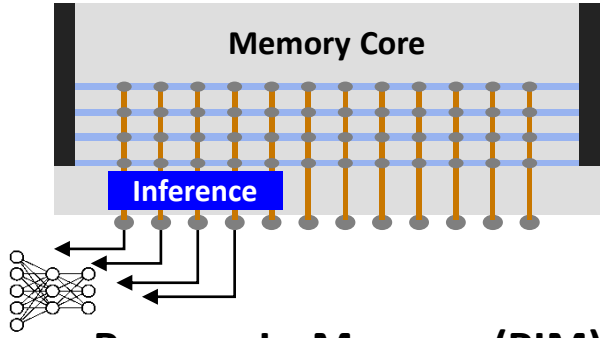
Server Inference



SRAM Stacking

- SRAM
- 2~4 stacked dies
- Bandwidth: 2x ~ 4x

Server Training

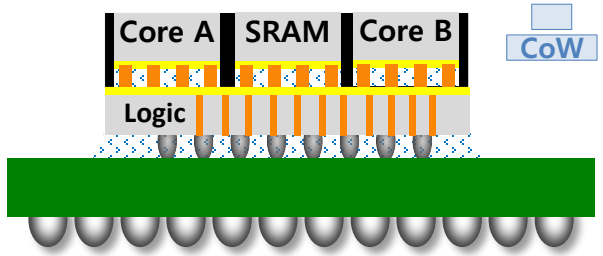


Process-In-Memory (PIM)

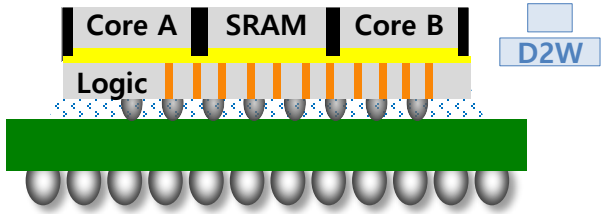
- HBM-base
- 2~8 stacked die
- Bandwidth: 2x ~ 16x

Future 3D System Integration Schemes

Heterogeneous Integration



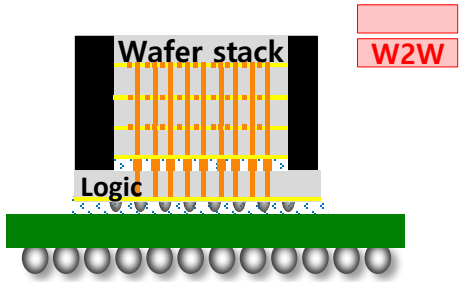
CoW



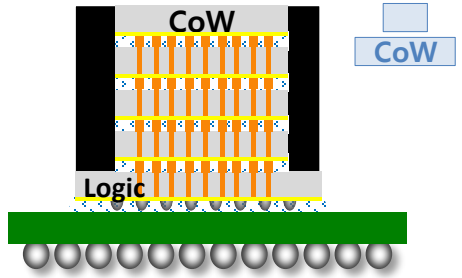
D2W

- ✓ High-end System Integration (different node, different die, PMIC, etc)
- ✓ Die partitioning

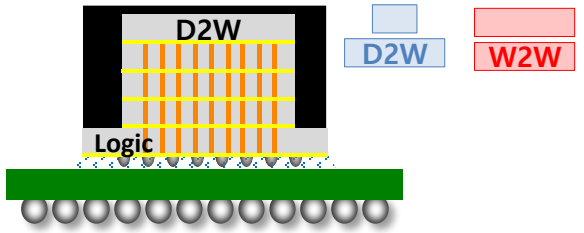
Homogeneous Integration



W2W



CoW



D2W + W2W

- ✓ New memory
- ✓ Same die multi-stack

W2W : High Throughput

CoW : High Yield & High Throughput

Conclusions

- **Package integration continues to drive new computing architectures**
  - From 1D to current 2.5D, and moving onto 3D…
- **Silicon interposer and HBM serve current AI training needs**
  - 3D integration serves further training requirements
- **Inference may require new memory solutions**
  - Small high bandwidth memory or
  - Low latency SRAM devices

Speaker Bio: Madhavan Swaminathan

Professor Madhavan Swaminathan is the John Pippin Chair in Microsystems Packaging and Electromagnetics in the School of Electrical and Computer Engineering and Director, Center for Co-Design of Chip, Package, System (C3PS), Georgia Tech. He is the author of 450+ refereed technical publications, holds 30 patents, primary author and co-editor of 3 books, founder and co-founder of two start-up companies and founder of the IEEE Conference Electrical Design of Advanced Packaging and Systems (EDAPS) sponsored by EPS.

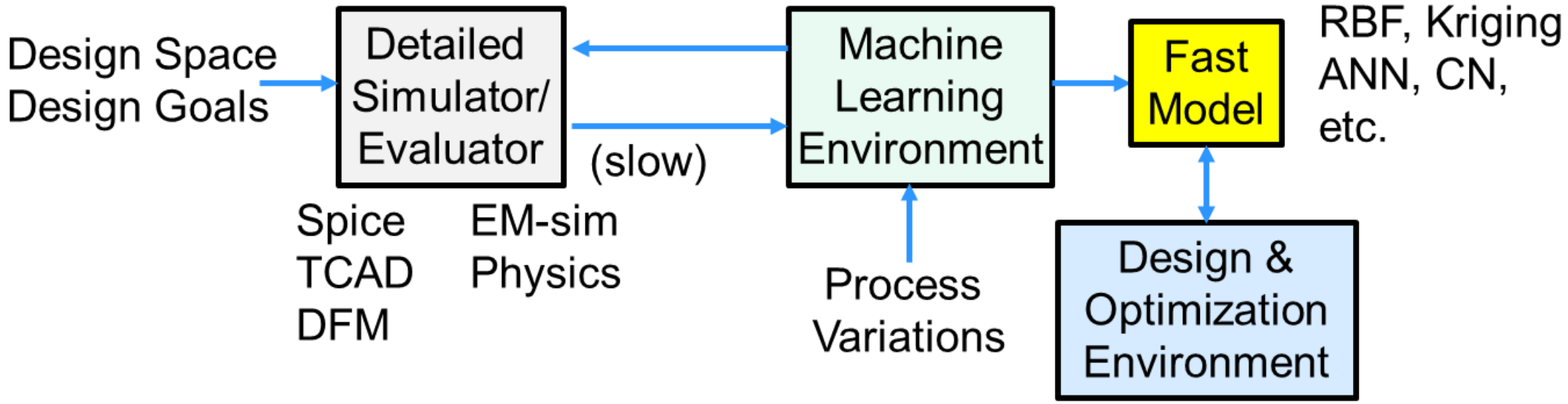
Limits of Modern EDA

- The need for design “re-spins” has not been eliminated
- Many of the observed failures during qualification testing are the direct result of an insufficient modeling capability
  - Sources of such failures include mistuned analog circuits, signal timing errors, reliability problems, and crosstalk [*]
- Simulation-based design optimization has had only limited success
  - Simulation “in-the-design-loop” often too slow and leads to impractical designs

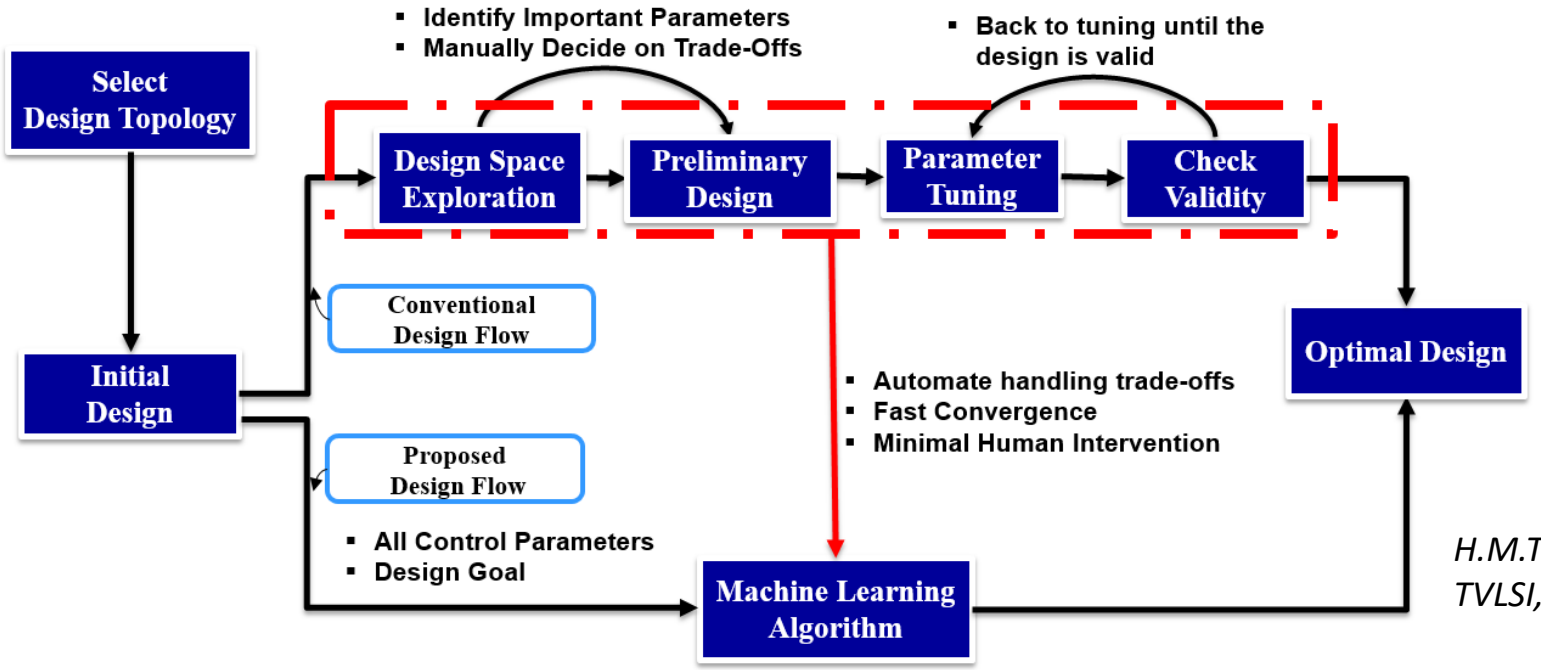
[*] Harry Foster, “2012 Wilson Research Group Functional Verification Study,” http://www.mentor.com/products/fv/multimedia/the-2012-wilson-research-group-functional-verification-studyview

Model-Based Design Paradigm

- Fast-to-evaluate “learned” model replaces detailed slow model in design, and design optimization/feasibility problems



Optimization using ML

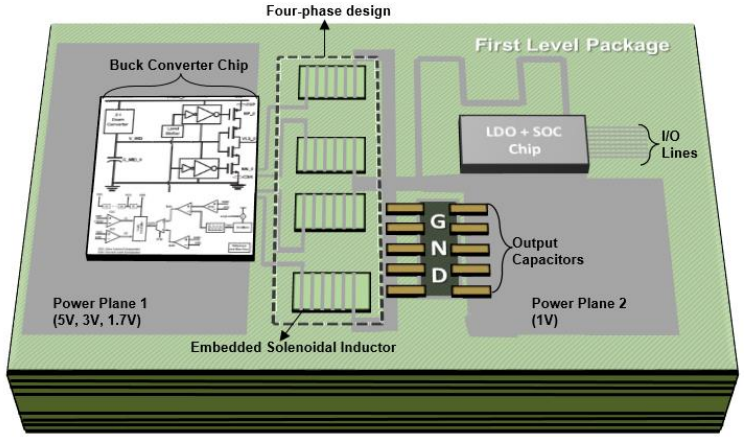


H.M.Torun, M.Swaminathan et al, TVLSI, 2018

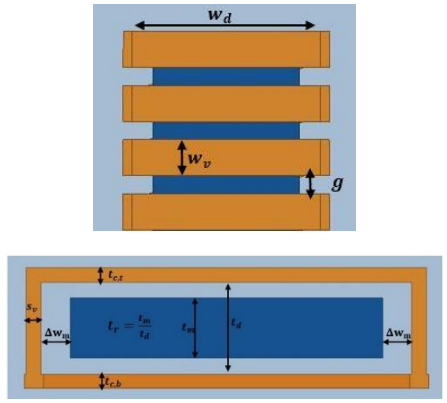
- Main objective is to use ML based optimization to automate the design cycle and minimize human intervention on optimization and tuning of control parameters of integrated systems.
- Active Learning:
  - uses zero training data.
  - ensures convergence to global optima while minimizing the required CPU time.
    - capability of handling very large number of parameters (50+).

Integrated Voltage Regulator Optimization

Overall IVR Architecture



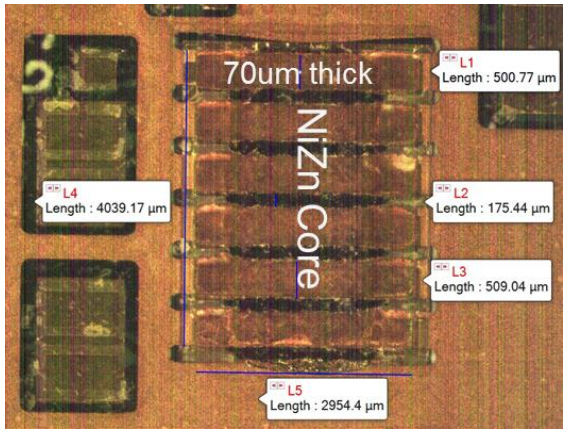
Embedded Solenoidal Inductor



	Hand Tuned	Non-Linear	GP-UCB	IMGPO	TSBO
Inductor Area	11.3 mm^2 (+56.1%)	25.19 mm^2 (+79.6%)	5.18 mm^2 (%0.4)	6.64 mm^2 (%28.1)	**5.16 mm^2**
Peak Efficiency	79.4%	78.6%	84.9%	84.4%	**85.1%**
CPU Time	N/A	>185 min (+72.9%)	117.33 min (+57.4 %)	115.6 min (+56.7 %)	**50.1 min**

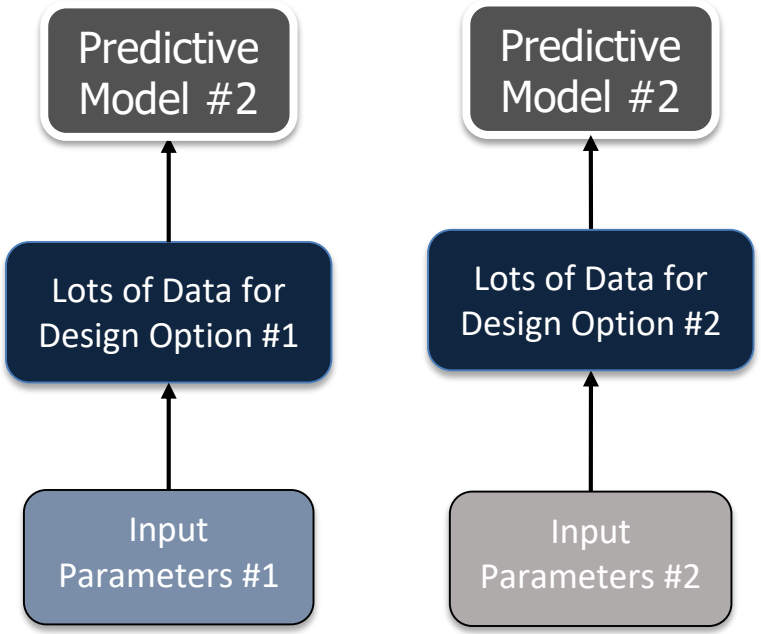
- Integrated Voltage Regulators are used to increase efficiency and conserve power in microprocessors (Ex: Intel Gen 4)
- Objective is to maximize IVR efficiency while minimizing inductor area
- IVR efficiency is affected by inductor and buck converter.
- Assuming LDO, PDN and LOAD is fixed.
- Solenoidal Inductors with magnetic cores are used
- Multiple trade-offs: ESR, DC resistance, inductance, lateral area
- Tune inductor control parameters to maximize efficiency (8 – 10 dimensions)

Embedded Solenoidal Inductor

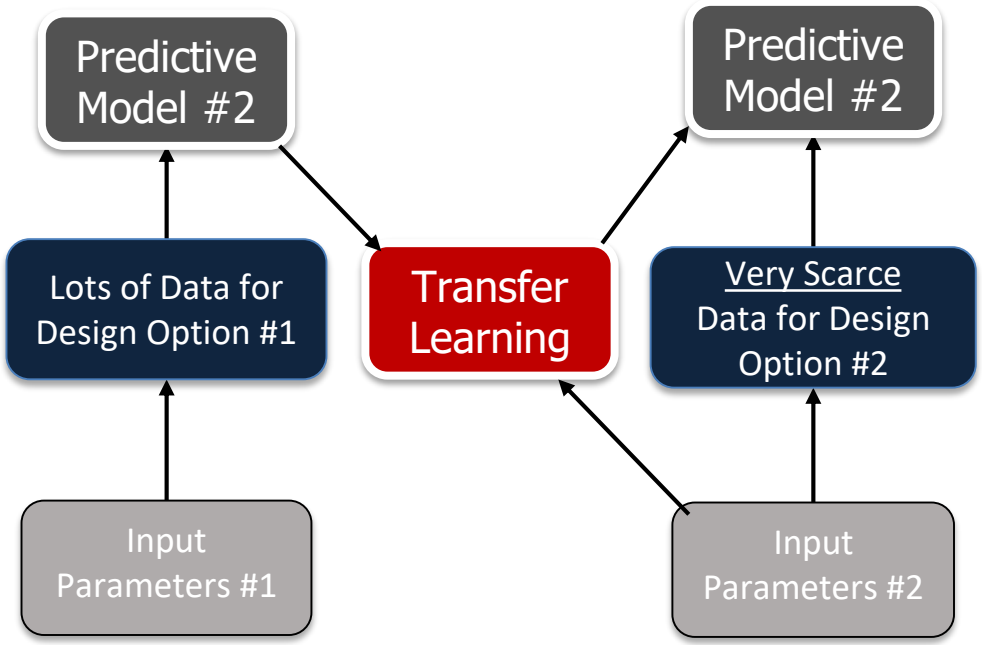


Hand Tuned: S. Mueller et al., ECTC'16.
Optimized:H. M. Torun et al., TVLSI '18

Design Space Exploration (using Transfer Learning)



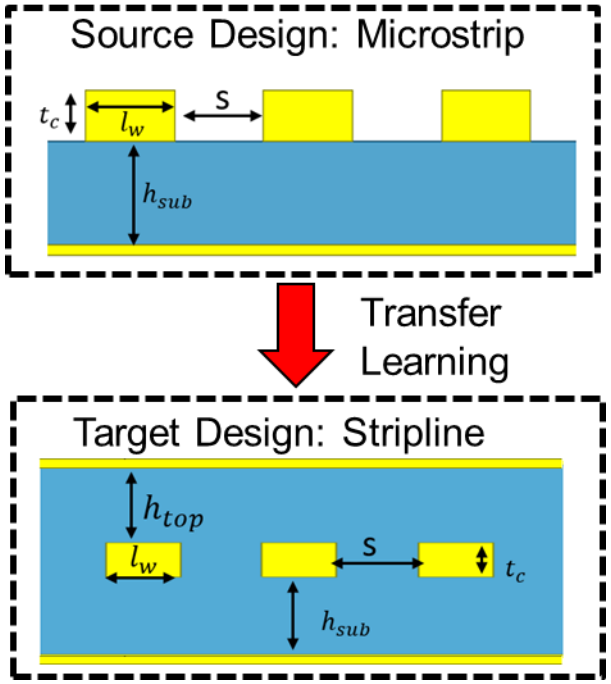
Conventional Method
(Different Model for Different Designs)

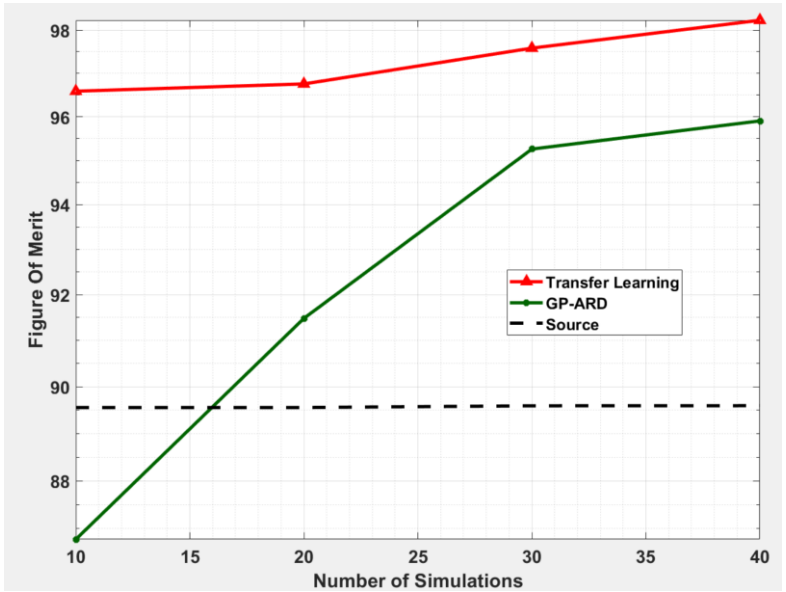


Proposed Method
(Re-use previous model to derive new model)

- Design space exploration involves developing models for many different topologies.
  - Ex: Sigle-ended vs differential signaling, shielded vs unshielded signal lines etc.
- Different topologies can share information that can be exploited using transfer learning to significantly reduce CPU time and effort to derive new models.

Model transfer from Microstrip to Stripline





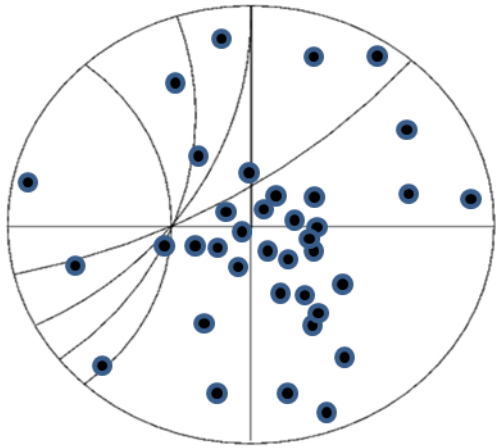
Parameters		Min	Max	
$\mathbf{l_w}$	Trace Width	$0.4\ \mu m$	$3\ \mu m$	Shared Parameters
$\mathbf{t_c}$	Trace Thickness	$0.4\ \mu m$	$3\ \mu m$	
$\mathbf{s}$	Spacing	$0.4\ \mu m$	$3\ \mu m$	
$\mathbf{h_{sub}}$	Substrate Height	$1\ \mu m$	$5\ \mu m$	
$\mathbf{f}$	Frequency	$0.1\ GHz$	$20\ GHz$	
$\mathbf{h_{top}}$	Substrate Height Offset	$1\ \mu m$	$5\ \mu m$	Additional Parameter for Stripline

- The goal is derive a model to predict frequency dependent RLGC parameters for both microstrip and stripline structures.
- Model for microstrip line has already been developed and validated to have high accuracy (assumption – prior data)
- The model for microstrip is then re-used using transfer learning to derive a new model for stripline.
- Preliminary results show transfer learning approach significantly reduces CPU time to derive the model for stripline compared to different models for each structure.

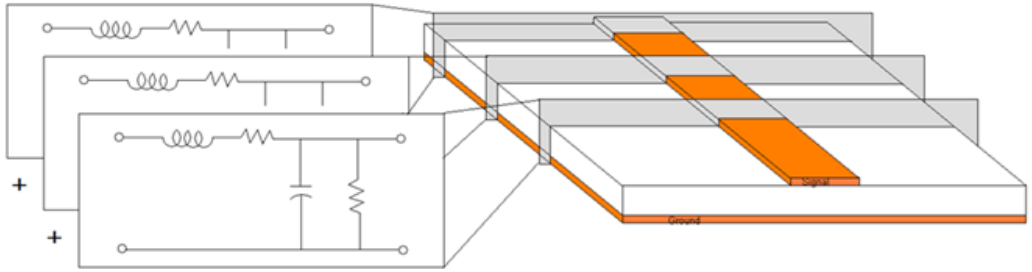
Design Space Exploration using ML

High-Dimensional Data

Mapping to Low Dimensional Space



- ❑ Inference
- ❑ Optimization
- ❑ Hardware Security
- ❑ Sensitivity
- ❑ Yield
- ❑ Manufacturability



And to you Skeptics….

Machine Learning can help

Eliminate the frustrations of Design and Simulation by AUGMENTING the engineer but never REPLACING the engineer. Engineers are the thinkers! Computers are the doers! Machine Learning is the enabler!

Q&A