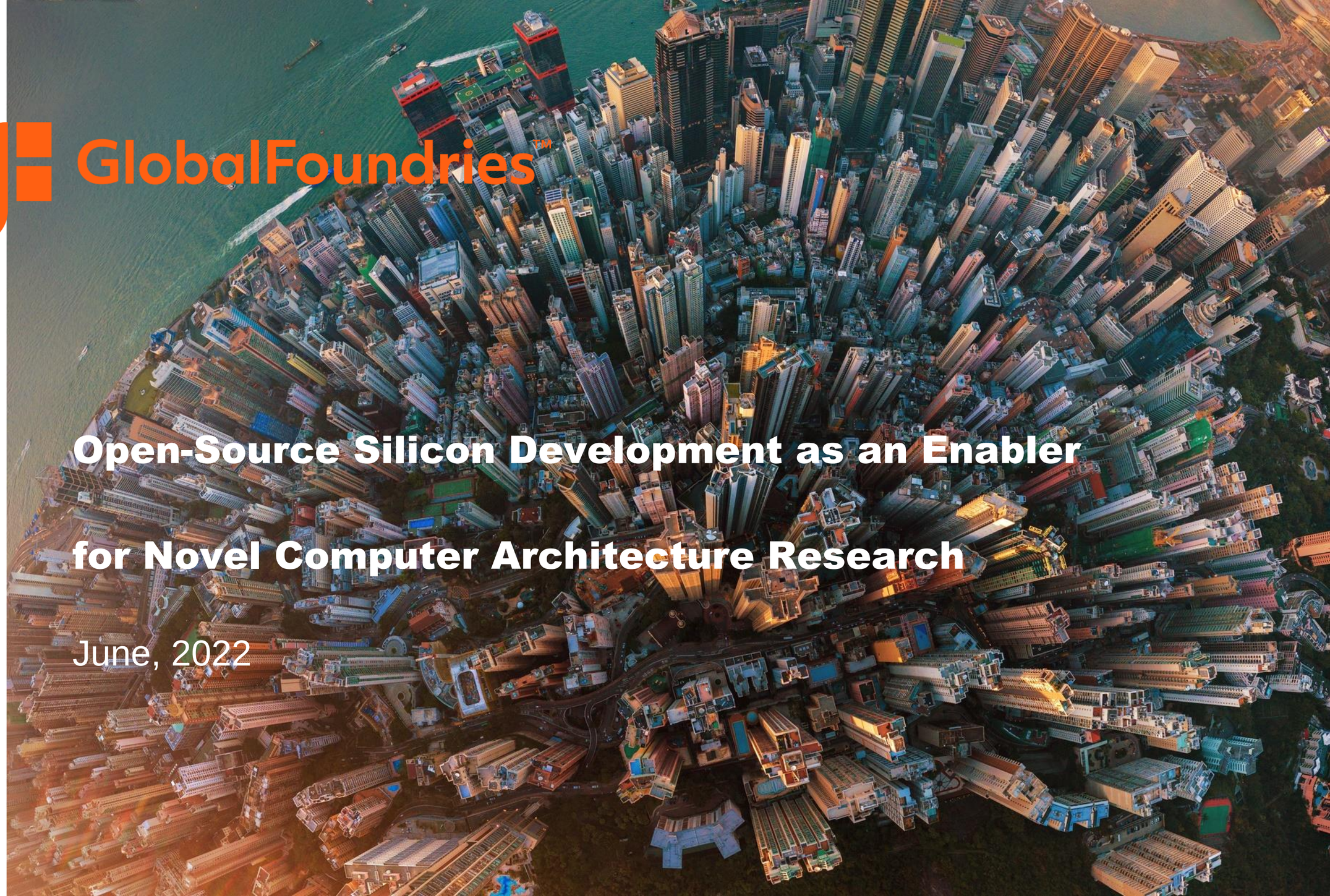


GlobalFoundries™

Open-Source Silicon Development as an Enabler for Novel Computer Architecture Research

June, 2022

Open-Source Silicon Development as an Enabler for Novel Computer Architecture Research

Shashank Nemawarkar
Director, Computer System Architectures,
GlobalFoundries

Open-Source Computer Architecture Research (OSCAR) Workshop
Saturday, June 18, 2022 - New York City (co-located with ISCA 2022)

Agenda

1. Motivation
2. Open-source silicon
3. Platforms and Design IPs
4. Tool enablement
5. Technology enablement
6. Bringing all together

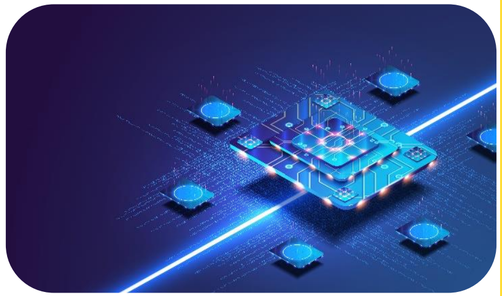
Why open-source silicon for computer architecture research?

- Domain gains from architecture through technology are bounded
- Only a few teams can afford internal development on vertical optimizations across domains
- Results from known, reproducible platforms provide clear gains from research ideas
- Silicon proven ideas get faster adoption in research and industry
- Open-source efforts democratize architecture research

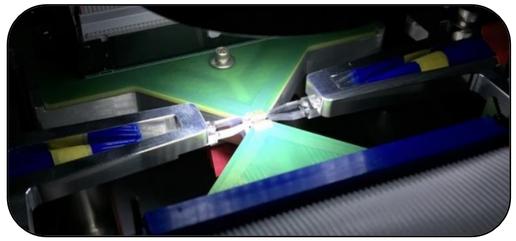
Motivation

- Exponential growth needed for computation:
  - Performance: Machine learning needs up to 3 orders increase every 3-5 years!
  - Storage: 20-30% growth per year*
  - Efficiency: Power/energy/cost be better than other metrics!
  - Domain proliferation:
    - Data center through edge, IoT, RF, Displays, Auto, and more
- Technology node gain: 20-40% power, 10-20% performance, and slowing
- Computer architecture ties applications to realizable systems
  - Research needed for exponential “efficiency” gains

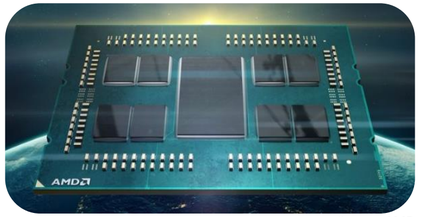
Quantum Photonic Computing



Photonic Super Computing

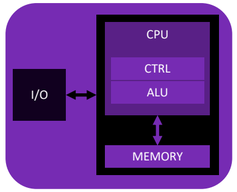


Source: Lightmatter

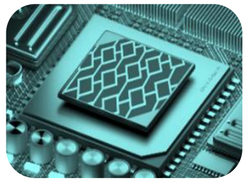


Source: AMD

- 2.5/3D
- Chiplets



Von Neumann



AI TPU
w/ systolic array

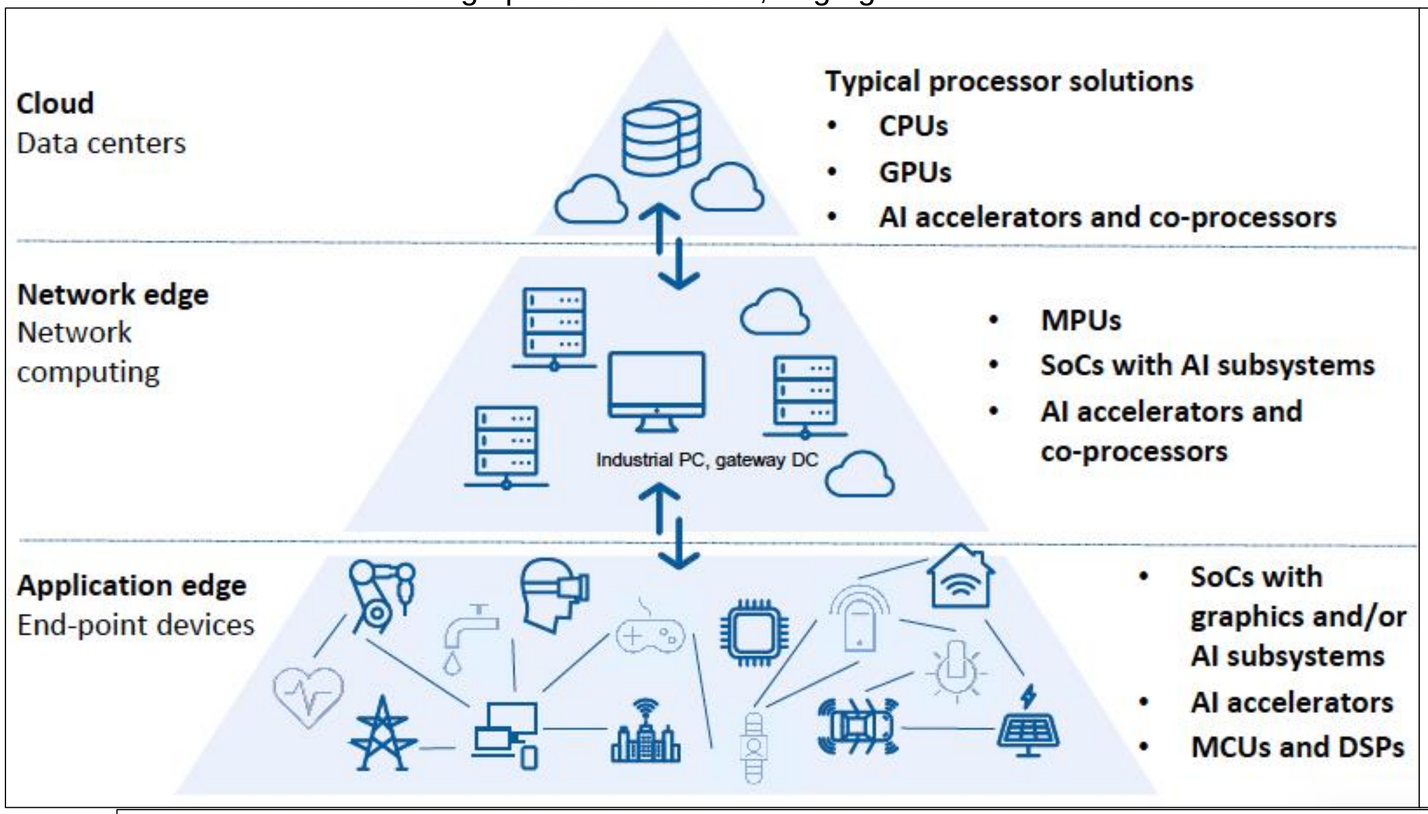
In-memory computing

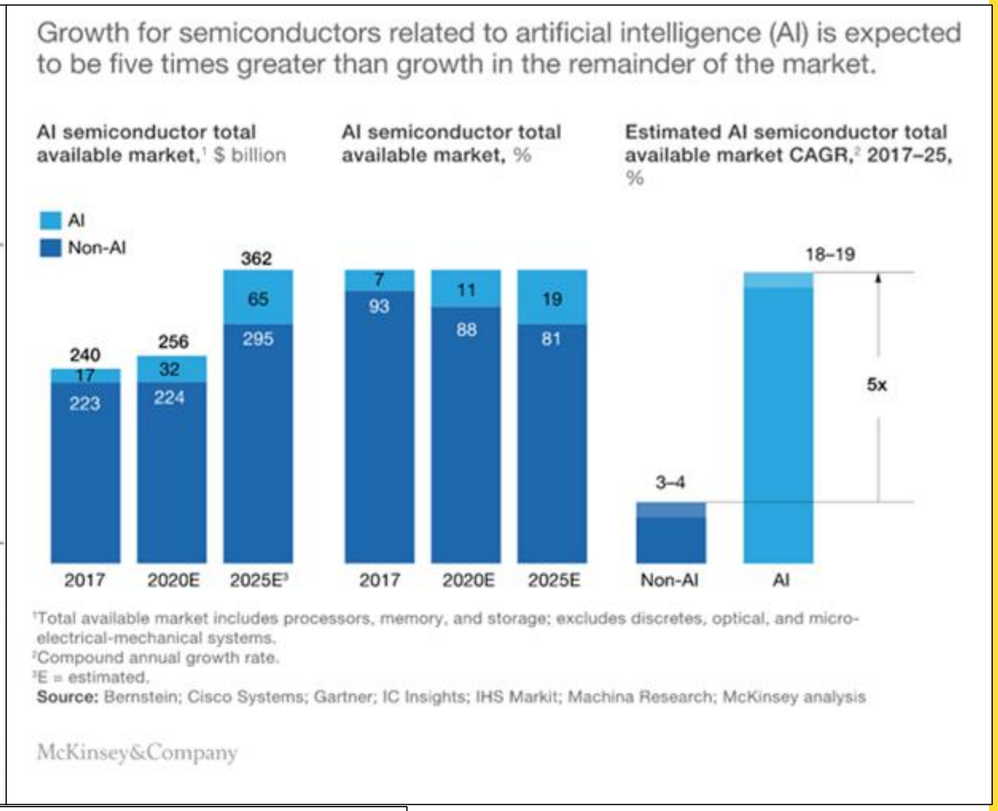
Architecture needs to exploit insights from algorithms through technologies

Storage: Statia 2022.

Compute is pervasive

- AI getting across domains
- CPU, GPUs provide limited upside
  - Decision making– serial execution, small gains
  - Number crunching– parallel execution, huge gains
- Address application diversity
- More computation in less (area, power, cost)
- Compute what is important (accuracy, approximate, probabilistic…)





Compute at the most efficient place

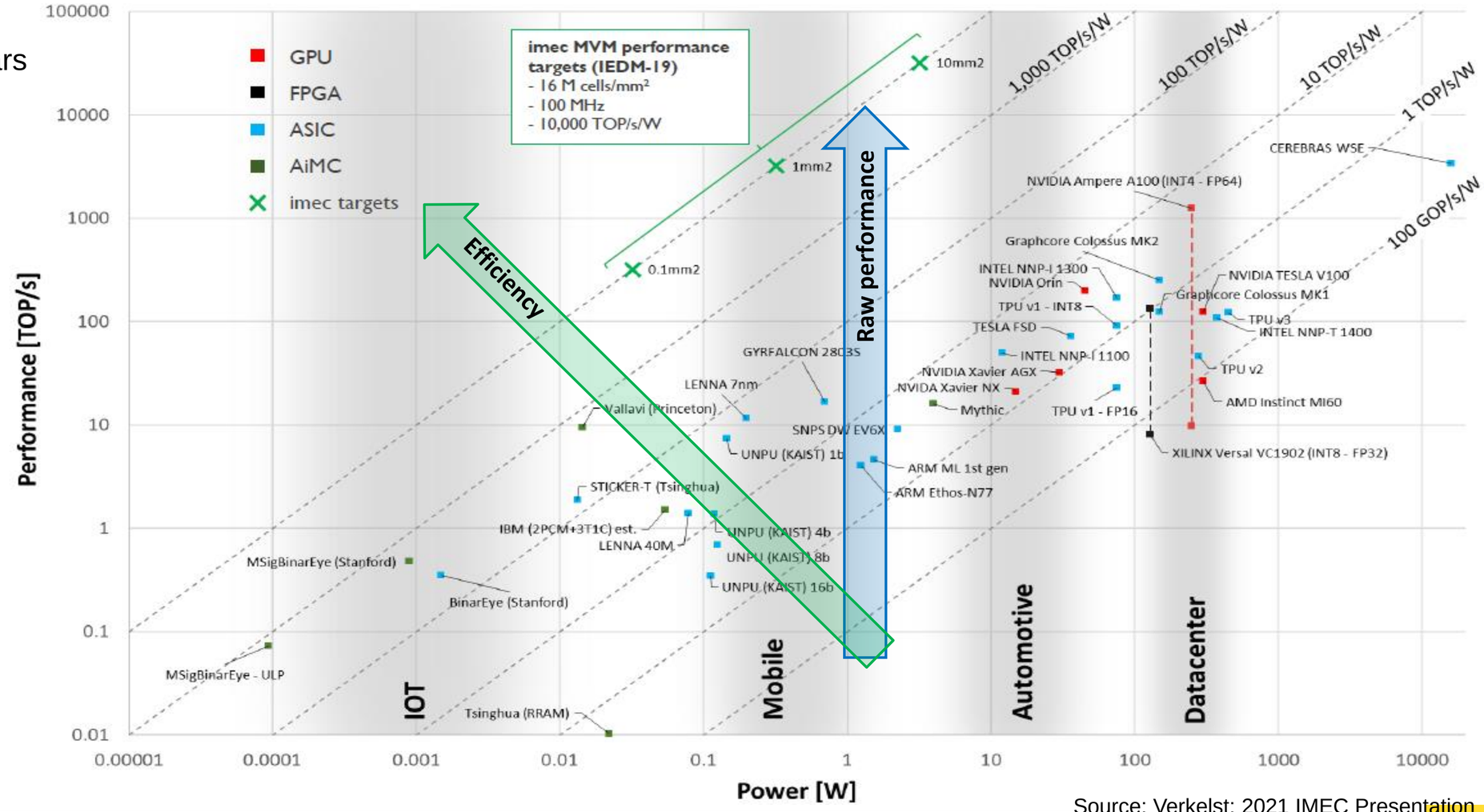
Deloitte Analytics, McKinsey

Machine Learning Efficiency Trends

Trends:
- 10-1000x improvements in 3-5 years
- High computation and communication demands
- Regular patterns, but traditional xPU/memory hierarchies fail

Perf/Power:
- Analog: 10 TOPS/W => 0.1pJ/OP
- Digital: ~6 TOPS/W => 0.16pJ/OP

Targets:
- IOT: 500-10000 TOPS/W
- Mobile: 100 TOPS/W
- Automotive: 10 TOPS/W
- Data Center: 10 TOPS/W



Source: Verkelst: 2021 IMEC Presentation

ML and IoT are breaking the walls between algorithms, architecture, design, and technology

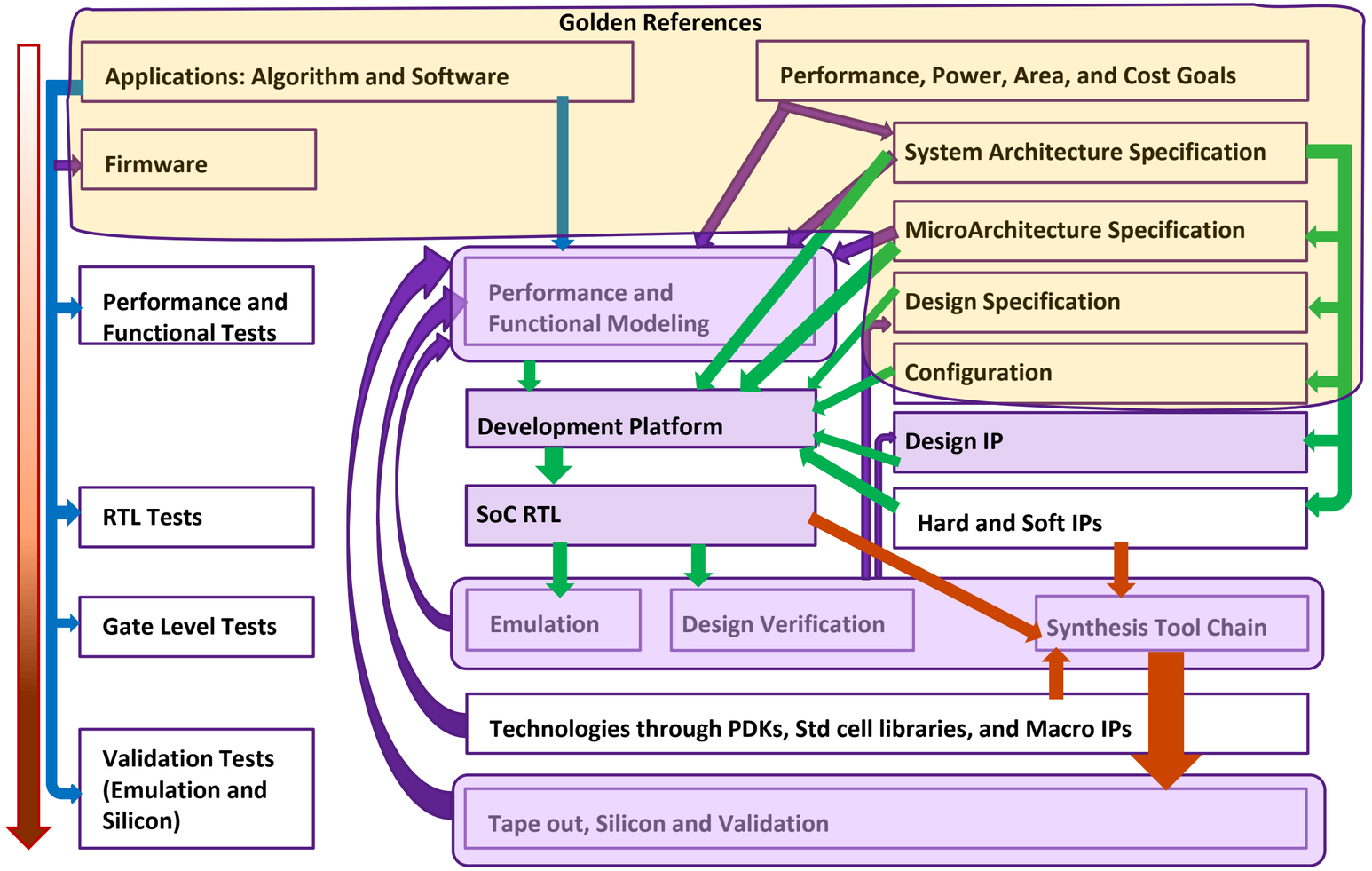
Agenda

1. Motivation
2. **Open-source silicon**
3. Platforms and Design IPs
4. Tool enablement
5. Technology enablement
6. Bringing all together

Open-source silicon development flow

Application through Silicon Validation

- Maximum development efforts:
  - SW Stack and tests
  - Purple implementation blocks
    - Performance
    - Design
    - Test
    - Synthesis to Silicon
- Architecture:
  - Glue for Algorithm through technology
  - Functional/performance
  - Reset/clock/debug



Frameworks, platforms, configurability and IP availability speeds up the design efforts

Personal Journey in Hardware

Shashank

Late 80s

Microelectronics and Systolic Array Architectures

Early Grad Studies

Internal platforms

Mid 90s

General Purpose Compute: Multithreaded Architectures, Design, Modeling

Doctoral Studies

Early Rapid Prototyping, basic simulators

Late 90s -Mid 00s

Processor and Server Architectures and Design

IBM & Startups

Rapid Prototyping, licensed IPs, and Open-source SW

Late 00s -Mid 10s

Network processors, SSD Controllers, uarch and Design

LSI/Avago/Seagate

Custom emulators, Early open-source and proprietary IPs

Late 10s

Mobile processor uarch and performance

Samsung

Commercial emulators, Open-source platforms, Open-source simulators

2020s

Compute architectures, designs, PPAC, methodologies, and technology guidance

GlobalFoundries

Full open-source silicon, open-source SW stack

HW Industry

Accelerators, chiplet, heterogeneous architectures

Full custom designs: high performance processors, low power applications, high efficiency

Full custom chips: Research

Open Source and Proprietary IP based Custom designs

Discrete Off-The-Shelf Chips based ssystems

Rapid prototyping with FPGAs and emulators

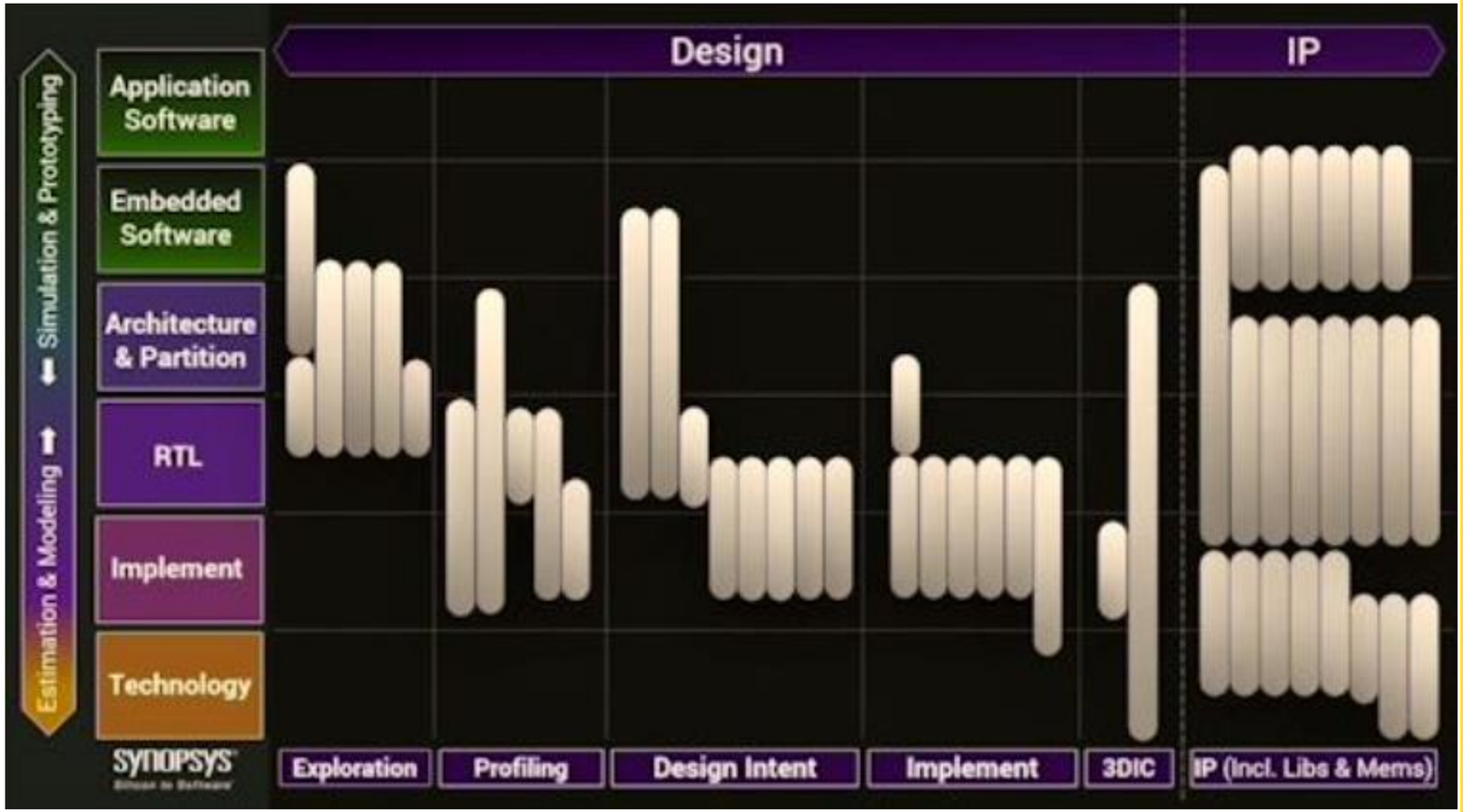
Simulators

Analytical Models

Architecture Research and IP Design

Hot Chips 2021: Synopsys DSO.ai presentation, Aart de Geus

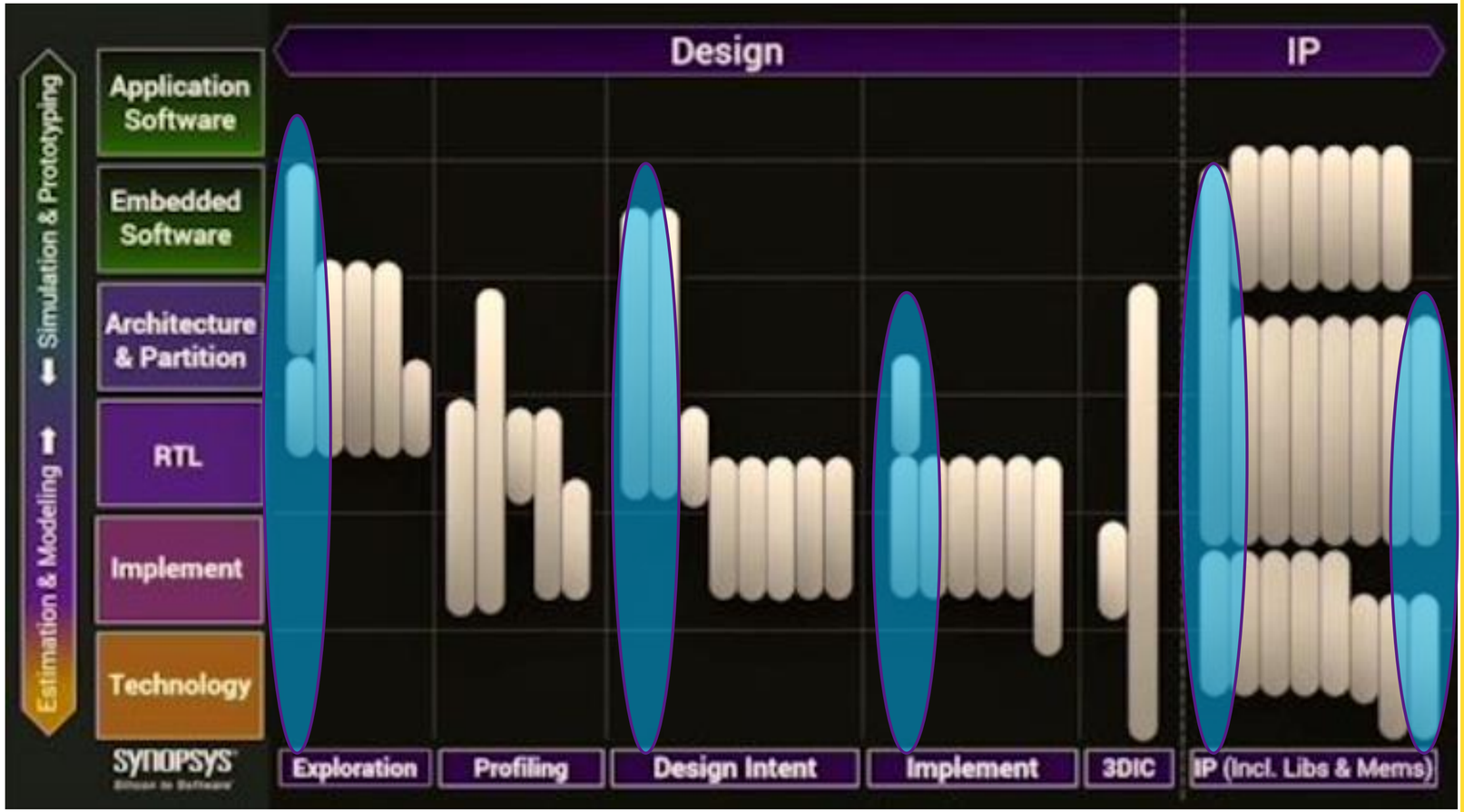
- Traditional design optimization
  - Explore SW to architecture mapping
  - Profile uarch to RTL
  - Create RTL design and tests
  - Synthesize and optimize design
  - (New) Heterogenous partitioning
  - Optimize using std libs and memories
- AI improves each step
  - Google chip design project to improve floorplan and synthesis
  - Whole methodology is ripe for innovation!



Architecture Research and IP Design

Full stack optimization from application through technology

- Technology visibility at application level provides non-organic gains
  - Newer devices, circuits and IPs
  - Newer architectures
  - Newer application of HW
  - Algorithmic impact
  - Example: NVMs, Low voltage, Low temperature logic.
- Reduced pressure on technology node improvement
- Increased state space for Computer Architecture and Algorithm Research



Aart de Geus, Synopsys, Hot Chips 2021

Agenda

1. Motivation
2. Open-source silicon
3. **Platforms and Design IPs**
4. Tool enablement
5. Technology enablement
6. Bringing all together

Open-source platforms

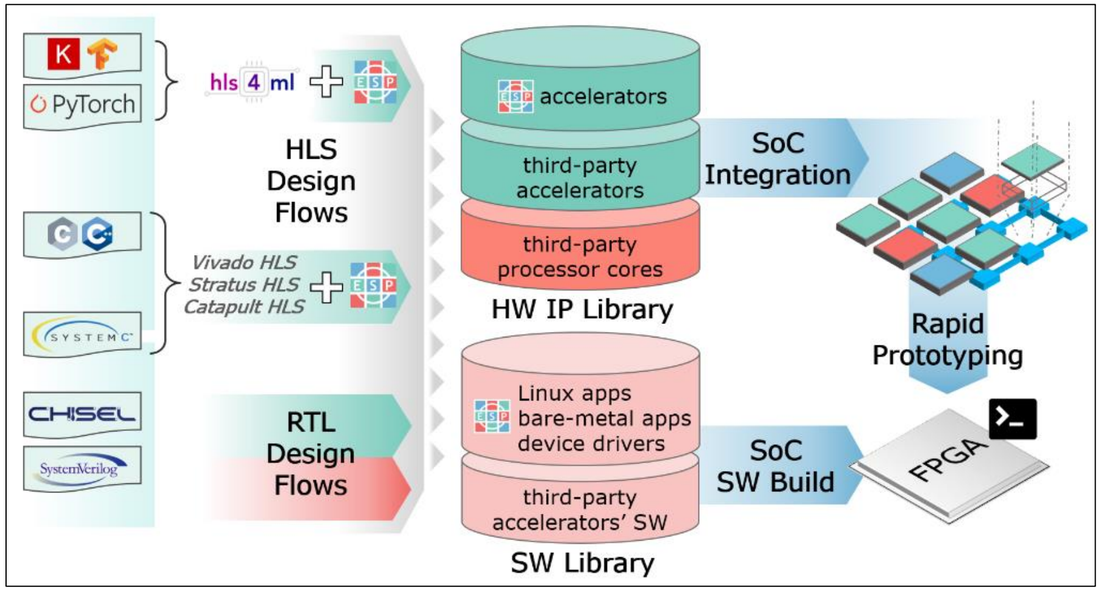
Multiple platforms available to enable algorithm to design

- Platforms enable IPs and ecosystem usage:
  - Example Columbia ESP, Simon Fraser NEEDLE
- Ability to stitch a System/SoC/IP using
  - Internal designs, IPs or external IPs
  - Internal or external interconnect IP
  - Foundational IP and interface IP
- Allows development of custom design with other vendor IPs, and basic infrastructure
  - Multiple starting points: high level language and synthesis or configuration scripts to generate RTL
- RISC-V Ecosystem:
  - Open source: processor, accelerator, NoC, external interface IPs from one or more sources
- ARM Ecosystem:
  - Proprietary, but can be licensed
  - Extensively used in the industry

Embedded Scalable Platform (ESP)

Columbia University, with academic and industry partners

- High level Synthesis for algorithms to RTL design
- Rapid prototyping: performance evaluation
- Area, power and energy need Silicon Validation!
- Std cell libraries, memories and technology for full stack optimization

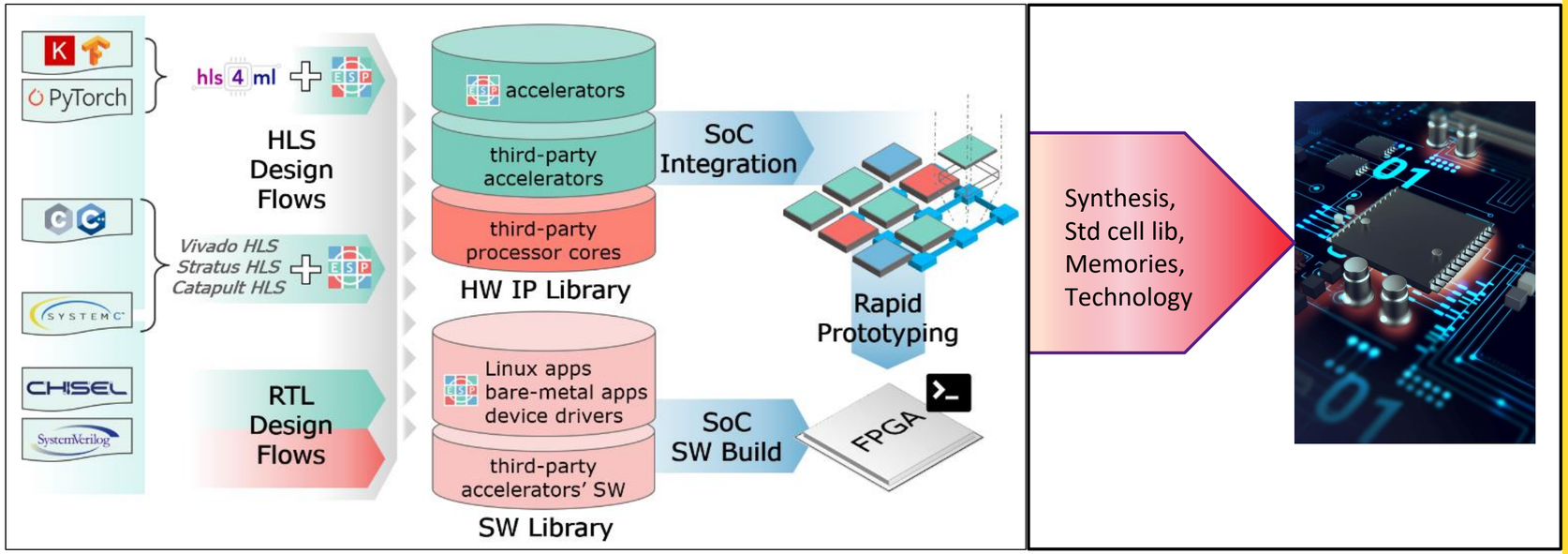


Frameworks, platforms, configurability and IP availability speeds up the design efforts

Embedded Scalable Platform (ESP)

Columbia University, with academic and industry partners

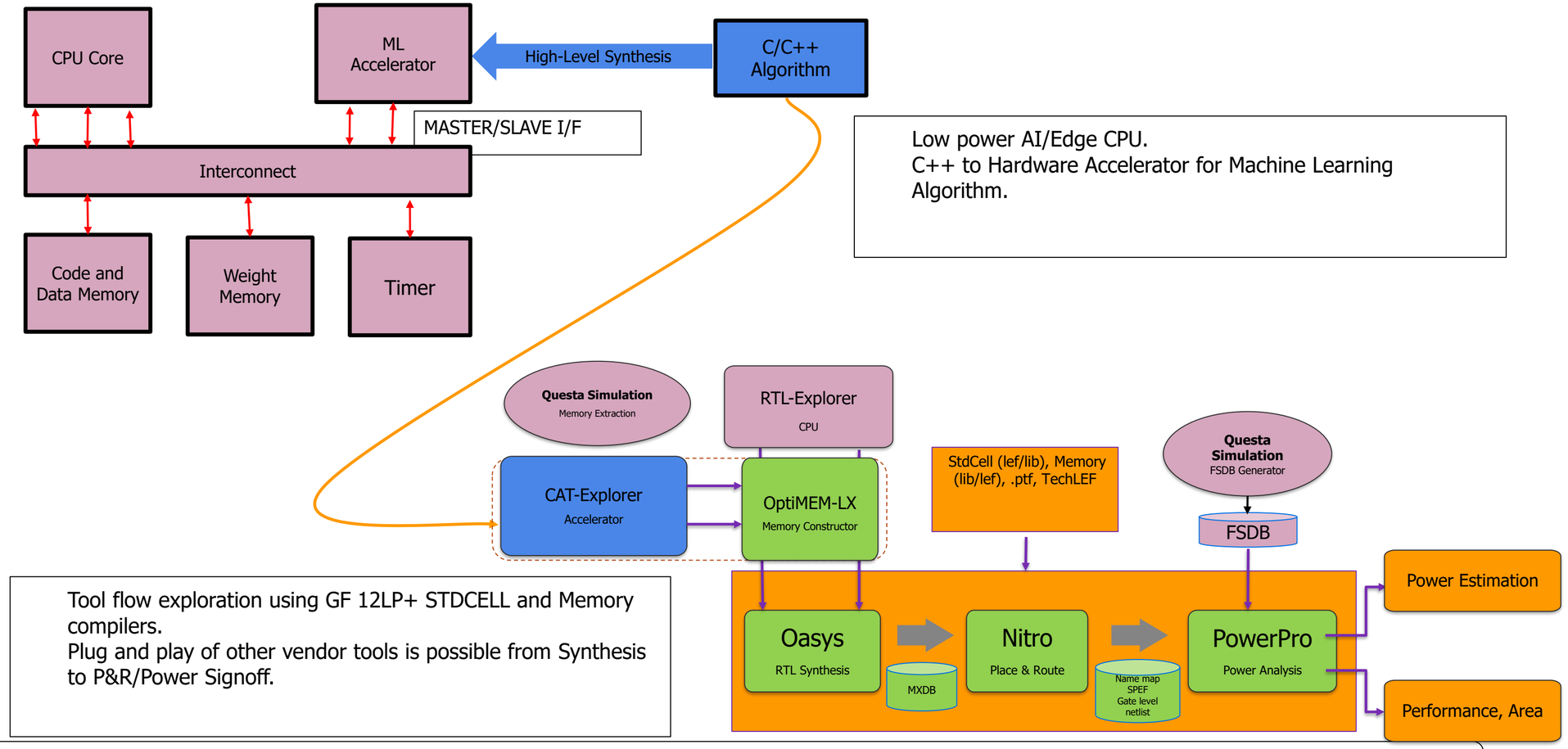
- High level Synthesis for algorithms to RTL design
- Rapid prototyping: performance evaluation
- Area, power and energy need Silicon Validation!
- Std cell libraries, memories and technology for full stack optimization
- Silicon to realize true benefits of the differentiated architecture



Frameworks, platforms, configurability and IP availability speeds up the design efforts

ESP with Siemens HLS and GF Technology

Algorithm to Silicon



Exemplary Framework enabling with Open-Source and Proprietary IPs and Flows

Open-source IPs

Rich availability of IPs leads to greater use of platforms

- Design IPs:
  - RISC-V cores: SiFive, ETH
  - ML: Nvidia
  - SoC: platforms to stitch IPs together
- HLS like XLS help create different accelerators
- Interconnects:
  - On-chip: multiple options including custom
- IOs:
  - PHY layer from a standard,
  - Custom flavors of data, link and transport layers
  - Chiplets in a package, across chips, in a rack, across racks
- Analog Mixed Signal IPs
  - Use of standard based IPs is preferrable
- Cell library and PDK
  - PLLs, ROs,
  - Std cells

Custom designs for architecture ideas need significant peripheral IP support

Proprietary IPs in Open-source Silicon

Open-source hardware development should allow

- A mix of open-source and proprietary IP as-needed
  - Satisfy short term or critical needs
  - Optimized design needed to show the real benefits for architecture ideas
  - New IPs are developed only when there is a need

- Proprietary AMS and selected digital designs help comparisons with commercial offerings

- Proprietary tools:
  - Expected to play significant role for efficient designs
  - Open-source tools becoming richer and numerous with time

Proprietary IPs may be needed to show architecture benefits against the best-in-class design

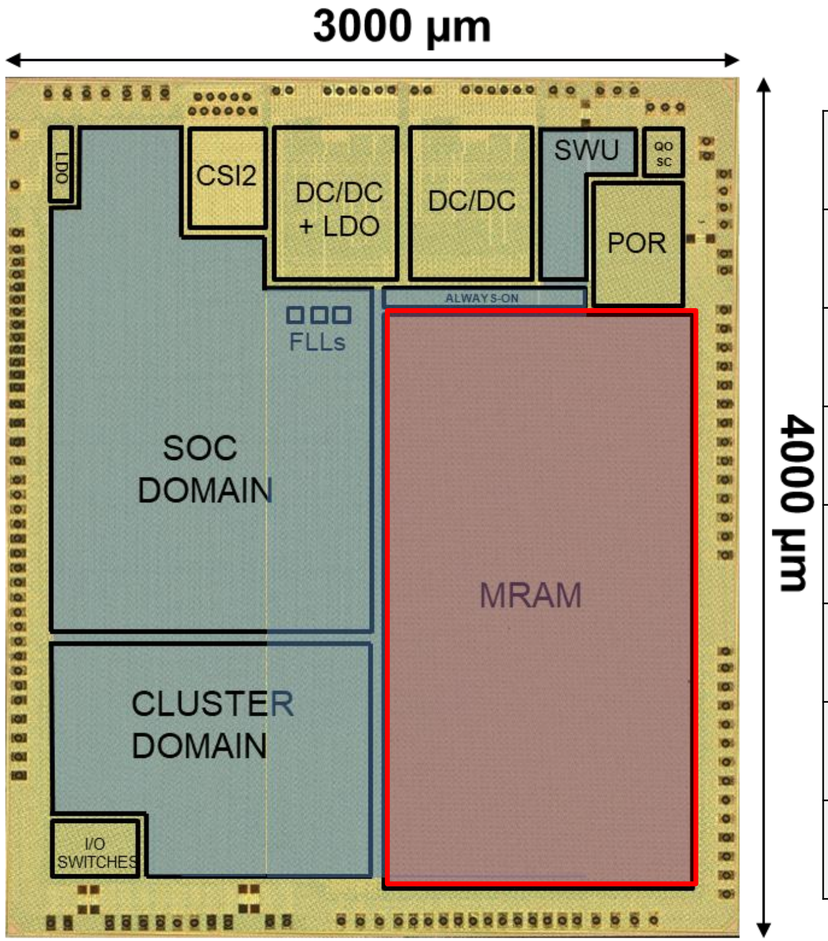
Agenda

1. Motivation
2. Open-source silicon
3. **Platforms and Design IPs**
4. Tool enablement
5. Technology enablement
6. Bringing all together

ETH VEGA: Extreme Edge IoT Processor

- RISC-V cluster (8cores +1) 614GOPS/W @ 7.6GOPS (8bit DNNs), 79GFLOPS/W @ 1GFLOP (32bit FP appl)
- Multi-precision HWCE(4b/8b/16b) 3×3×3 MACs with normalization / activation: 32.2GOPS and 1.3TOPS/W (8bit)
- 1.7 µW cognitive unit for autonomous wake-up from retentive sleep mode

- **Fully-on chip DNN inference with 4MB MRAM (high-density NVM with good scaling)**



Technology	22nm FDSOI
Chip Area	12mm^2
SRAM	1.7 MB
MRAM	4 MB
VDD range	0.5V - 0.8V
VBB range	0V - 1.1V
Fr. Range	32 kHz - 450 MHz
Pow. Range	1.7 µW - 49.4 mW

In cooperation with GREENWAVES TECHNOLOGIES GAP9

ETH: Heterogeneous Accelerators

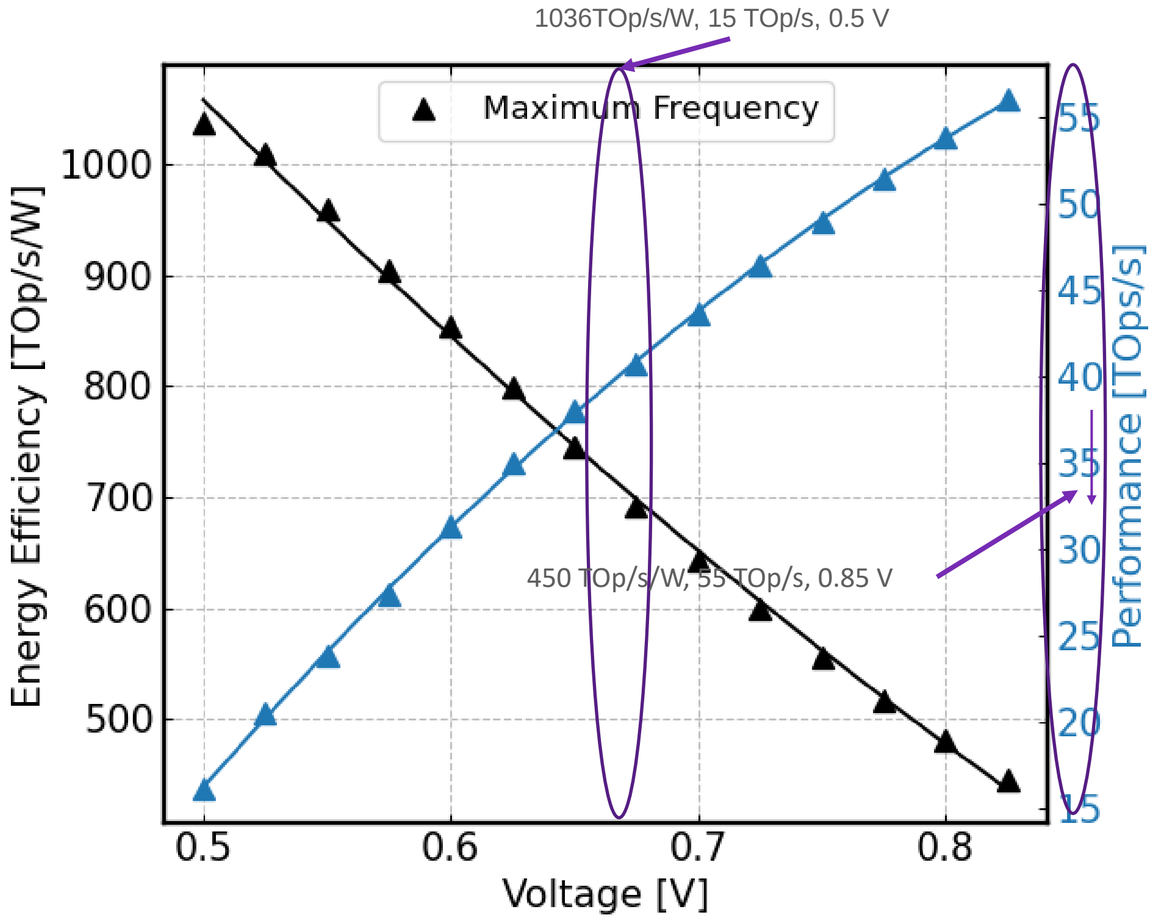
The *Kraken*: TCNs and SNNs at The Extreme Edge

- RISC-V Cluster (8 Cores + 1)
- CUTIE – dense ternary neural network accelerator
- SNE – energy-proportional spiking neural network accelerator
- DVS Interface for hardware support of event-based vision
- PULPO – Floating point linear algebra accelerator

Technology	22nm FDSOI
Chip Area	9 mm^2
SRAM SoC	1 MB
SRAM Cluster	4128 KB
VDD range	0.55V - 0.8V
VBB range	0V - 1.1V

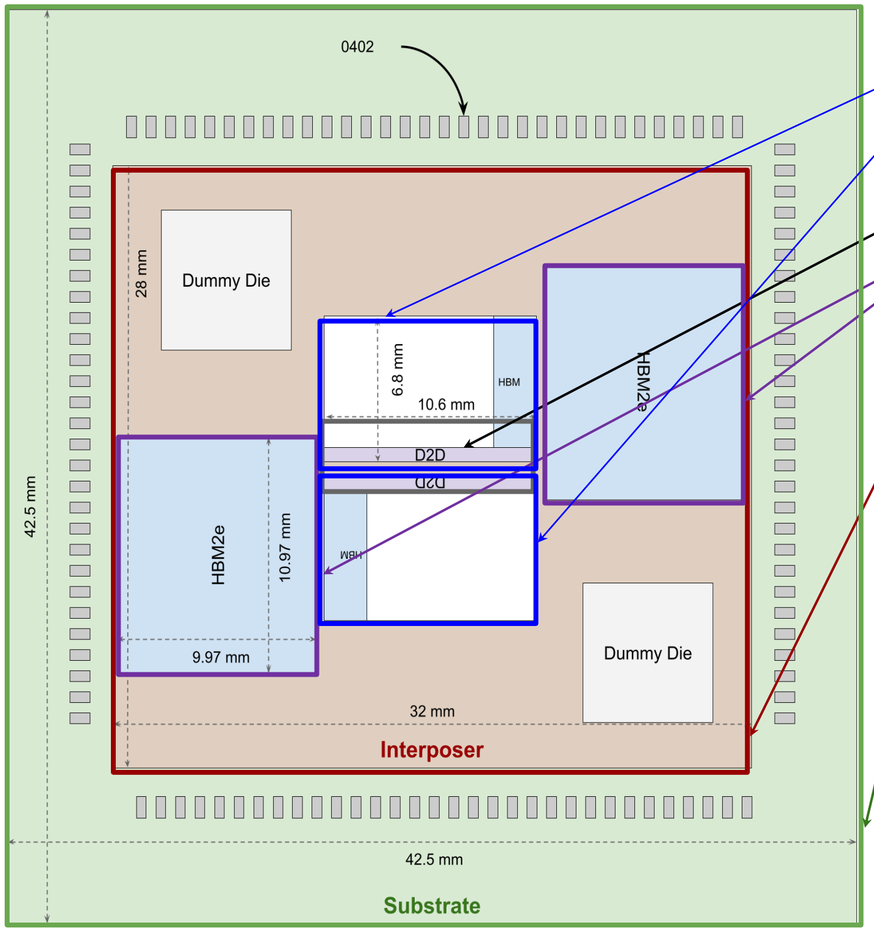
ETH Kraken: Results for CUTIE

- **Key implementation results**
  - Area: ~3 mm^2
  - Voltage: 0.5 – 0.9 V
- **Inference on CIFAR-10 - Ternary**
  - Accuracy: **86%**
  - Energy per inference: **2.72µJ**
- **Inference on CIFAR-10 - Binary**
  - Accuracy: 82%
  - Energy per inference: 4µJ!
- **Achievable Efficiency and Performance:**
  - Peak Core Energy Efficiency: **1036 Top/s/W**
  - Peak Throughput: **55 TOp/s**



- **DVFS controls are helpful for performance vs efficiency**
- **Both accelerators can work together for inference: CUTIE TNN and SNE SNN**

ETH: Leveraging Chiplets: Occamy



Dual-chiplet

- Area: **~70mm^2**

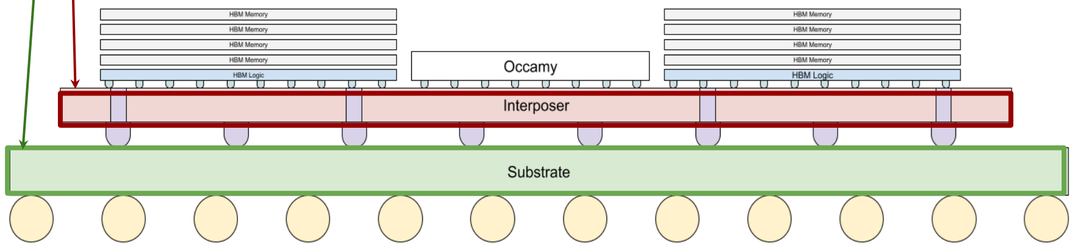
Chip2chip link

HBM2e memories

Interposer

- passive

Substrate



- Cores: target 1GHz (typ)
- 2 AXI Interconnect Subsystems (multi-hierarchy)
  - 64-bit
  - 512-bit with “interleaved” mode
- Peripherals
- Linux-capable manager core CVA6
- 6 Quadrants: 216 cores/chiplet
  - 4 cluster / quadrant:
    - 8 compute +1 DMA core / cluster
    - 1 multi-format FPU / core (FP64,x2 32, x4 16/alt, x8 8/alt)
- 8-channel HBM2e (8GB)
- D2D link (Wide, Narrow)
- System-level DMA
- SPM (2MB wide, 512KB narrow)

- Open source cores and interconnects with Proprietary IPs
- Complex project: from open source SW stack down to HW implementation and validation

Agenda

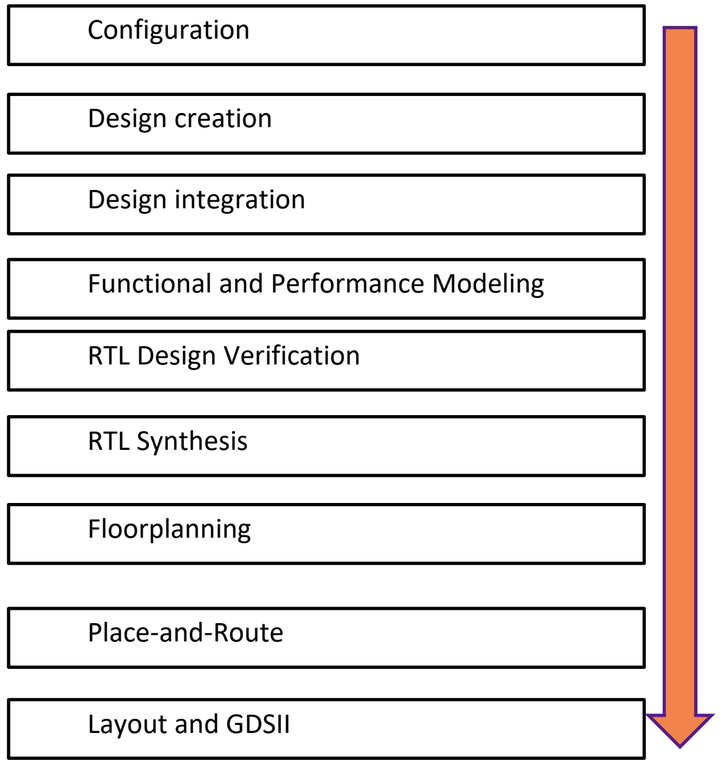
1. Motivation
2. Open-source silicon
3. Platforms and Design IPs
4. **Tool enablement**
5. Technology enablement
6. Bringing all together

Tools for open-source development

Tool and design costs increase exponentially with advanced nodes

Sample of tool chain

- ESP: Platform to create SOC with different open-source IPs and custom designs to generate RTL
  - Includes essential interconnect, interface IPs
  - Proprietary IPs can be used as well
- Performance and functional modeling:
  - SPARTA: Performance and functional modeling framework
  - Qualcomm AIMET Toolkit for ML accelerators
- OpenHW: Set of essential IPs like RISC-V cores
  - Multiple groups creating RISC-V DV infrastructure
  - Proprietary DV tool chain helpful for commercialization
- OpenROAD provides an EDA tool chain from Synthesis to GDSII
  - Combines OpenLANE automation for RTL to GDSII
  - Generates ground rule conforming designs



Prof. Kahng, UCSD, IEEE VLSI Symposium on Tech and Circuits, 2022. https://theopenroadproject.org

Tools: SPARTA

Sparta is a framework

Much like STL or Boost, Sparta contains a set of C++ classes, structs, and utilities used to build a performance and/or functional simulators

Sparta is a Discrete Event Driven (DES) Framework

Work is done by defining and scheduling events during simulation

Simulation is considered finished when there are no more events to execute or an explicit halt is requested

Sparta contains components for

Defining a simulation instance and hierarchy

Defining units/resources representing modeling components

Defining parameters for runtime behavior modification

Defining dynamic configurations and architectures

Defining counters, statistics, reports, logging, triggering, notifications, etc

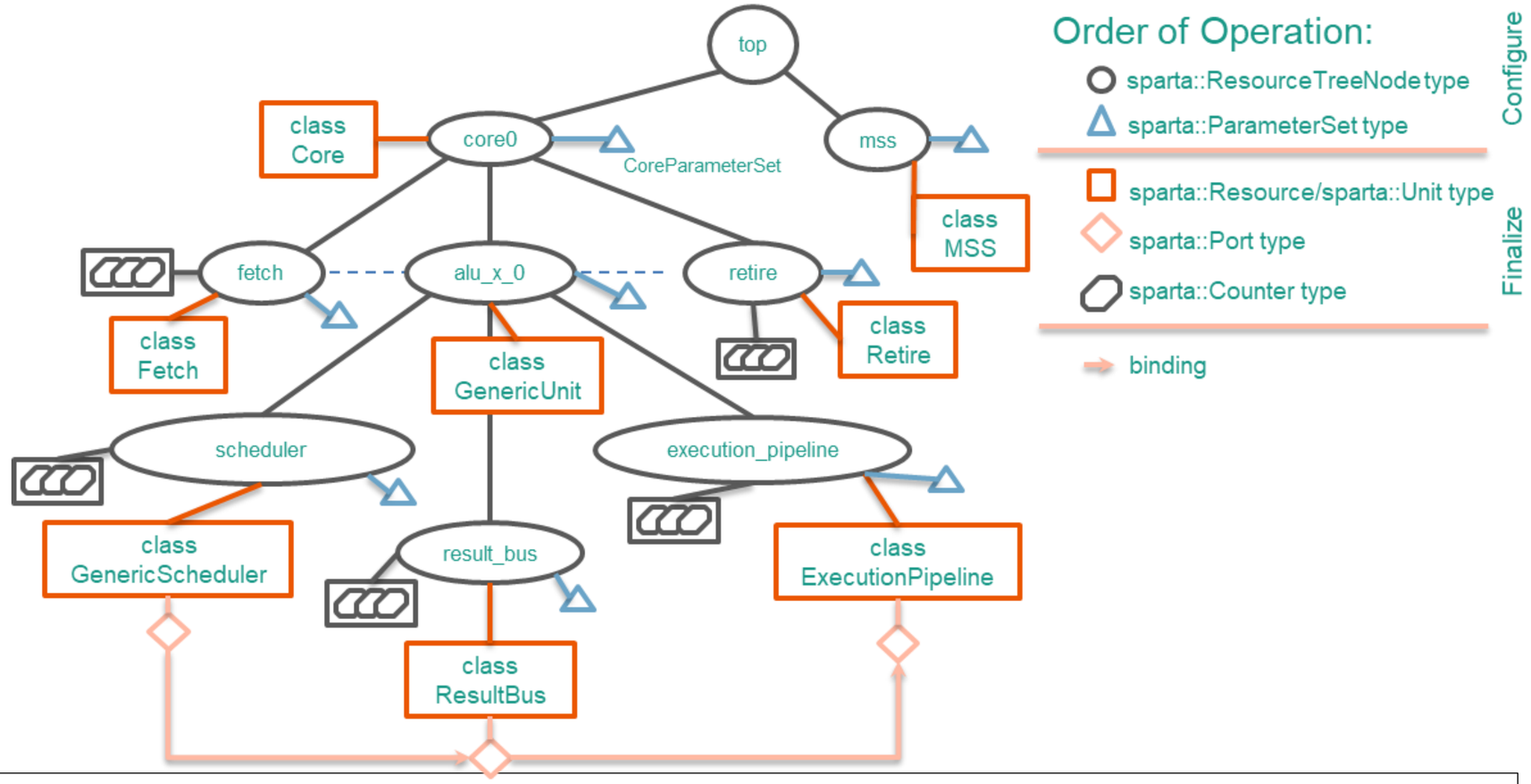
Defining timed events, ordering of those events, and scheduling of those events

Open-source Performance modeling framework: ARM, RISC-V and AMD/Samsung GPU

Knute Lingaard: SPARTA Presentation Link on Github : https://github.com/sparcians/map

Tools: SPARTA Simulation Layout

- Developed over a decade through Freescale, Samsung, SiFive
- Detailed and heavy duty SoC modeling framework
  - Single and multi-core, cache and memory hierarchies,
  - Extensive configurability
  - Data science hooks



Simulations, visualizations, data collections for commercial SoCs

Knute Lingaard: SPARTA Presentation Link on Github : ttps://github.com/sparcians/map

Agenda

1. Motivation
2. Open-source silicon
3. Platforms and Design IPs
4. Tool enablement
5. **Technology enablement**
6. Bringing all together

Technology enablement: Tapeout

Typical EDA Methodologies, Tapeout and Silicon

- Physical realization on a technology node: Primary concerns are:
  - Does the node meet cost expectation?
  - Are PDKs, std cell libraries, memories available on the node?
  - Are all digital and AMS IPs available, whether open source or proprietary?
  - Are the integration criteria and special concerns for all IPs met?
    - Metal stack, power supplies, power islands, etc.

- Next objectives: EDA tool chain to generate GDSII through synthesis on a technology node
  - Pipe cleaning of steps starts before the design is frozen
  - Stages of design freeze to accommodate reduced level of design changes
  - **Golden design** (taped out) continues through verification for potential bugs and work arounds
  - 2-4 months to real silicon depending on foundry, technology, space availability, etc.

- Google, OpenROAD[1] partnership helped 200+ tapeouts:
  - Foundries: Skywater, GF, Intel
  - Nodes: 130nm, 90, 65, 16, and 12

Open-source consortiums taking small, significant steps on silicon realizations

1, Prof. Kahng, UCSD, IEEE Symp on VLSI Tech and Circuits, 2022.

New device technologies

Architectural impact

- Design-Technology-Co-Optimization (**DTCO**) and System-Technology-Co-Optimization (**STCO**):
  - What are the improvements in technologies?
  - What is the impact on transistors, silicon features and physical characteristics?
  - How does it impact the PDKs, standard cell libraries, memories etc.?
  - Can design IPs be optimized around newer development: feature size, power, performance?
  - **Does the roll up of characteristics impact architecture?**
    - Example: a frequency of operation can affect the implemented algorithm (e.g., branch prediction redirection)

- What are the **new devices** of interest?
  - Do they improve the characteristics of existing devices?
  - Are there fundamental characteristics to be exploited?
    - NVMs is a big research area; and their characteristics differ widely
    - Analog-in-Memory-Compute and digital counterpart affect architectural exploitation
    - Can architecture blocks be created or changed to benefit **new or modified functions**?

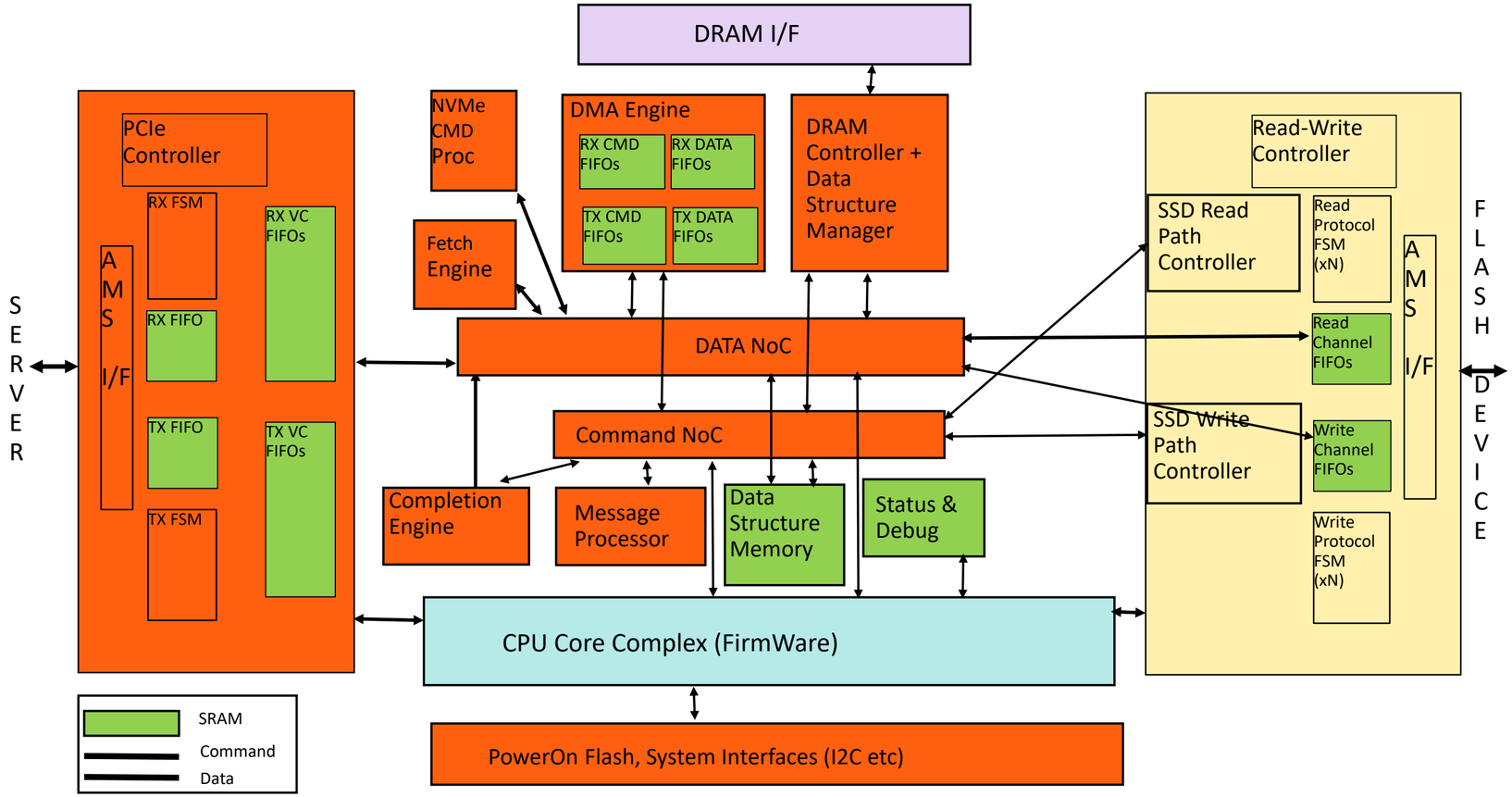
Better architecture using new technologies, devices and circuits to achieve system goals

1, Prof. Kahng, UCSD, IEEE Symp on VLSI Tech and Circuits, 2022.

Technology: Untouched data movements

Example: Solid State Device Controller

- Affects system level FoMs
- Significant performance impact
- Protocol and interface design research
- Considerations:
  - New technology
  - SW stack optimizations
  - New applications
    - ML
    - Database operations
    - Memory level parallelism exploitation



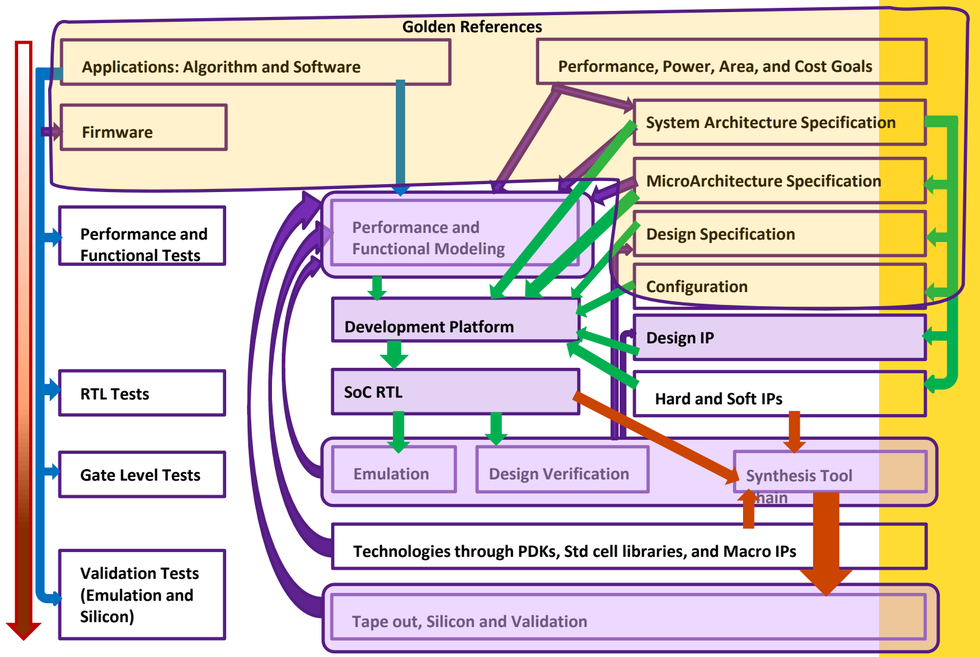
Example of research areas typically untouched for performance benefits

Agenda

1. Motivation
2. Open-source silicon
3. Platforms and Design IPs
4. Tool enablement
5. Technology enablement
6. **Bringing all together**

A Running thread: Software Stack

- Application, the algorithm to implement and its translation to the workload is on of the most important piece of hardware optimization
  - Compilation to map to a desired microarchitecture
  - Tests for performance and power models
  - Functional, performance and power tests on RTL design
  - Basic gate level tests for synthesized netlists
  - Validation tests for emulation and silicon
- Application can exploit newer technologies or newer characteristics to modify how the objectives are met
  - Affects algorithm, compilation, data structure, data layout, data movement for computation and storage
- On design completion, use of the hardware by algorithm and compiler are the main tools available to a user

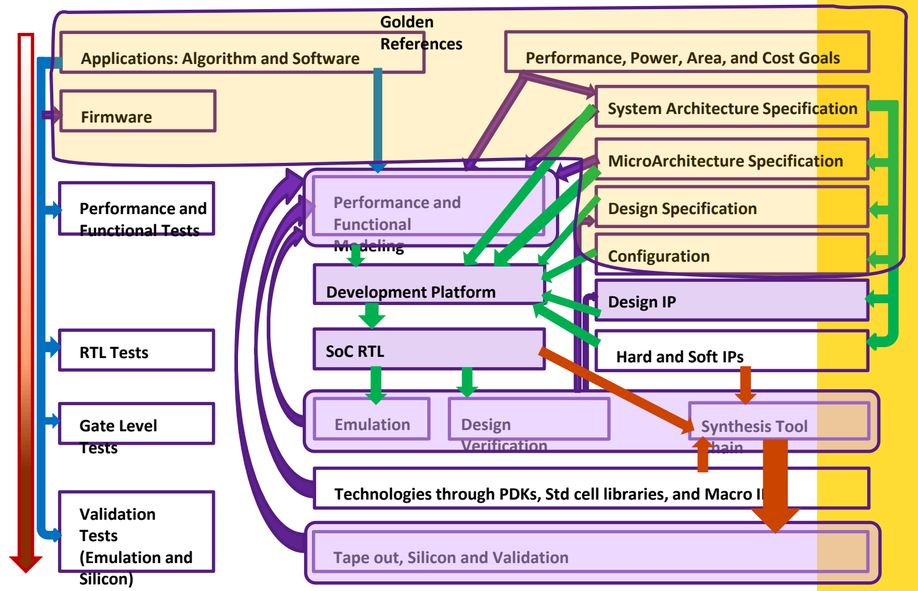


SW stack provides benefit from new technologies before design, optimization guidance during design, and optimal mapping after design

Bringing all together

Biggest performance and efficiency gains need compute and storage to spend the least latency and energy for data transfers and evaluation. Can “the architecture idea” deliver?

- Open-source tools and IPs are maturing to create RTL designs and realize in silicon
- Each architecture idea needs to be vetted against best-in-class designs for its true potential
- A mix of global and local optimization iterations in the design flows is needed for large efficiency gains
- New technologies can benefit change of algorithm, system and microarchitecture for orders of magnitude gains
  - Accelerators and memory hierarchies with their sizes and functions can be heavily optimized for target applications



GF helps silicon realizations of architectural proof-of-concepts as enabler of research ideas and beyond

GF Technology Development with Ecosystem

~1400

technologists in dedicated research teams

>30K

wafers per year dedicated to development

>50

universities, government partners and other research institutes partnered in collaborative efforts

>150

differentiated programs built on 25+ world class platforms

Global manufacturing footprint

Burlington, VT, USA
Fab 9

Wafer size: 200mm
Technology: RF SOI, SiGe

Malta, NY, USA
Fab 8

Wafer size: 300mm
Technology: FinFET, NVM, RF SOI, SiPh

Dresden, Germany
Fab 1

Wafer size: 300mm
Technology: FDX™, NVM, HV, BCDLite®

East Fishkill[1], NY, USA
Fab 10

Wafer size: 300mm
Technology: HP CMOS, RF SOI, SiPh

Singapore[2]
Fab 7 / GIGA+ / New Fab 2023

Wafer size: 300 & 200mm
Technology: BCD/BCDLite®, HV, NVM, DDI, RF SOI, LP SiGe

Notes:
(1) We plan to transition our facility in East Fishkill to ON Semiconductor by the end of 2022.

Open Source Links in the Presentation

Platforms:

- ESP: https://esp.cs.columbia.edu
- NEEDLE: https://github.com/sfu-arch/

Tools:

- SPARTA: Performance and functional modeling framework: https://github.com/sparcians/map
- OpenROAD (and OpenLANE): https://theopenroadproject.org/
- OpenLane (now as part of OpenRoad offering): https://github.com/The-OpenROAD-Project/OpenLane

IPs:

- RISC-V International: https://riscv.org/
- ETH Zurich: RISC-V PULP Platform: http://pulp-platform.org

Thank You