2020 DESIGN AND VERIFICATION™ DVCON CONFERENCE AND EXHIBITION UNITED STATES

Next Generation Verification for the Era of AI/ML and 5G

Frank Schirrmeister, Pete Hardee, Larry Melling, Amit Dua, Moshik Rubin

Cadence Design Systems, Inc.

cadence®

accellera SYSTEMS INITIATIVE

© Cadence Design Systems, Inc - All Rights Reserved

Agenda

- The era of 5G and AI/ML
  - Challenges
  - What they mean for verification
  - Typical Designs
- Verification options to *enable* 5G/AI/ML
- AI/ML Inside/Outside for Verification
- Summary - Outlook

© Cadence Design Systems, Inc - All Rights Reserved

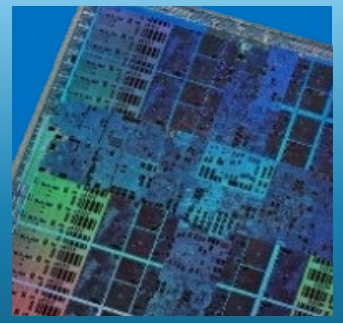
THE ERA OF 5G AND AI/ML

© Cadence Design Systems, Inc - All Rights Reserved

Electronics Innovation

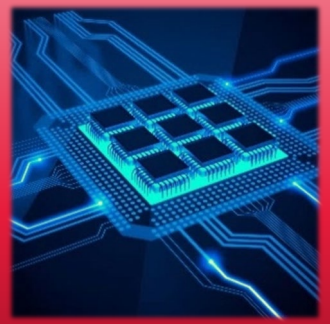
2020

Edge-to-Cloud and 5G

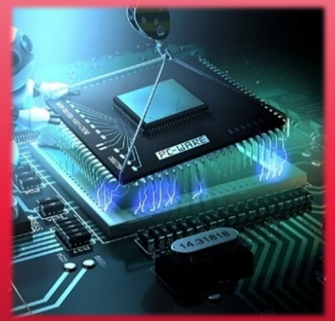


Specialized SoCs

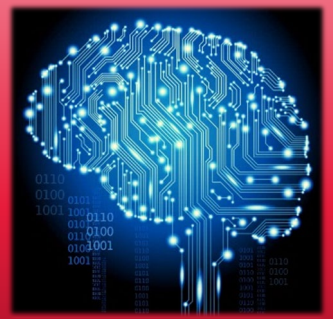
Artificial Intelligence



Parallel and Distributed

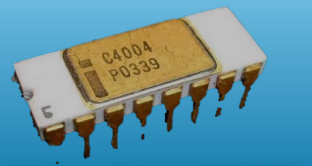


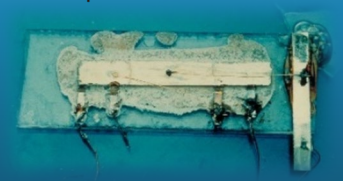
Full-flow Solutions

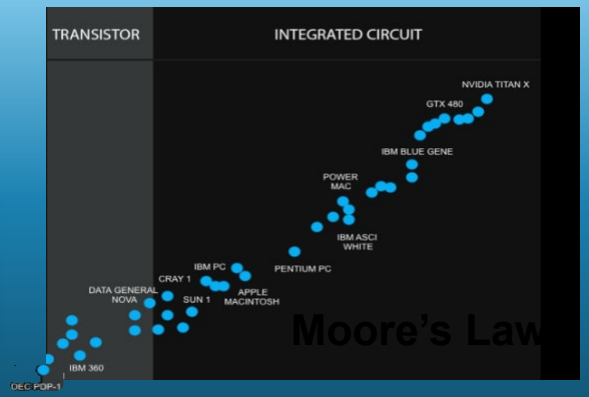


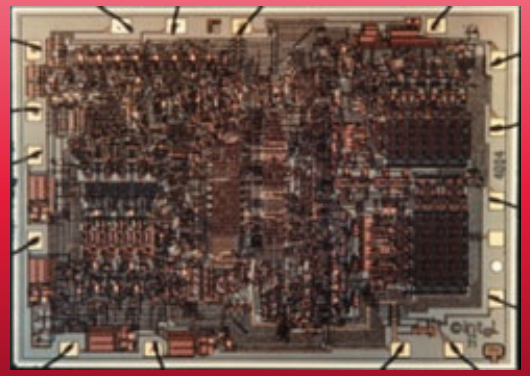
Deep Learning

1960









Semiconductors and Systems

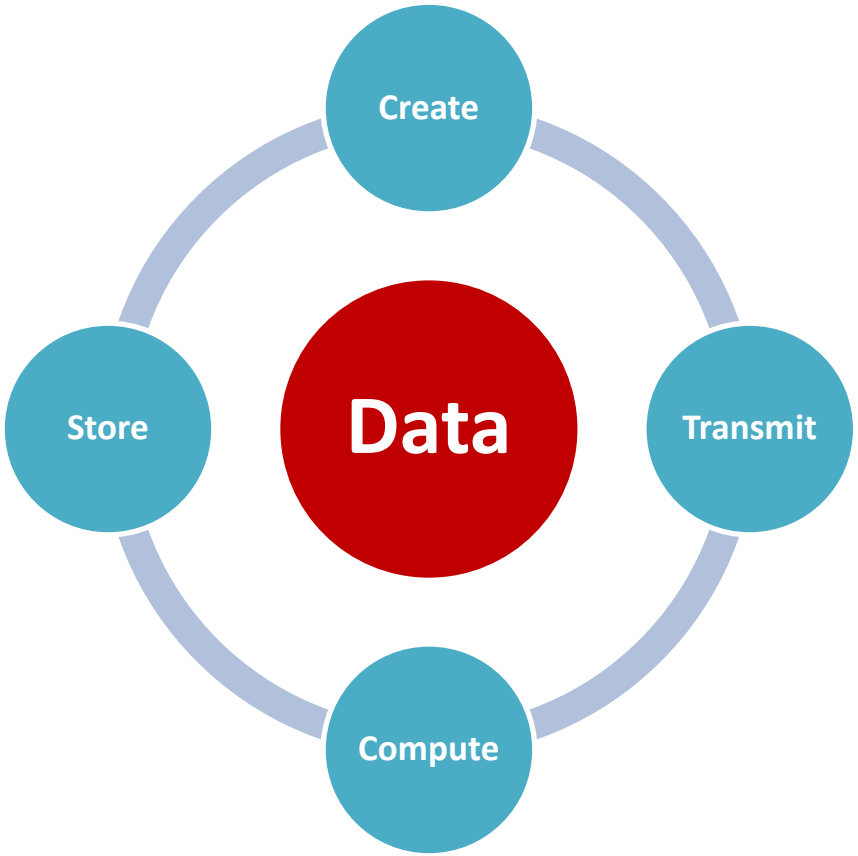
EDA system design enablement

© Cadence Design Systems, Inc - All Rights Reserved

A Data-driven World

Sensors Everywhere
Internet of Things

Memory Technology
(HDD, DRAM, NAND)

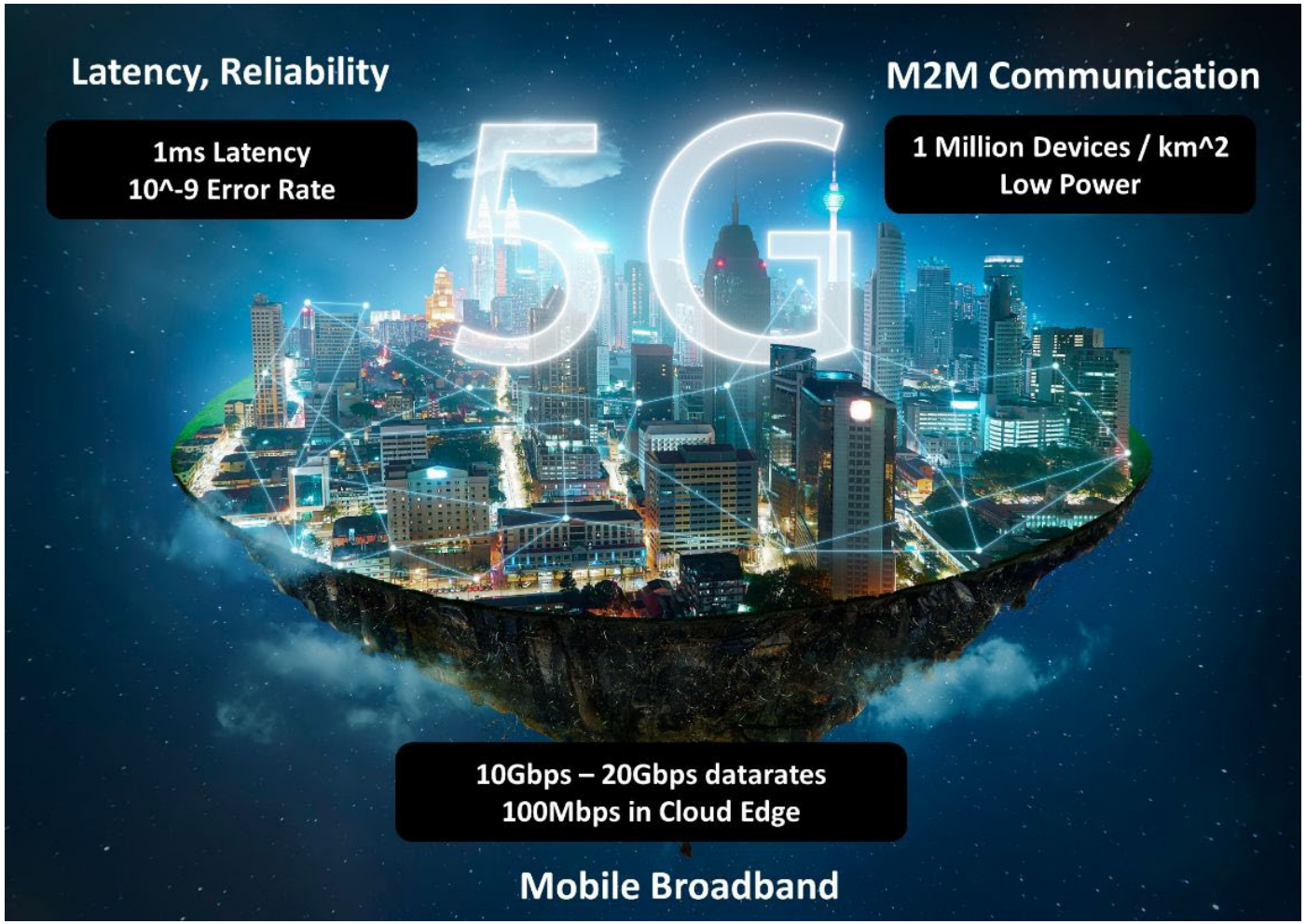


Wireless & Wired Infrastructure
5G

Datacenter, Servers
ML/AI

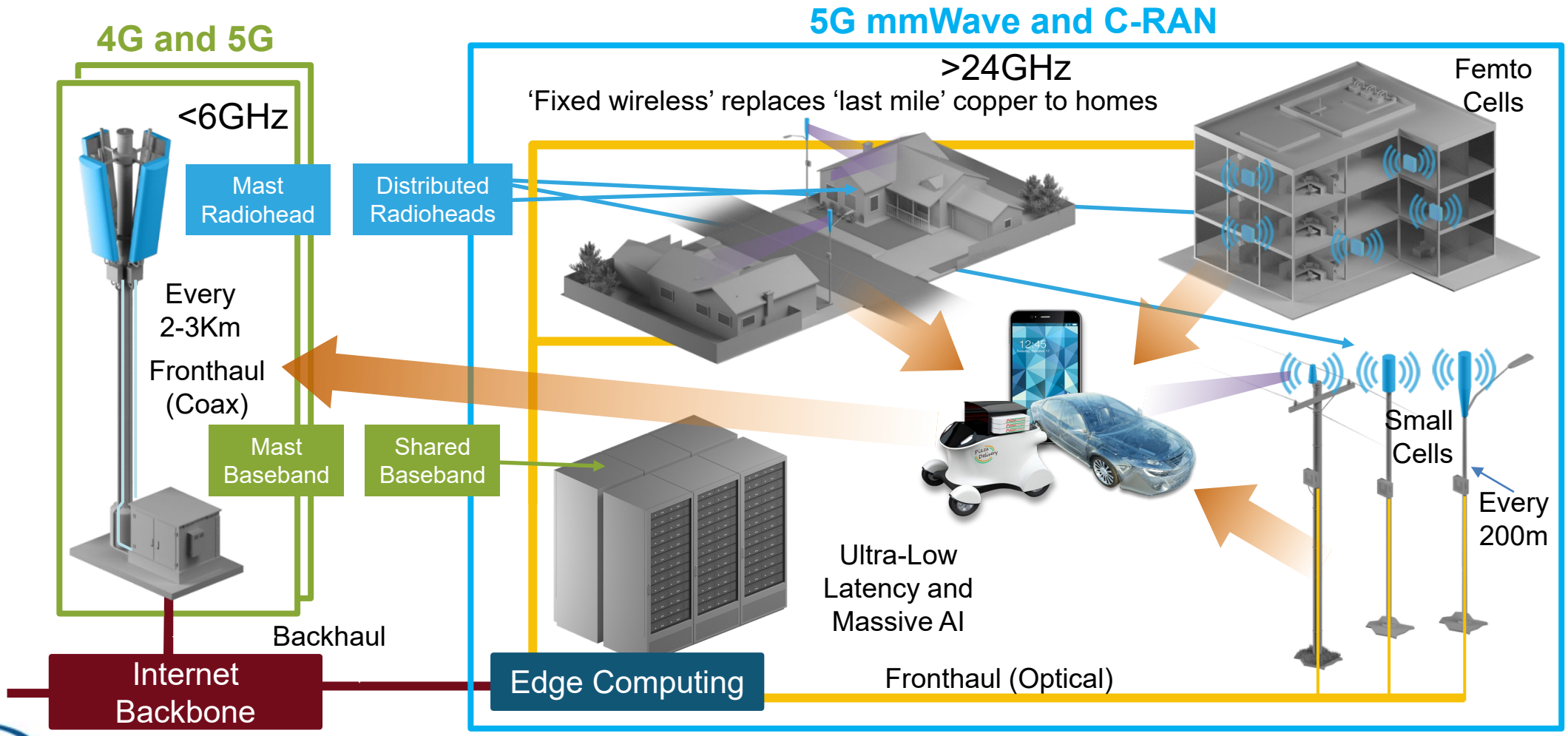
© Cadence Design Systems, Inc - All Rights Reserved

The Promise of 5G – Where to Start?



© Cadence Design Systems, Inc - All Rights Reserved

5G — New Bandwidth, New Subsystems



© Cadence Design Systems, Inc - All Rights Reserved

5G Requirements 1/2

Enhanced mobile broadband

- **Needs**
  - Faster speed, Lower latency, Greater capacity
  - On-the-go, ultra-high-definition video, virtual reality, and other advanced applications.
- **Design characteristics**
  - traditional large (>200MG), complex designs
  - requiring full-chip execution for SW - Emulation

Internet of Things

- **Needs**
  - Existing networks struggling
  - 5G unlocks (IoT) - more connections at once
  - Additional monthly revenues for carriers
  - IoT revenues smaller because of low usage
  - 5G competes against Wi-Fi and Zigbee.
- **Design characteristics**
  - much smaller (<32MG), very power sensitive
  - performance system dependent
  - multi-device simulation / emulation for QoS and performance validation

© Cadence Design Systems, Inc - All Rights Reserved

5G Requirements 2/2

Mission-critical & control

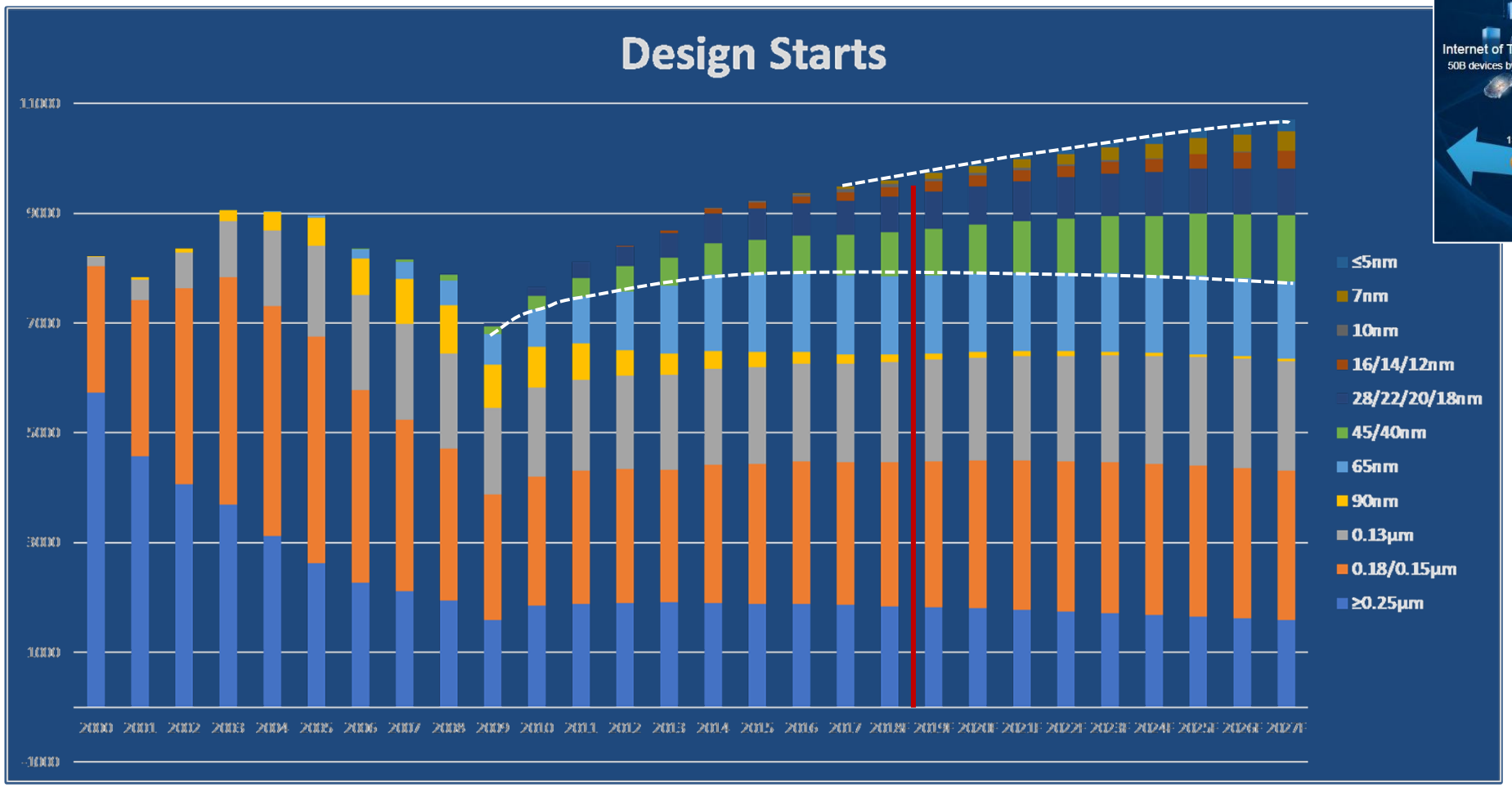
- **Needs**
  - absolute reliability in medical, vehicle safety
  - Latency limiting factor
  - 5G delivering lower latency
  - New use cases in healthcare, utilities, traffic management, and other time-critical contexts
  - Operators expect only incremental revenue
- **Design characteristics**
  - Small-to-medium designs (<200MG)
  - Functional safety drives need for system emulation

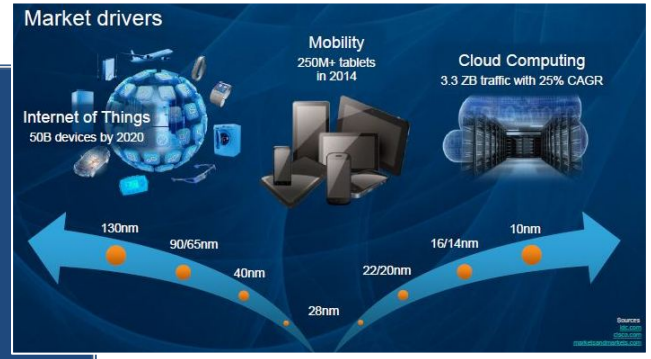
Fixed wireless access

- **Needs**
  - 5G, millimeter wave spectrum, capable of delivering speeds of more than 100 Mbps to the home
  - Viable alternative to wired broadband
  - New revenue stream for wireless operators in areas with less fiber/cable access
- **Design characteristics**
  - Extension to traditional base-station developers
  - More complexity requiring system emulation
  - Design size expected to be large (>200MG)

© Cadence Design Systems, Inc - All Rights Reserved

Design Start Market: Bifurcation



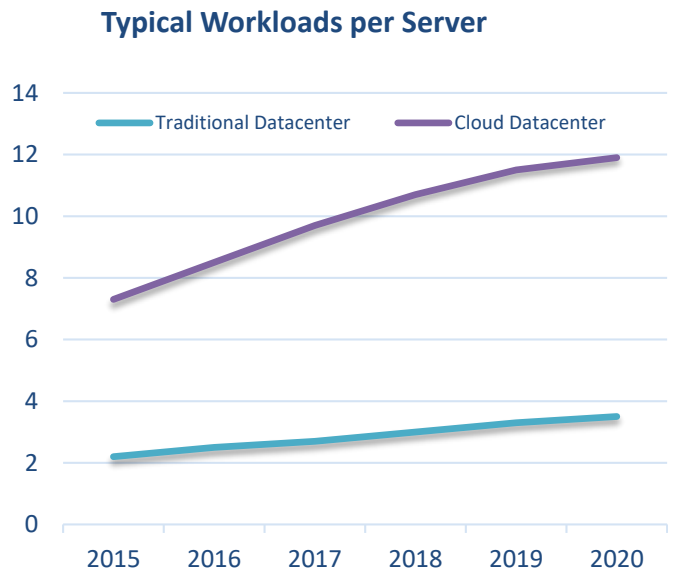


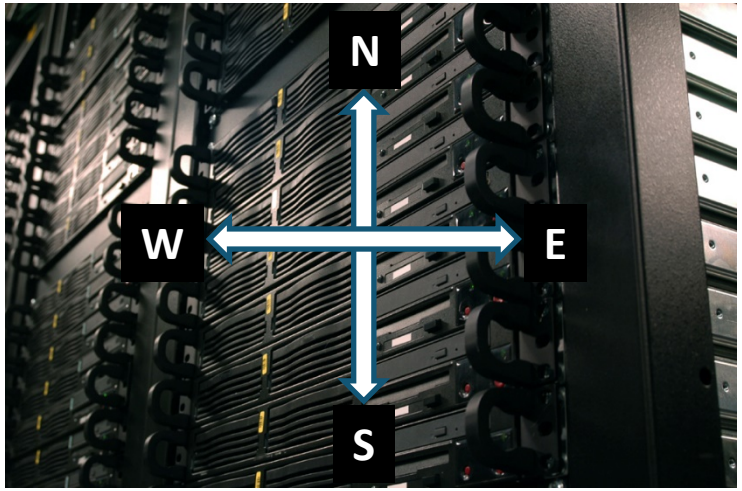
Source: IBS 2014 to 2018

© Cadence Design Systems, Inc - All Rights Reserved

Datacenter Opportunities

Workload-optimized, high-performance compute, connectivity, accelerators – AI/ML/DL





Hyperscale Optimization CPU

- Workload optimized
- Machine learning
- Deep learning
- Accelerator offloads

Rack-Level Connectivity

- Leaf /spine
- Memory pool (HBM)
- Connectivity / SiP
- Reduced latency
- Mesh / 3D-torus / fabric

Scale Out Clusters

- DNN
- SSD / NVMe
- Coherency
- VM / containers
- Mesh/3D-torus

© Cadence Design Systems, Inc - All Rights Reserved

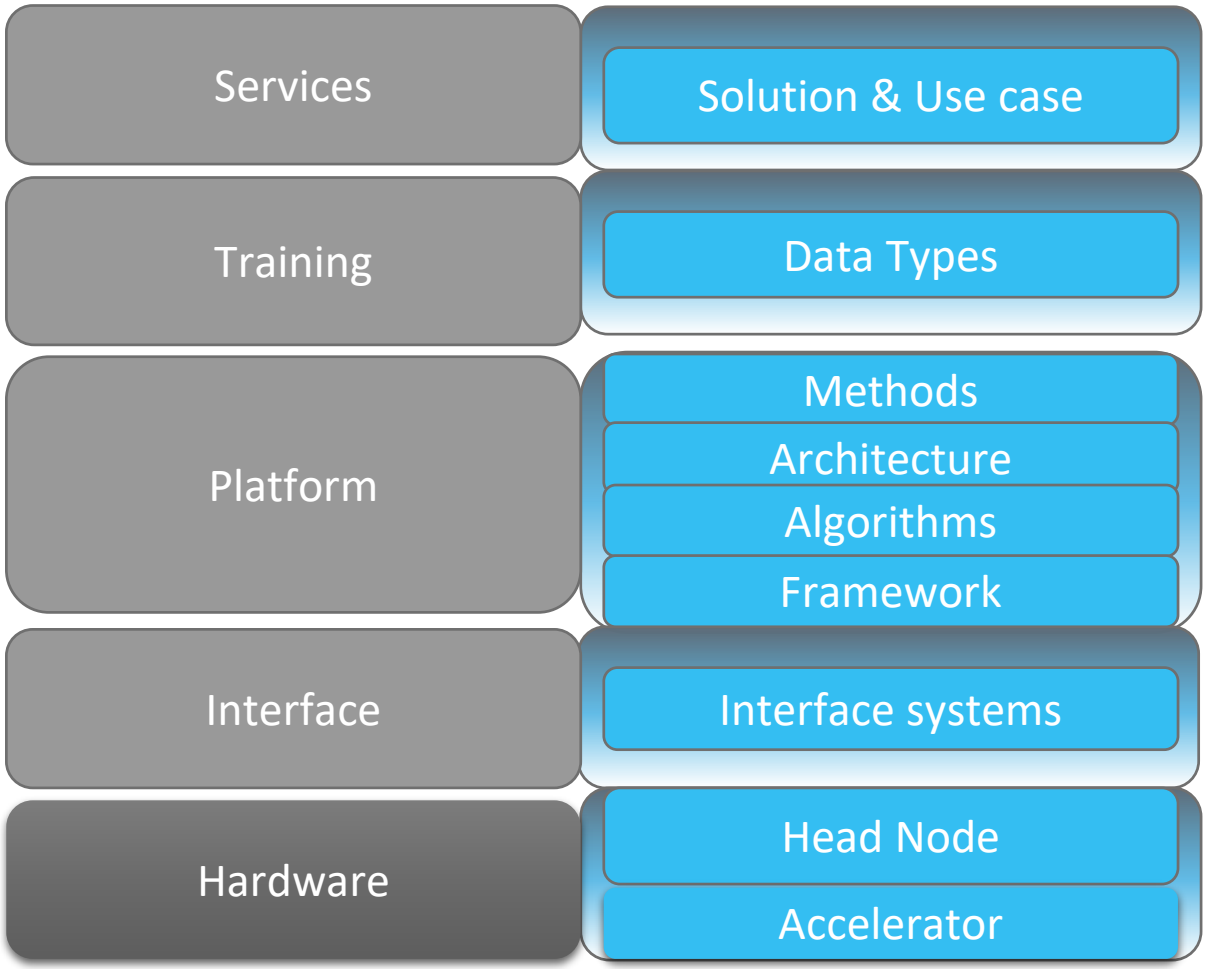
AI Chip Datacenter Technology

Opportunities in existing market	Emerging opportunities
• Accelerators for parallel processing, such as GPUs[1] and FPGAs[2]	• Workload-specific AI accelerators • Quantum computing, neuromorphic
• High-bandwidth memory • On-chip memory (SRAM[3])	• Quasi-volatile memory • Non-volatile memory (NVM) PCM, MRAM
• Potential growth in demand for existing storage systems as more data is retained	• AI-optimized storage systems • Emerging NVM (as storage device)
• Infrastructure for data centers	• Silicon photonics • Programmable switches

[1]Graphics-processing units. [2]Field programmable gate arrays. [3]Static random access memory.
Source: McKinsey analysis and Cadence

© Cadence Design Systems, Inc - All Rights Reserved

AI/ML Technology Stack



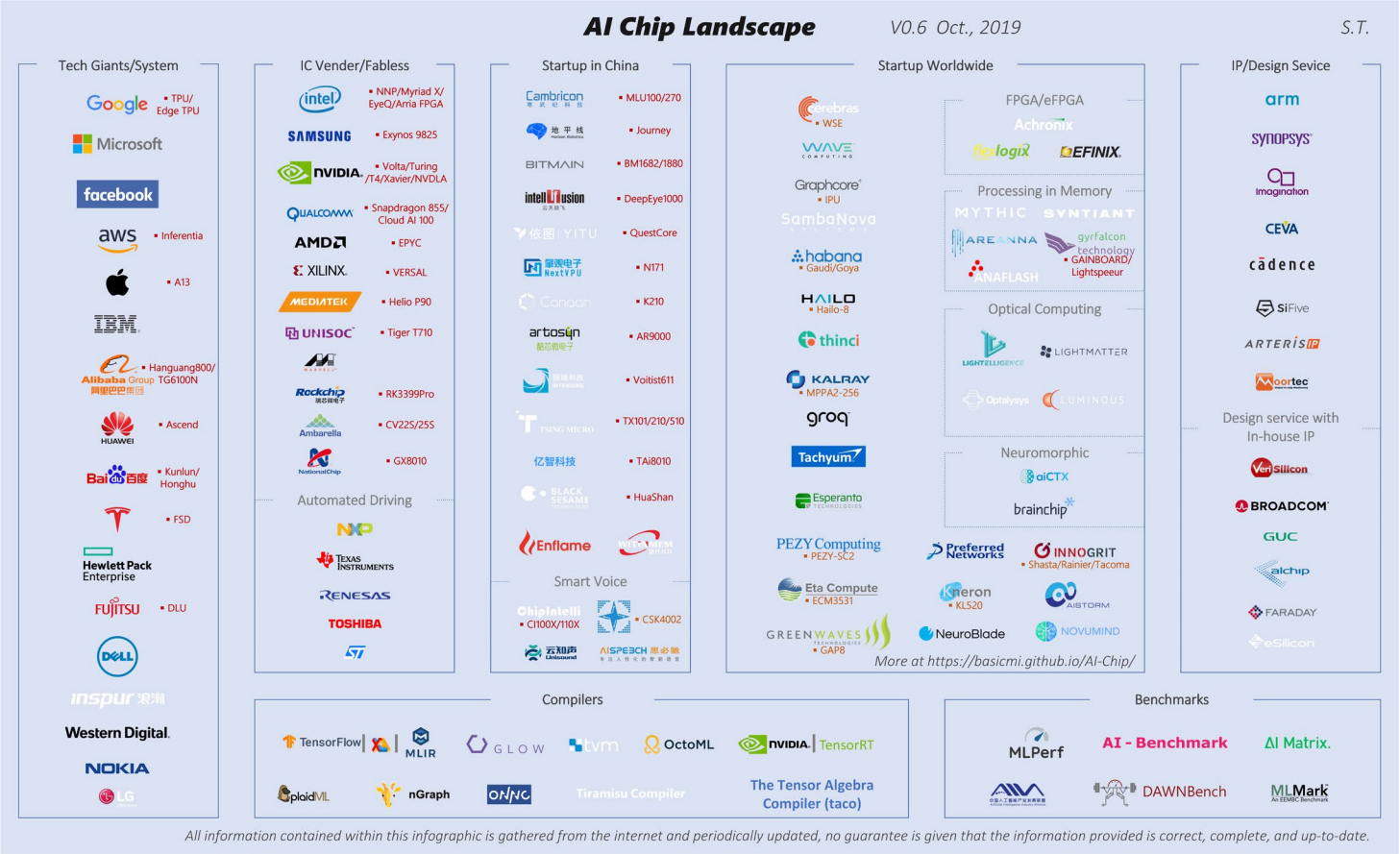
- Complex optimization across hardware, software, platforms and services
- Silicon designed to perform highly parallel operations required by AI enables simultaneous computations

"AI could allow semiconductor companies to capture 40 to 50 percent of total value from the technology stack, representing the best opportunity they've had in decades."

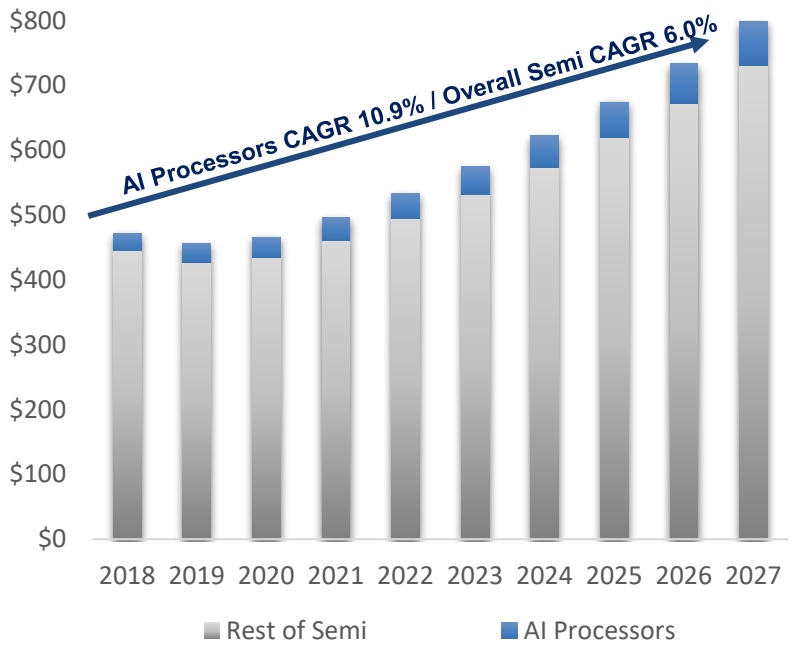
Source: McKinsey analysis and Cadence

© Cadence Design Systems, Inc - All Rights Reserved

AI Chip Landscape, Key Ecosystems



https://github.com/basicmi/AI-Chip

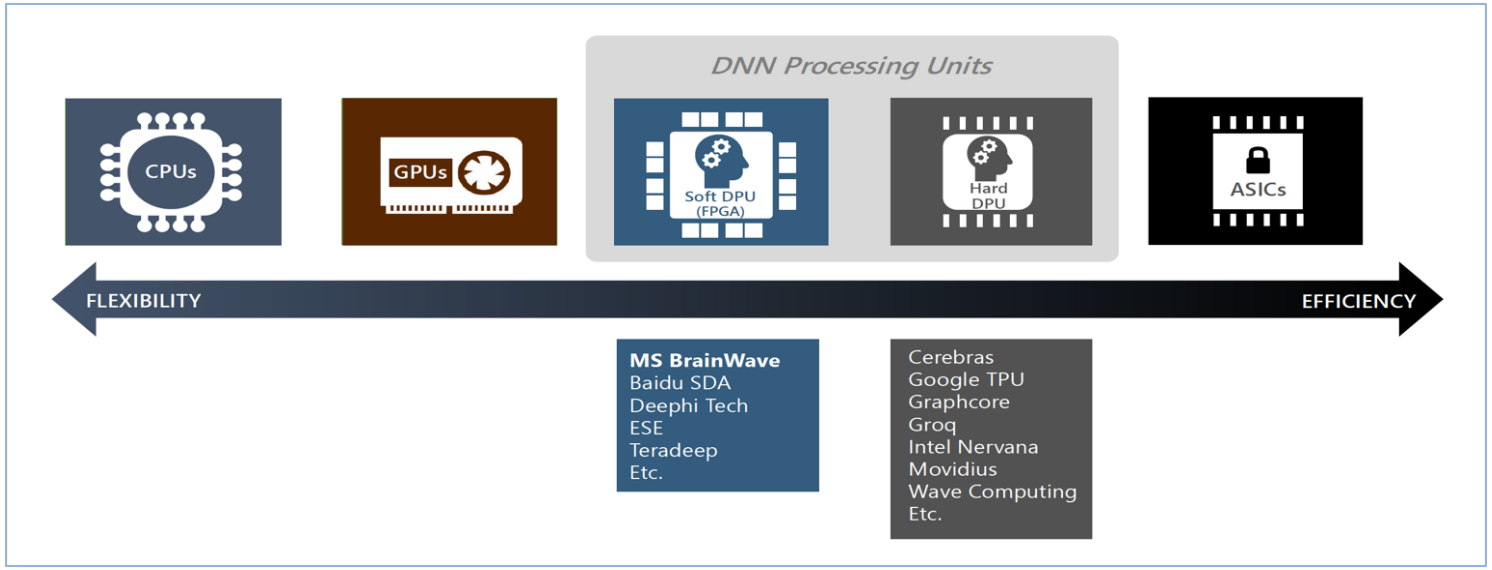


Source: International Business Strategies, Inc. (IBS) Semiconductor Market Data Centers, April 2019.

AI processors expected growth: 3X faster than total semi market

© Cadence Design Systems, Inc - All Rights Reserved

Many AI/ML Implementation Options



Source:
Microsoft. Hot Chips 2017

- Deep learning algorithms show a high degree of parallelism
  - General-purpose CPUs designed for sequential workloads
  - GPUs with massively parallel execution capability have popularized DNNs over the last few years
  - Proliferation of FPGAs and domain-specific accelerator ASICs for more efficiency

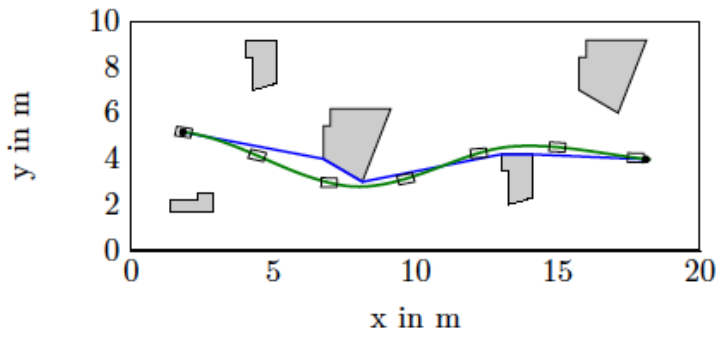
© Cadence Design Systems, Inc - All Rights Reserved

Decision Making is Shifting from MPC to AI

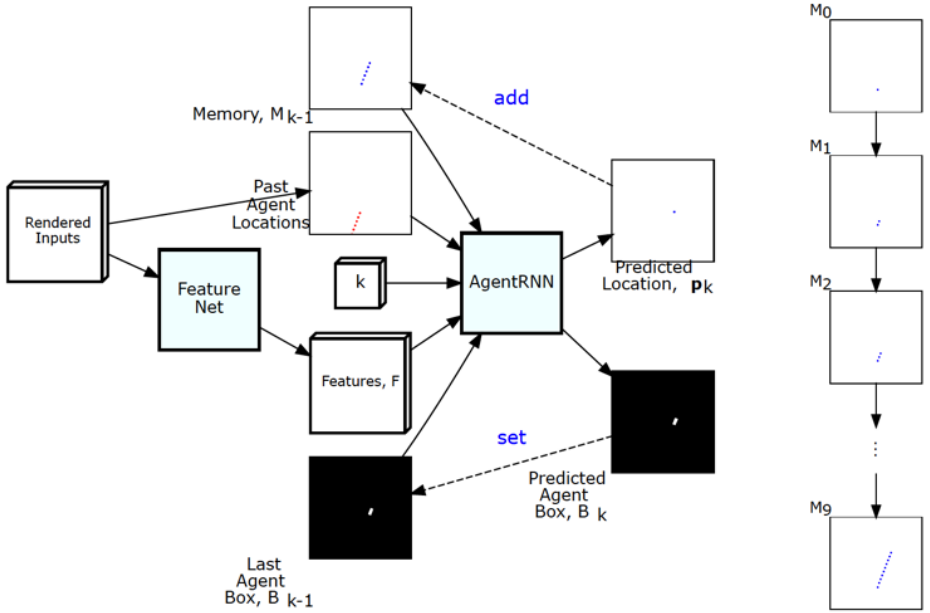
Example: Path Planning

Traditional:
Model Predictive Control (MPC)

MPC frameworks running on CPU to generate safe trajectories



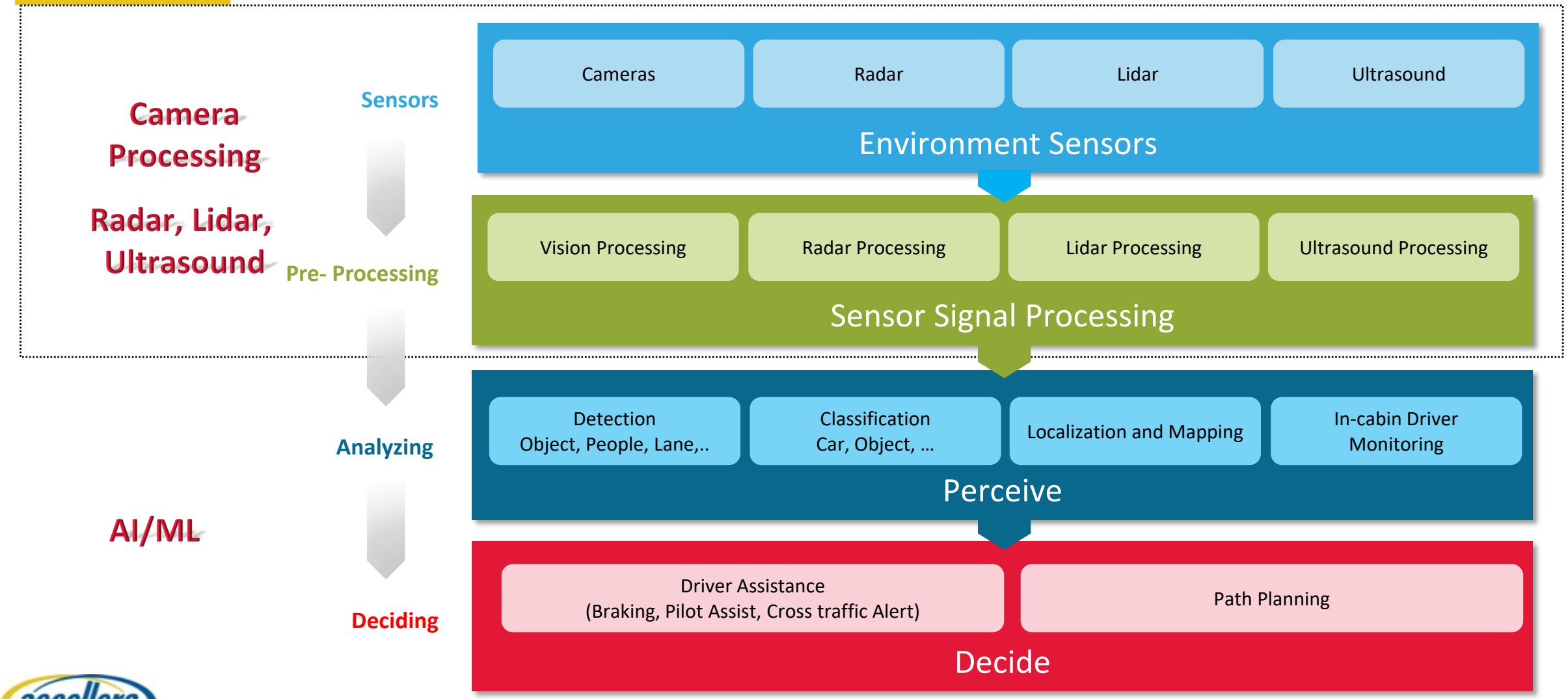
Present and Future:
ChauffeurNet - Google Waymo
Convolution and recurrent neural networks



https://arxiv.org/pdf/1812.03079.pdf

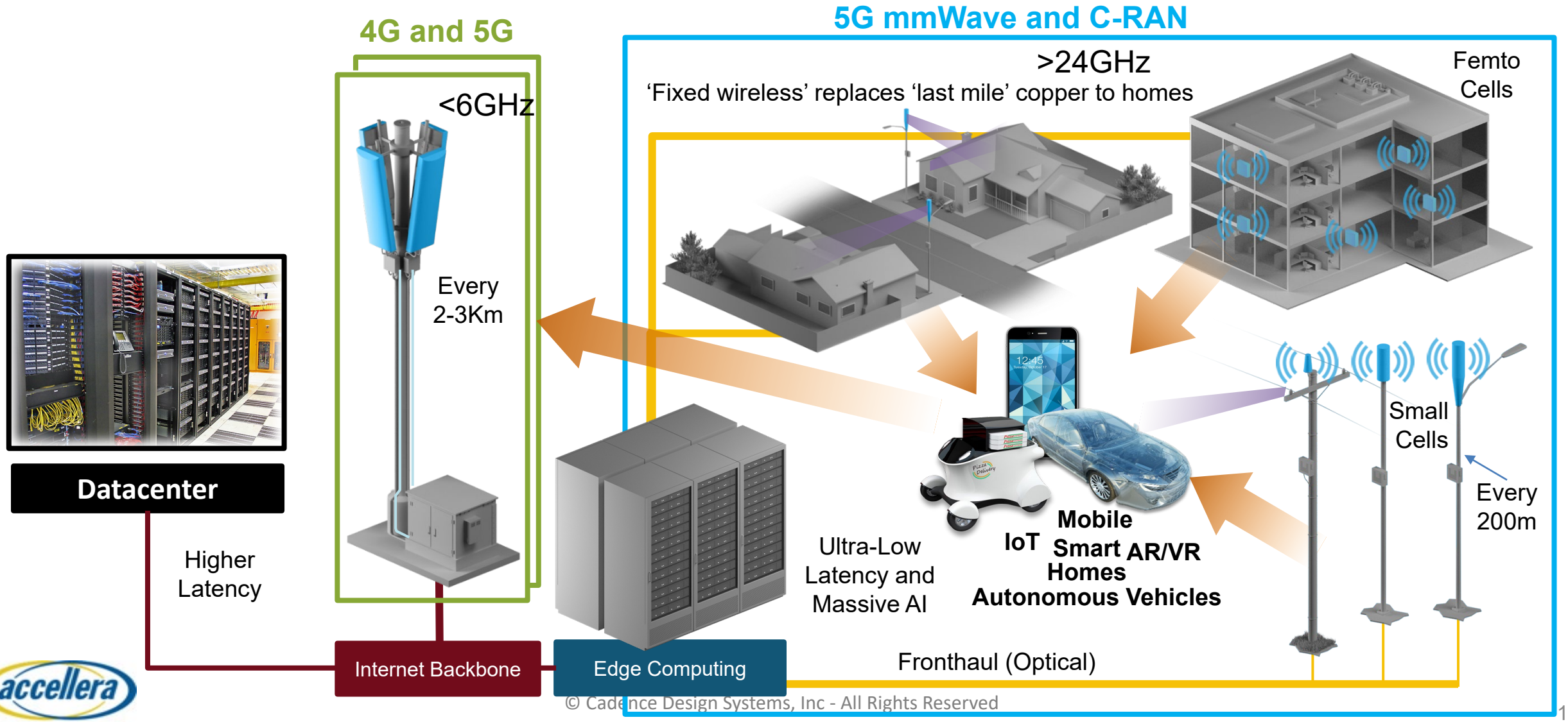
© Cadence Design Systems, Inc - All Rights Reserved

Accelerating AV and ADAS Experiences



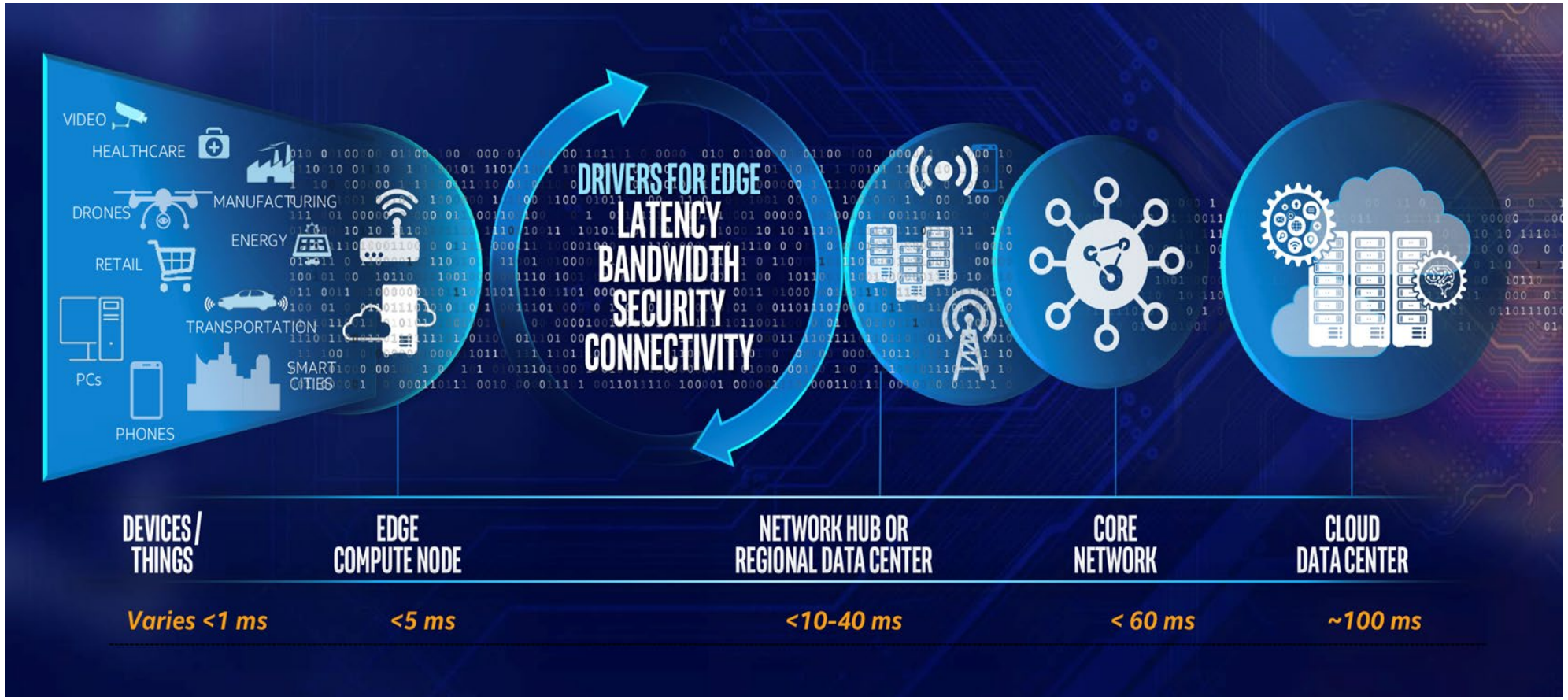
© Cadence Design Systems, Inc - All Rights Reserved

AI, ML and 5G are connected



© Cadence Design Systems, Inc - All Rights Reserved

Data Drives Edge Computing

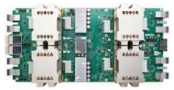


Source: Intel

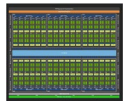
© Cadence Design Systems, Inc - All Rights Reserved

Location of data is crucial!

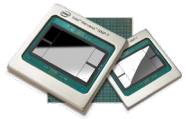
Datacenter Training
- High complexity NNs for ADAS
- Analysis to predict our behavior
- …



Google Cloud TPUs



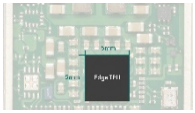
Nvidia GPUs



Intel Nervana NNP-T

Datacenter Inferencing
- More complex data - higher latency
- Train and deploy inferenced NN
- …

Edge Training
- Face recognition to unlock phone
- Update with localized data
- …



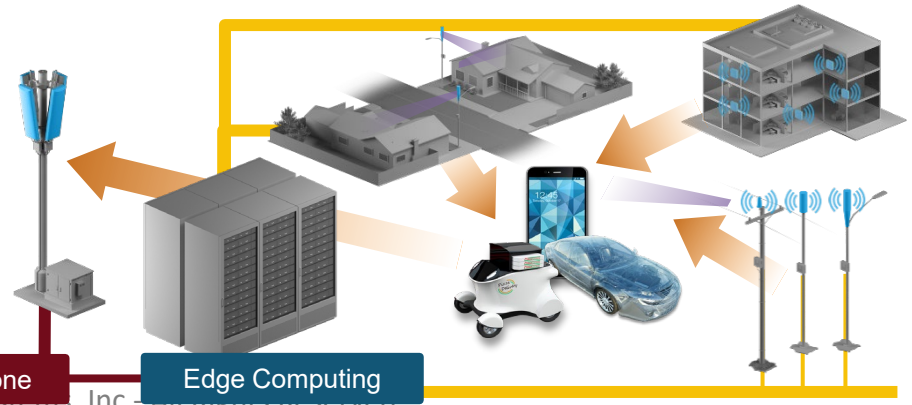
Google Coral

Nvidia Jetson Nano

Intel Up Squared

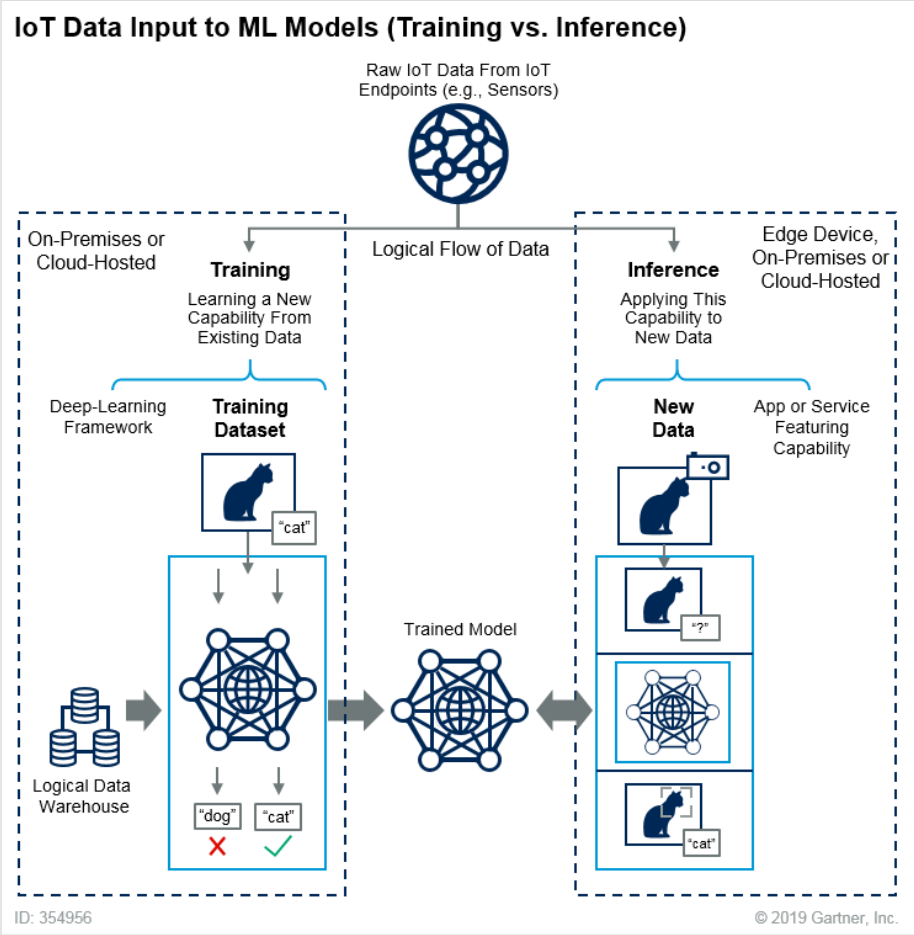
Edge Inferencing
- Real time decisions (Vision)
- Phone unlock
- …

Datacenter



© Cadence Design Systems, Inc - All Rights Reserved

Training & Inference at Datacenter & Edge



- AI in Datacenter
  - Scientists directly craft models for all situations
  - Model training is automated
  - Supervision is required in most cases to label data
  - Full fidelity data must be used for training
- AI at the Edge
  - Automatically adapt, deploy on available infrastructure
  - Performance needs to be achieved within the edge products' power and bandwidth constraints

© Cadence Design Systems, Inc - All Rights Reserved

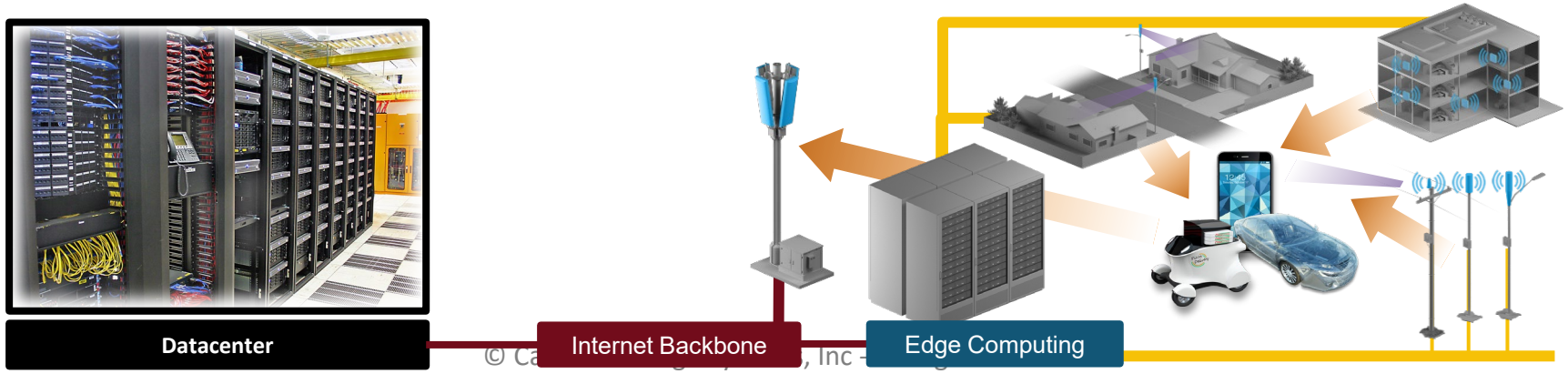
Design & Verification Requirements

Datacenter

- Big chip solutions for GPUs, NOCs, Workload specific AI accelerators
- High capacity verification, Advanced nodes
- Specific models for IP, verification
- Advanced flows Low power with high performance
- Workload optimization server flows (SBSA, …)

Edge

- Specialized IP, extendable processing
- Advanced IP models (verification, implementation)
- Comprehensive low power tool set for Digital and Analog IP



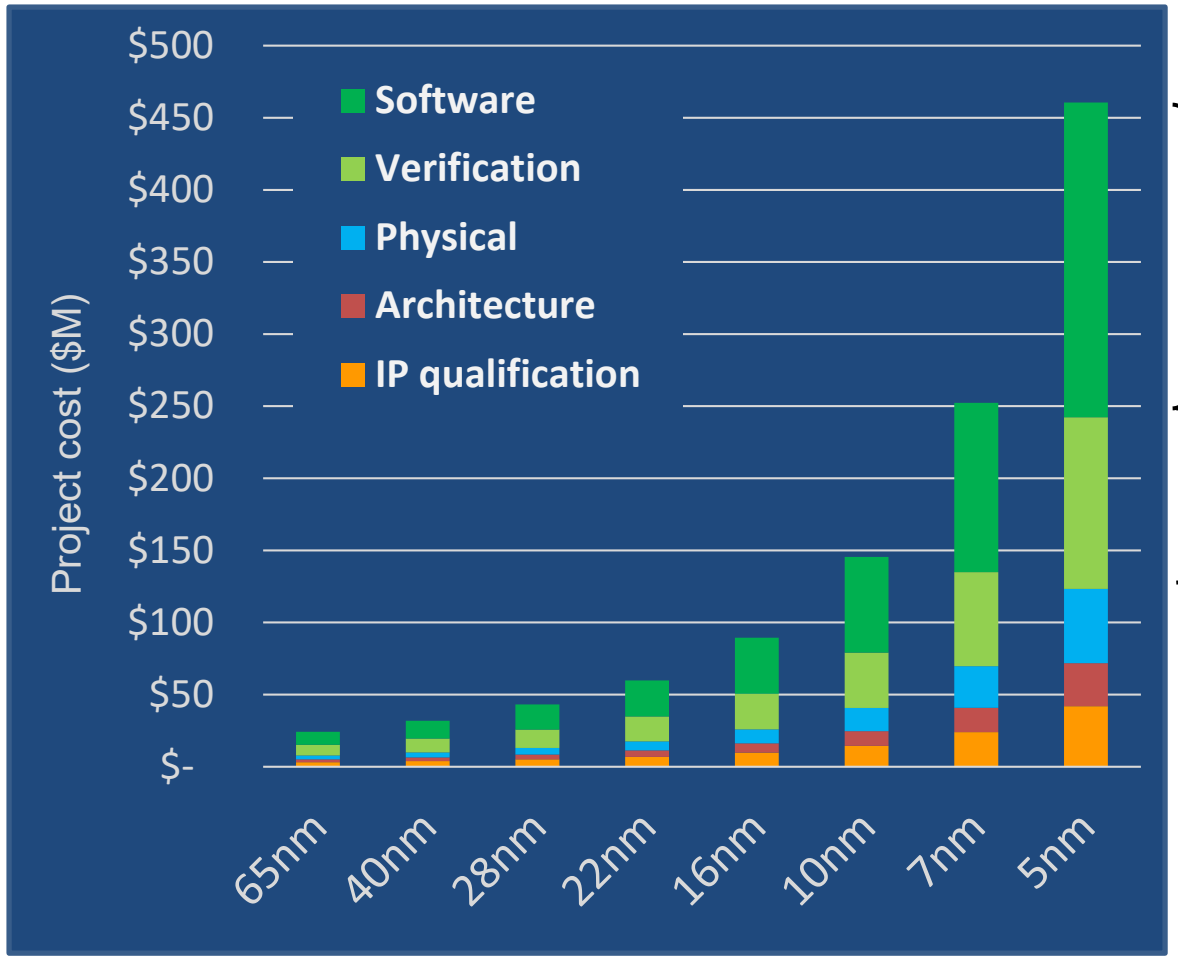
Design and Verification Challenges

- Diverse requirements
  - Training – High throughput, big designs
  - Inference – Flexibility, Lowest power
- Verification Challenges
  - Significant software content
  - Big Designs – Emulation throughput, debug
  - Physical and virtual interfaces , Virtualization
- Physical Design Challenges
  - Complex SystemVerilog Descriptions
  - 1st Floorplan uncertainty
  - Advanced node foundry limits and closure
- System Design Challenges
  - Mix of older and advanced nodes
  - New system design innovations with 2.5D/3D

Example: AI Processor on FPGA Prototype

© Cadence Design Systems, Inc - All Rights Reserved

Unifying Challenge: Verification and Software



Expanding **Software** Challenge

Expanding **Verification** Challenge

VERIFICATION & SOFTWARE

$$2^N$$

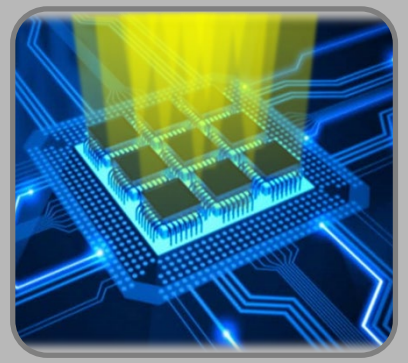
Source: IBS 2018

© Cadence Design Systems, Inc - All Rights Reserved

VERIFICATION OPTIONS TO *ENABLE* 5G/AI/ML

© Cadence Design Systems, Inc - All Rights Reserved

AI/ML Verification Requirements



Formal

- SoC **Scalability**
- **Smart Proof** technology
- Optimized **regressions**

Simulation

- **Fast simulation** for high-activity designs
- UVM **randomization**
- Fast elaboration for **replicated structures**
- Coverage **metrics**

Emulation

- **Billion-gate designs**
- Parallel Partition Compiler
- INT8 to INT64
- **Power / Performance**
- Memory **models**: HBMx
- **Senor model**: MIPI CSI

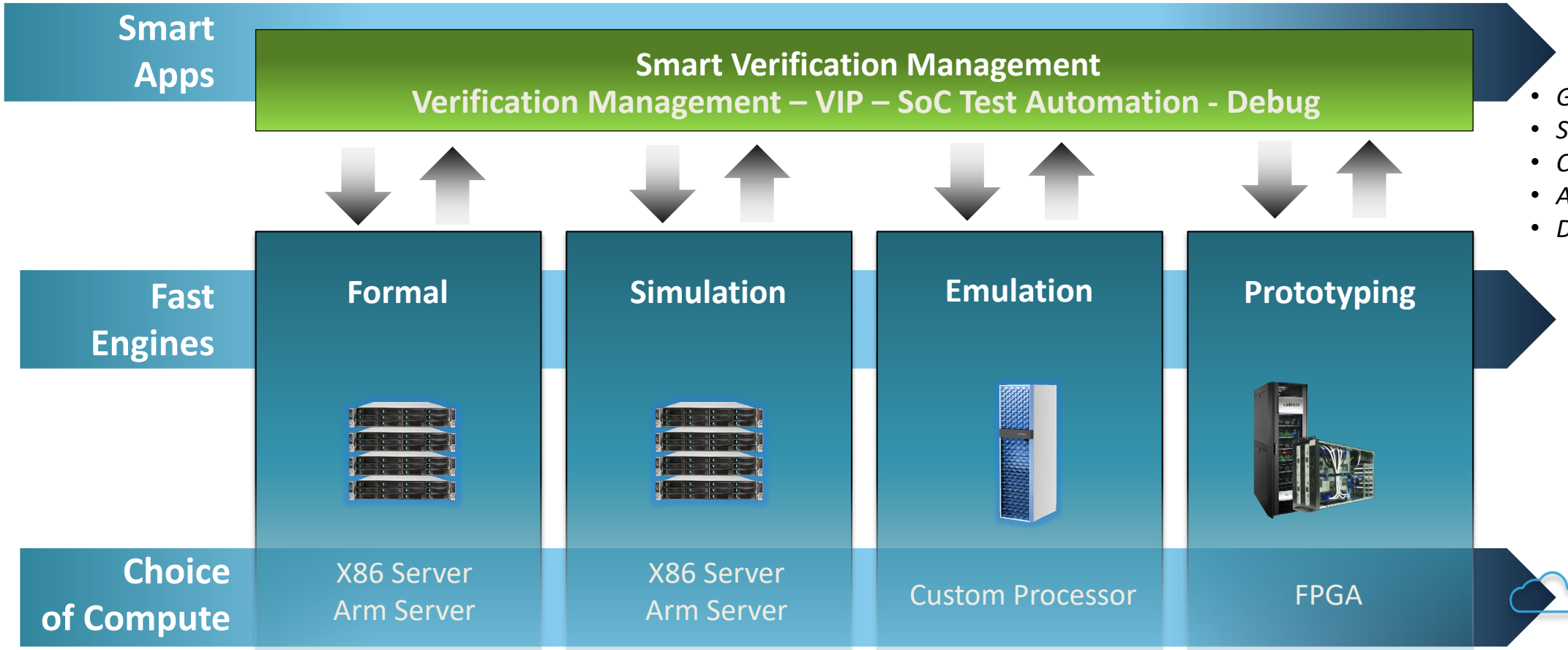
Prototyping

- **Scaling** to large designs
- Unified frontend
- ICE test suite: **Faster, data-extended (DL) regression**
- **SW driven validation** – deep learning data training refinement

© Cadence Design Systems, Inc - All Rights Reserved

Verification Throughput!

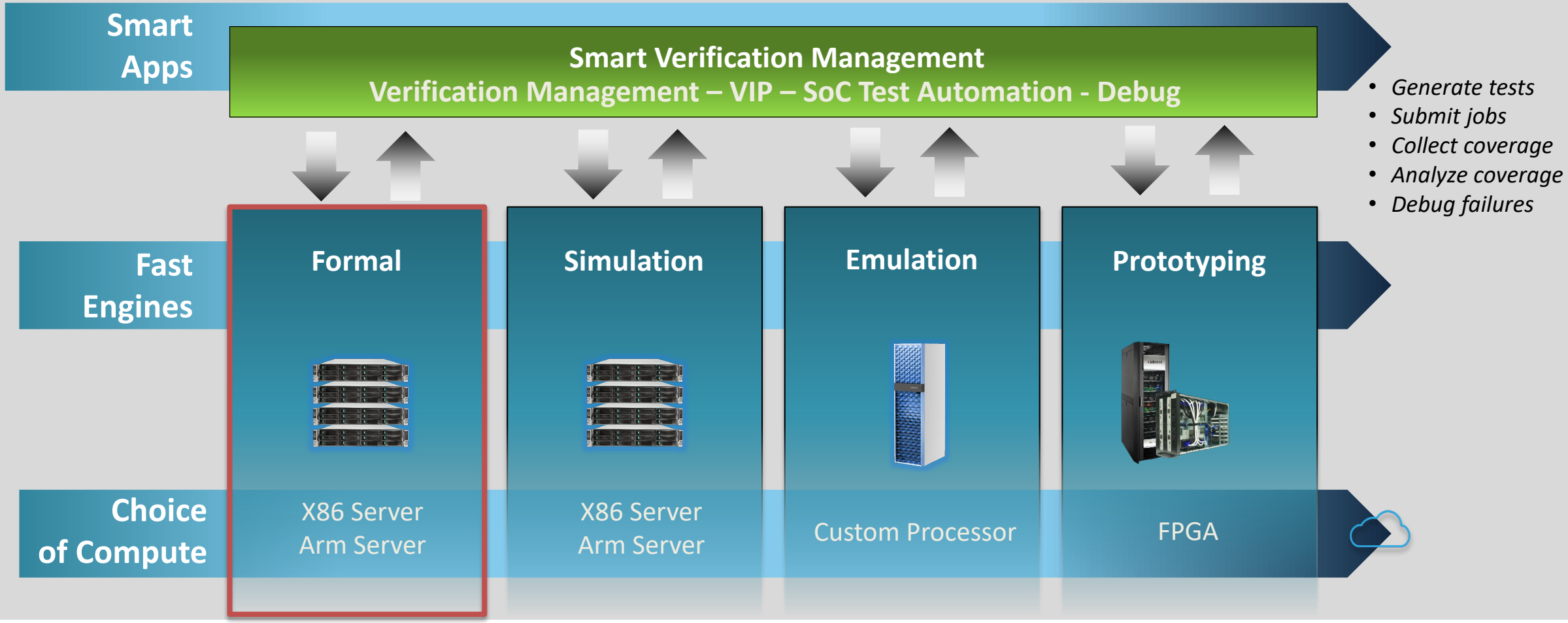
Find and fix the most bugs per $ invested in bare metal compute



- *Generate tests*
- *Submit jobs*
- *Collect coverage*
- *Analyze coverage*
- *Debug failures*

© Cadence Design Systems, Inc - All Rights Reserved

Formal Verification

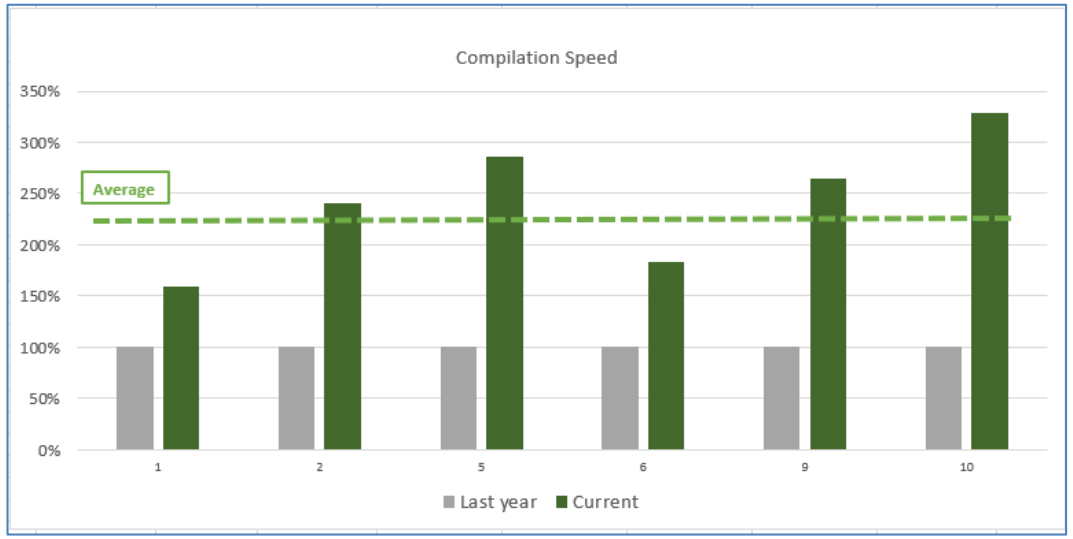


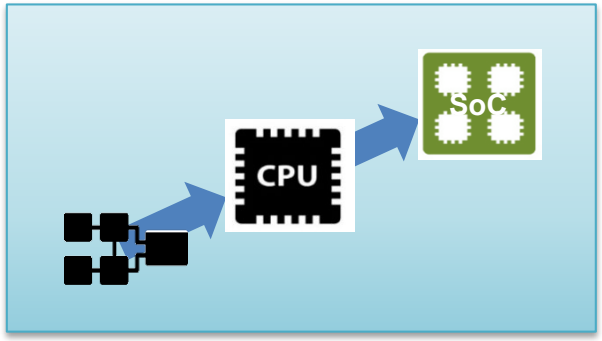
© Cadence Design Systems, Inc - All Rights Reserved

SoC Design Scalability

Compilation Performance

✓ **2X** faster in 2019

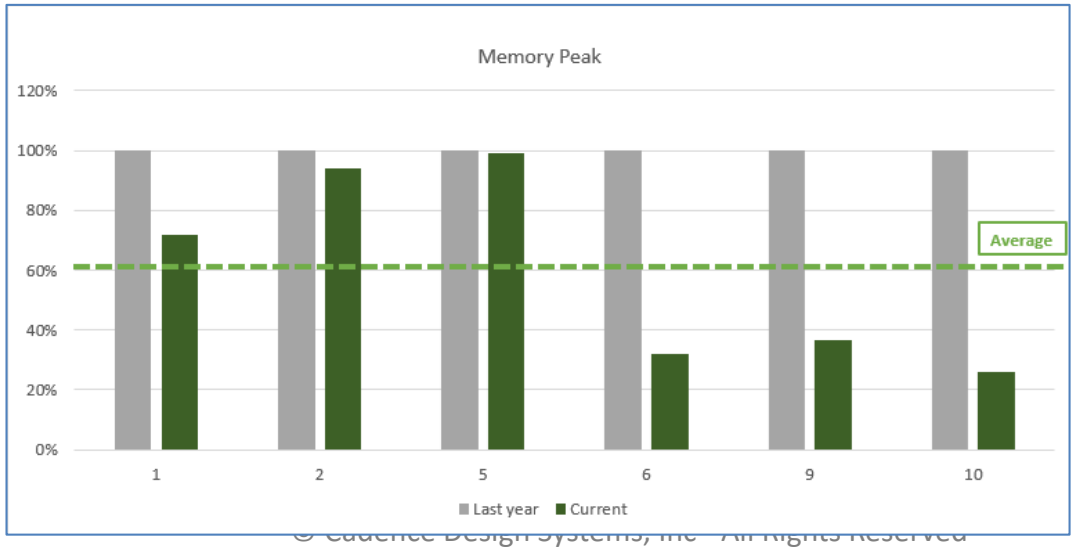




- 6 big designs from 4 different customers
- Min. compilation speed: 100M gates/hour
- Peak memory: less than 0.5 Kbyte/gate

Compilation Memory

✓ **40%** smaller in 2019



© Cadence Design Systems, Inc. All Rights Reserved

Smart Proof Automation Framework

Component & Data Management

Proof Profiling Data

- Regular read/write

Proof Caching

- Cache storage in single file
- Automatic cleanup of old cache data

Multi Advisor Proof Orchestration

- Forced `on`
- No overwrite on engine mode

Engine Algorithm Selection

- Automatic training and inference

Learning

Machine Learning

Optimizes subsequent runs/regressions

Optimizes out-of-the-box proofs

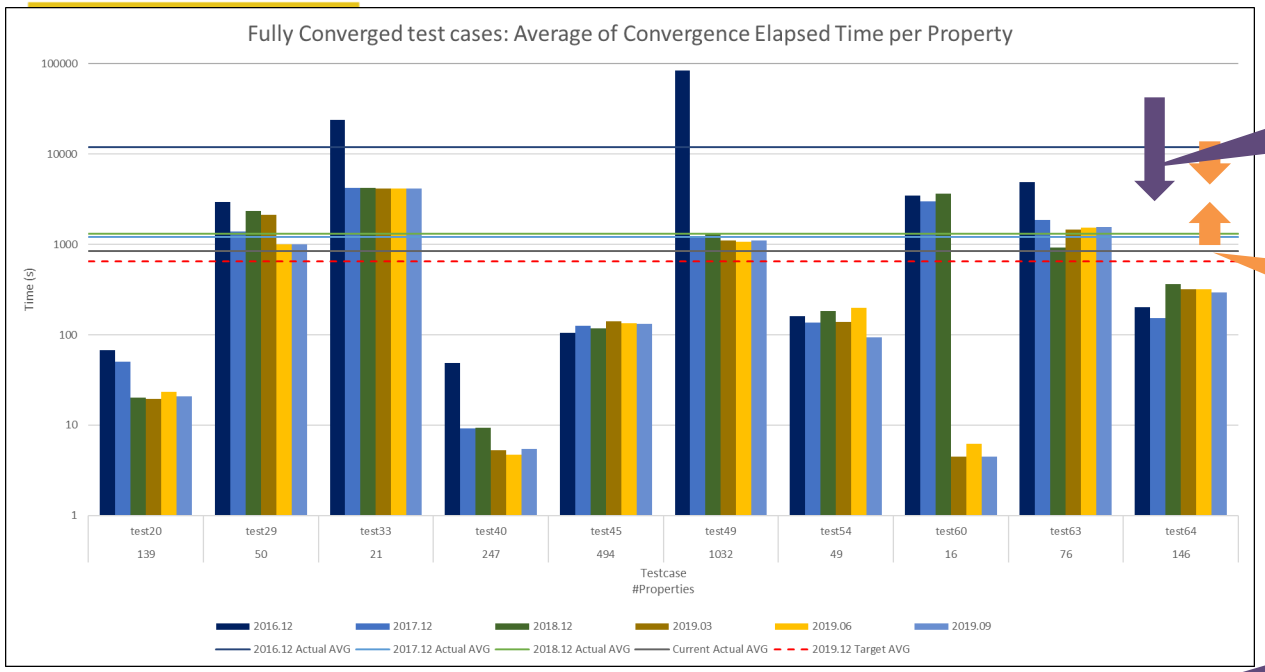
Find more bugs

Better convergence

Faster proofs

© Cadence Design Systems, Inc - All Rights Reserved

Property Performance and Convergence

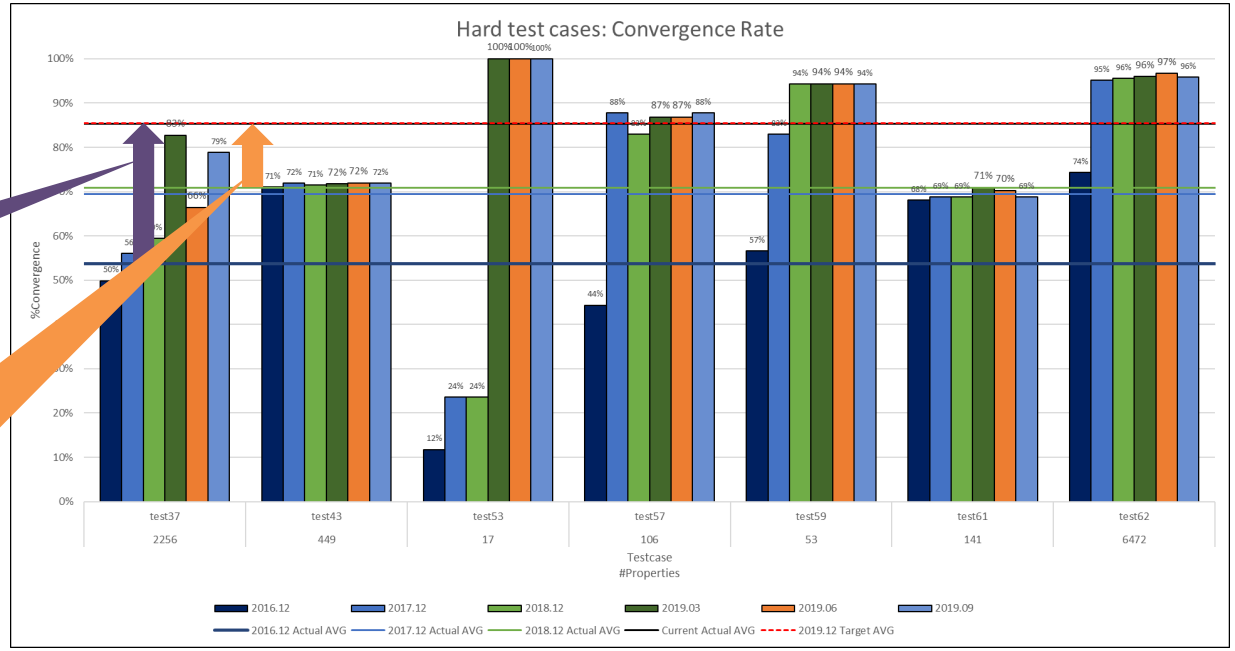


Avg. **14X** speedup in last 3 years

Avg. **1.6X** speedup in 2019

Core Engine Performance

- ✓ 10 fully-converging designs
- ✓ **14X** speedup last 3 years
- ✓ **1.6X** speedup in 2019



Avg. **3.2X** non-converged reduction in last 3 years

Avg. **2X** non-converged reduction in 2019

Core Engine Proof Success

- ✓ 7 "hard" designs: significant # undetermined properties
- ✓ **2X** reduction in undetermined properties in 2019

© Cadence Design Systems, Inc - All Rights Reserved

Proof profiling data (PPD) and proof caching

- Challenge
  - Running engines to reproduce previous tool results can often be extremely costly in terms of resources, especially in cases where the design/environment has minimally changed or not changed at all
  - Need ability to learn from past runs of a design, to optimize subsequent proofs
- Approach
  - Keep record of previous proof strategies that worked, so they can be tried again
    - `prove -save_ppd -with_ppd`
  - Be able to harvest previous proof results, when confirmed that they still apply
    - `set_prove_cache on`
- Benefits
  - Better convergence: faster proofs on properties determined before frees up resources to work on other properties
  - Smarter use of machine time
  - Helps with reproducibility of proof results

© Cadence Design Systems, Inc - All Rights Reserved

Proof profiling data (PPD)

- Saves knowledge collected during a prove command to be used as recommendations in subsequent runs

JasperGold®

	Type	Name	Engine	Bound	Time
✔	Cover	state_23	Q3	17 - 23	10.9
✔	Cover	state_24	U	22 - 24	80.8
?	Assert	property:0	B	25 -	0.3
?	Cover	all_positions	Bm	23 -	128.6
✔	Cover	all_but_central	L	23 - 24	97.0

prove –save_ppd

jgproject/sessionLogs/session_0/sessionPPD.ppd

JasperGold®

prove –with_ppd

```
INFO (IPF149): Starting PPD exploitation on 36 matching properties
```

"Best engines" orchestration

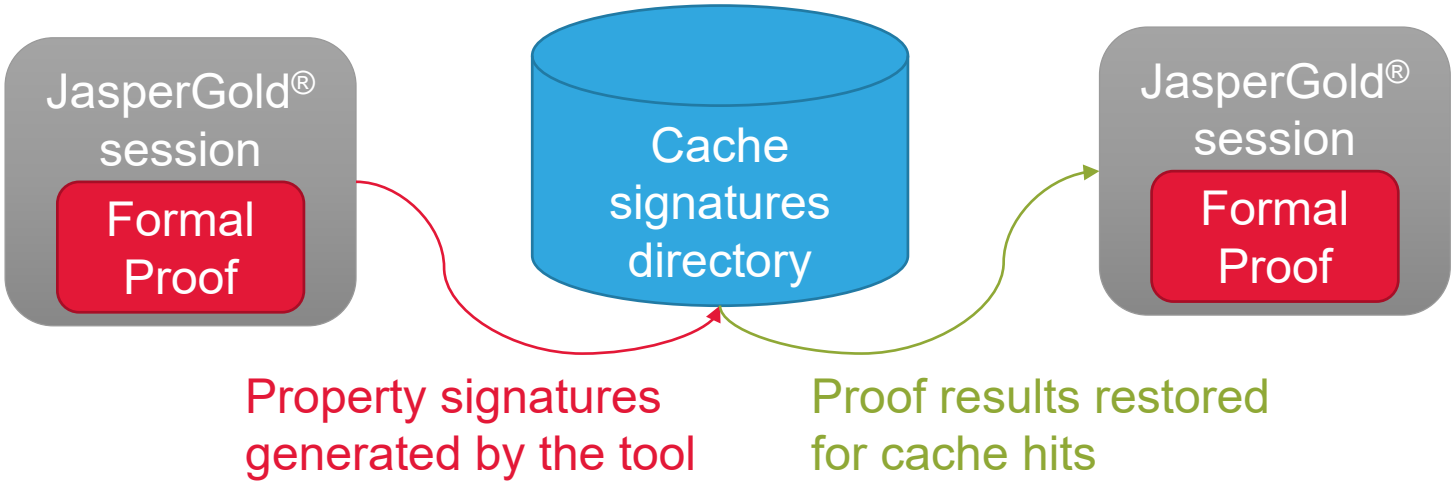
Regular orchestration kicks when results in PPD fail to be reproduced

- Benefits
  - Increase the chances of converging faster on properties determined before, and that did not have their results

© Cadence Design Systems, Inc - All Rights

Proof caching

- Restores proof results based on properties' signatures
- Current version sensitive to small changes to design and environment

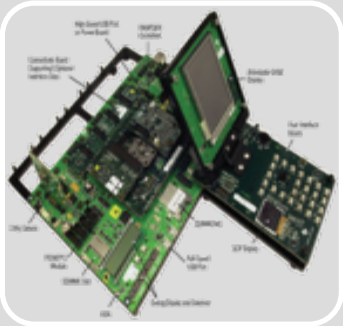


- Benefits
  - Save engine time when processing unchanged properties already determined before
  - Focus resources on properties that changed or that were never determined, improving convergence

© Cadence Design Systems, Inc - All Rights Reserved

A Continuum of Dynamic Engines

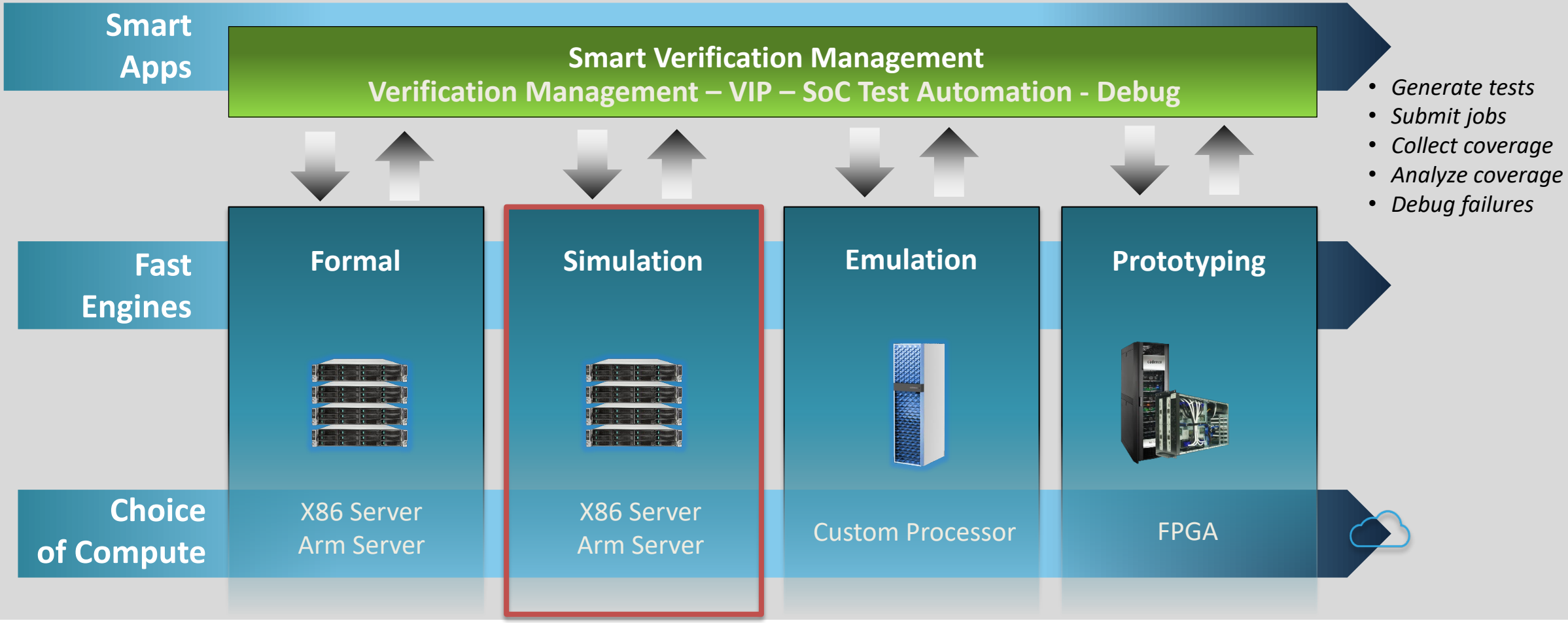
Verification and software platforms need to interoperate



SDK OS Simulation	Virtual Platform	HDL Simulation	Acceleration Emulation	FPGA Prototype	Prototyping Board
Highest speed	Almost @ speed	KHz Range	MHz Range	10's of MHz	Real time speed
Earliest in flow	Pre-RTL	Early RTL	Early RTL	Later RTL	Fully accurate
Ignores HW	Less accurate	Golden Reference	Min RTL mods	Some RTL mods	Actual Silicon
Easy replication	TLM HW Debug	Best HW debug	Detailed HW debug	Some HW debug	Difficult HW debug
Cross-compile	Great SW debug	Limited SW Debug	Great SW Debug	Great SW Debug	OK SW Debug
	Easy replication	Easy replication	Harder to replicate	OK to replicate	Easy to replicate
	Less HW detail	Mixed-abstractions	Datacenter access	Harder Bring-up	HW changes hard
	Slower with detail	Slow SW execution	Contested Resource		

© Cadence Design Systems, Inc - All Rights Reserved

Simulation



© Cadence Design Systems, Inc - All Rights Reserved

Simulation Performance

COMPILE Incremental Build Parallel Build Hierarchical build w/ Cloning *Up to 10X speed-up*

FULL REGRESSION THROUGHPUT

Relentless focus on performance

Continuous Core Performance Enh's

Save/Restart w/ dynamic test reload

LONG TEST LATENCY

Rocketick multi-core technology

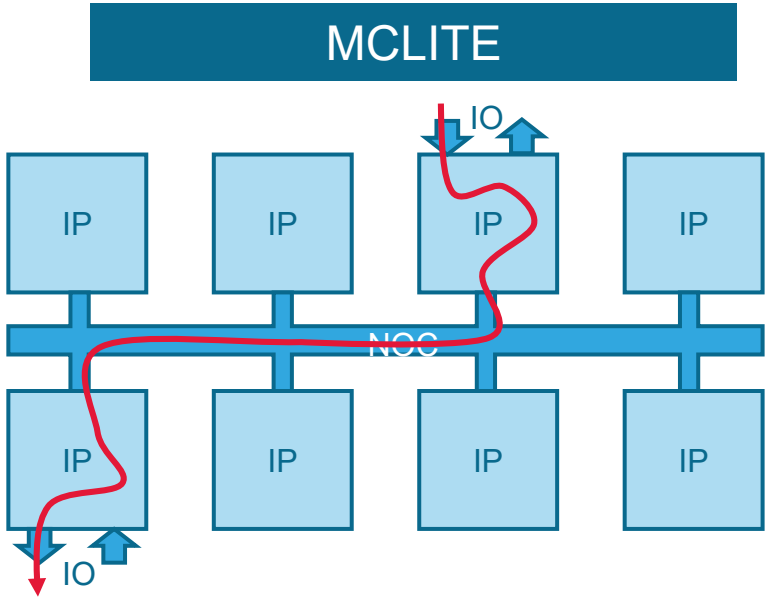
MC-Lite 1.2-1.8X

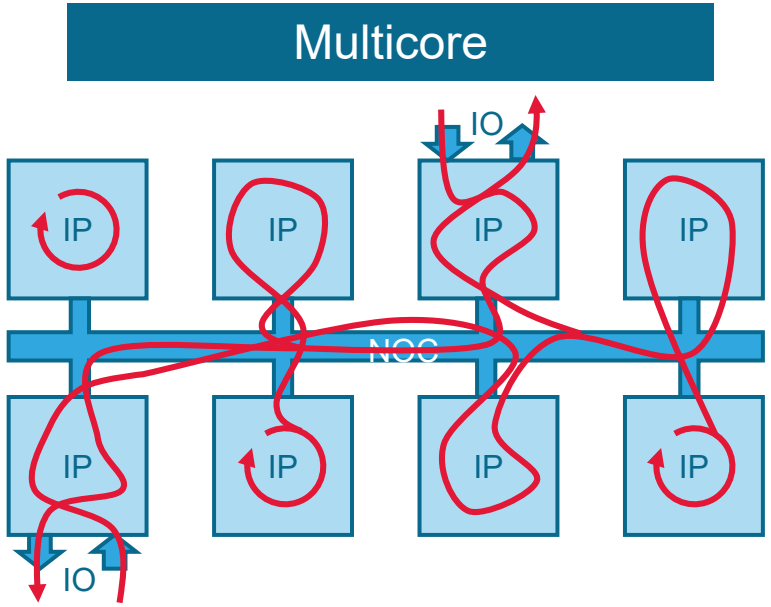
MC-Rocketick 3-5X

COMPUTE **X86** **Arm** **Cloud**

© Cadence Design Systems, Inc - All Rights Reserved

Multi-Core Performance





Can use on any long running test

- Uses the single-core build
  - Decision to use mclite deferred to runtime
- Same as single-core scheduling
  - Guaranteed congruent results with SC
- **1.2 – 1.8X gains**

Requires "high activity" DUT-heavy test

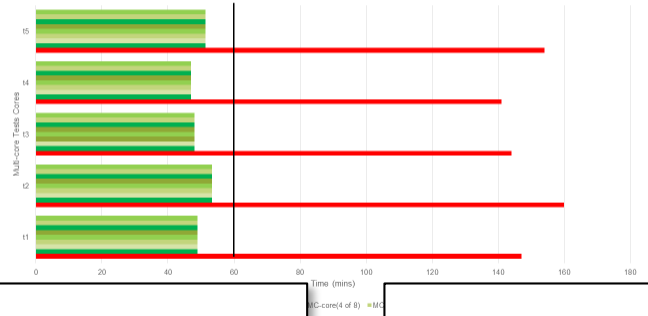
e.g. Gate Level ATPG

- Highly scalable parallel solution
  - Needs a special build
- Specialized (Rocketick) engine
  - Non-accelerable parts (testbench) on SC
- **3-5X gains**

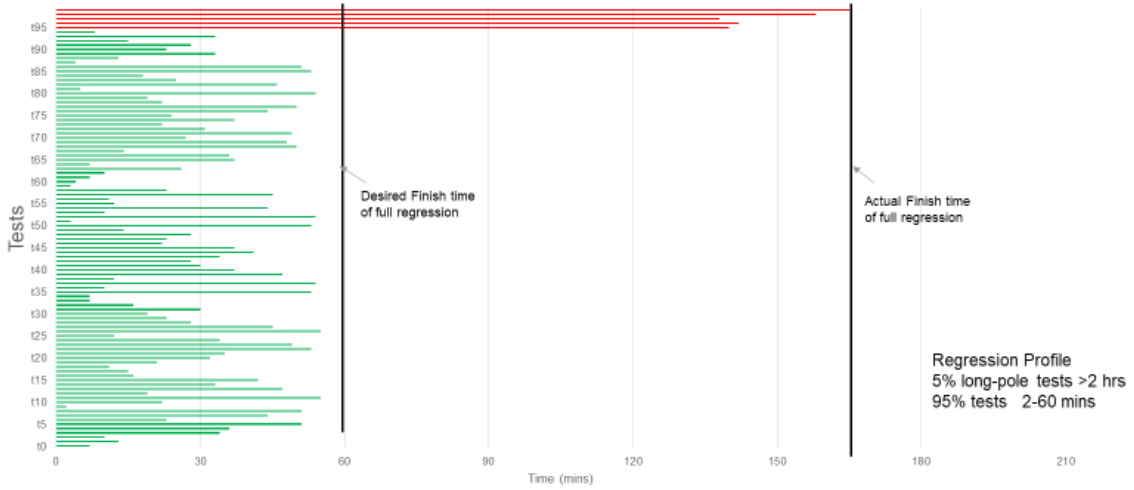
© Cadence Design Systems, Inc - All Rights Reserved

Multi-Core and Simulation Regressions

Multi-Core Simulation: 3X faster using 8-cores



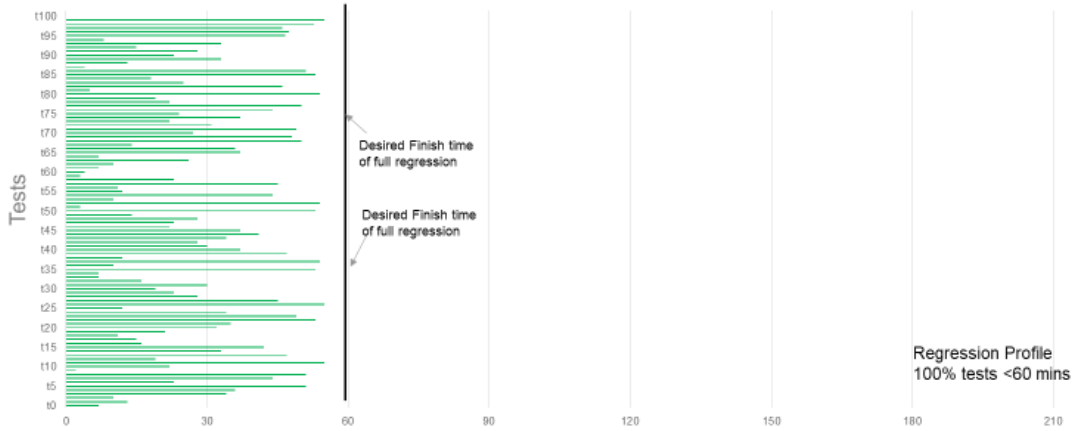
Simulation Regression: Xcelium Single-Core



20 © 2019 Cadence Design Systems, Inc. All rights reserved. cadence

Simulation Regression: Multi-Core 3X on 5% Long-pole tests.

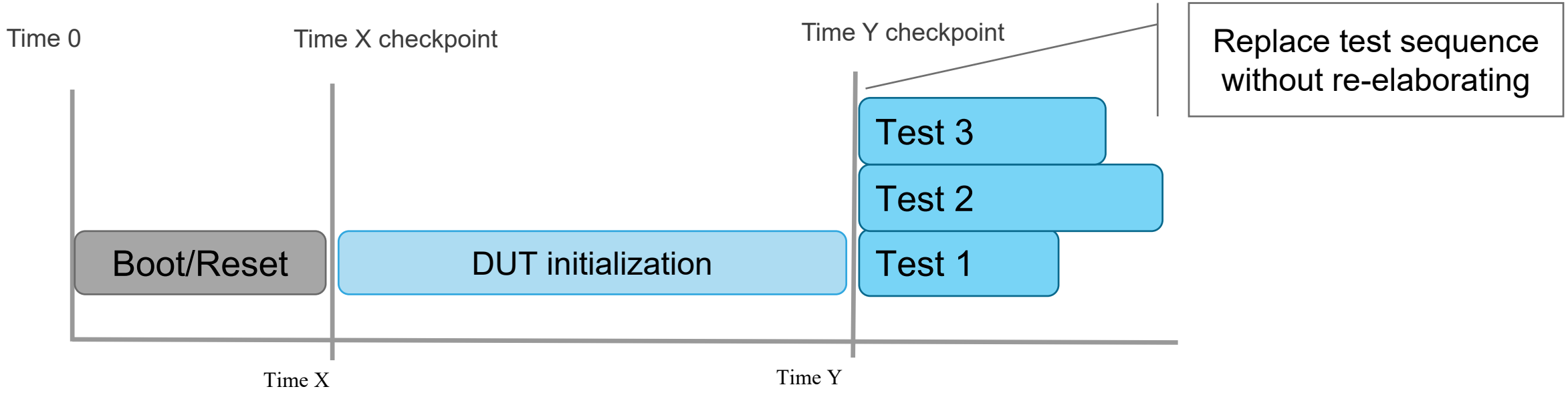
Actual finish time of regression = Desired finish time of regression



22 © 2019 Cadence Design Systems, Inc. All rights reserved. cadence

© Cadence Design Systems, Inc - All Rights Reserved

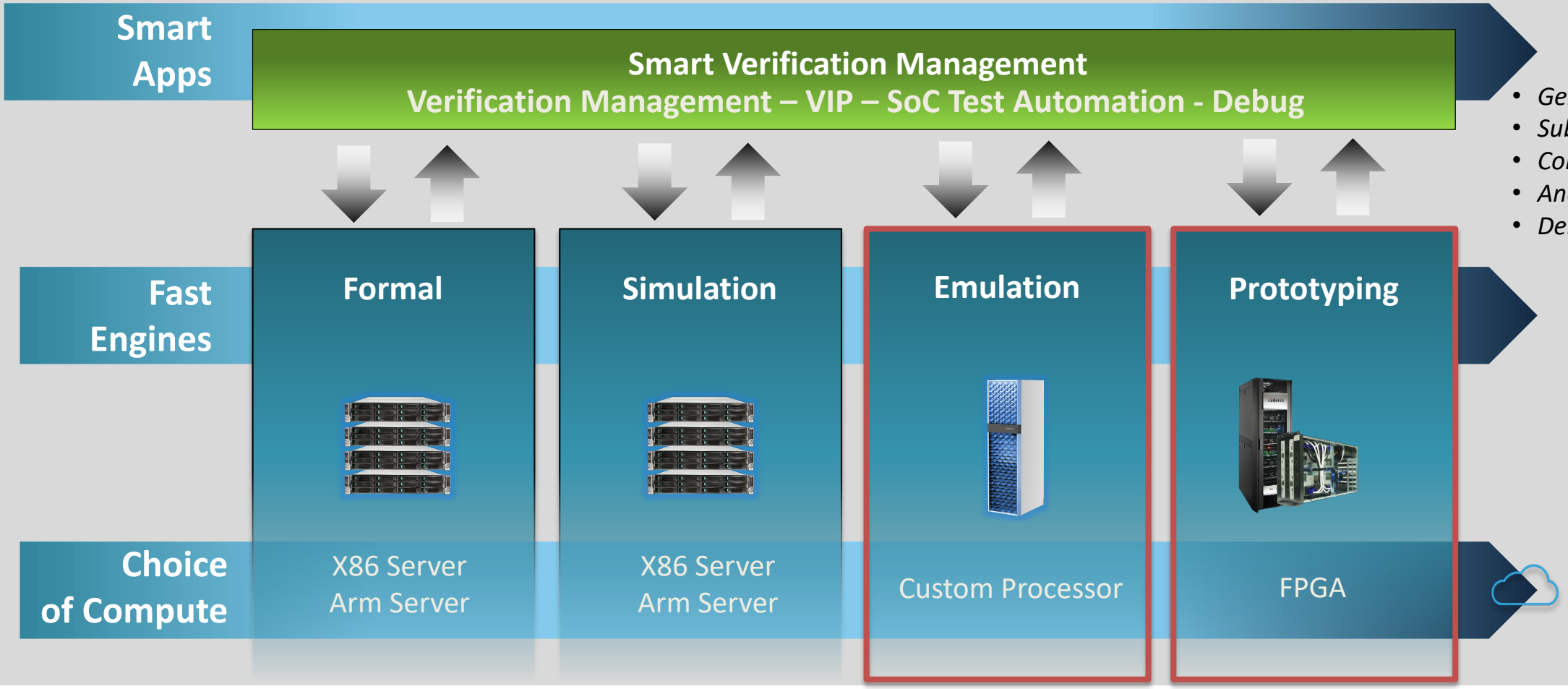
Dynamic Test Reload for SystemVerilog



- SystemVerilog/UVM Dynamic Test Load
  - Load new SystemVerilog package into saved snapshot
  - Call testbench functions when the snapshot is reloaded
- Dynamic Base Snapshot: Time zero snapshot or saved snapshot
- Dynamic Test Snapshot: Contains the incremental new SystemVerilog package

© Cadence Design Systems, Inc - All Rights Reserved

Hardware Assisted Development



- *Generate tests*
- *Submit jobs*
- *Collect coverage*
- *Analyze coverage*
- *Debug failures*

© Cadence Design Systems, Inc - All Rights Reserved

Emulation And Prototyping

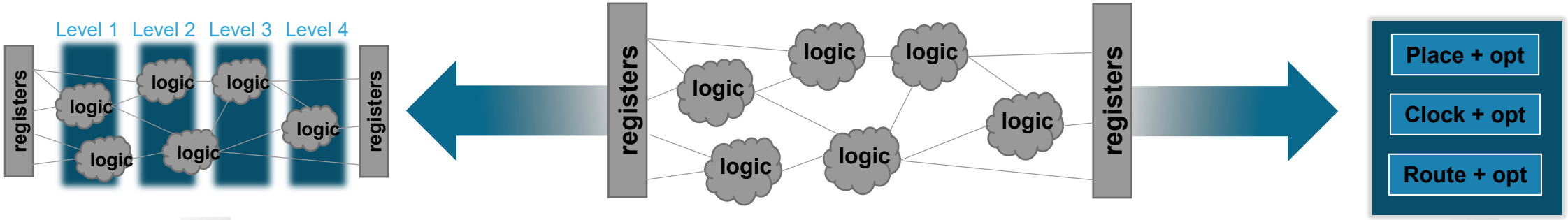
	Key Care Abouts
Emulation ***"Debug your design"***	• **Predictable fast build**: Rebuild ne • **SOC level capacity**: IP level sim ough already • **Fast and complete debug:** ls
Prototyping ***"Debug your Software"***	• **Build time and debug less important**: desig 1-2 weeks • **Highest performance**: software debug tion runs • **Lowest cost**: replicate one build fo lopers • **SOC level capacity**: more and more so at C level

Fast RTL verification & debug

Early SW development & HW regressions

© Cadence Design Systems, Inc - All Rights Reserved

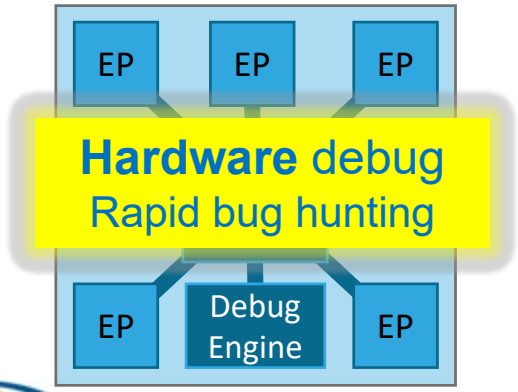
Processor Based Emulation and FPGA Based Prototyping



Simple left-to-right stream processing

ASIC-style full Place & Route

Emulation Processor



1MHz

Fast compile

Predictable compile ("If it compiles it runs")

"Full Vision" debug

Xilinx FPGA

5-10MHz

Slower compile

Compile may need tuning to close timing/routing violations

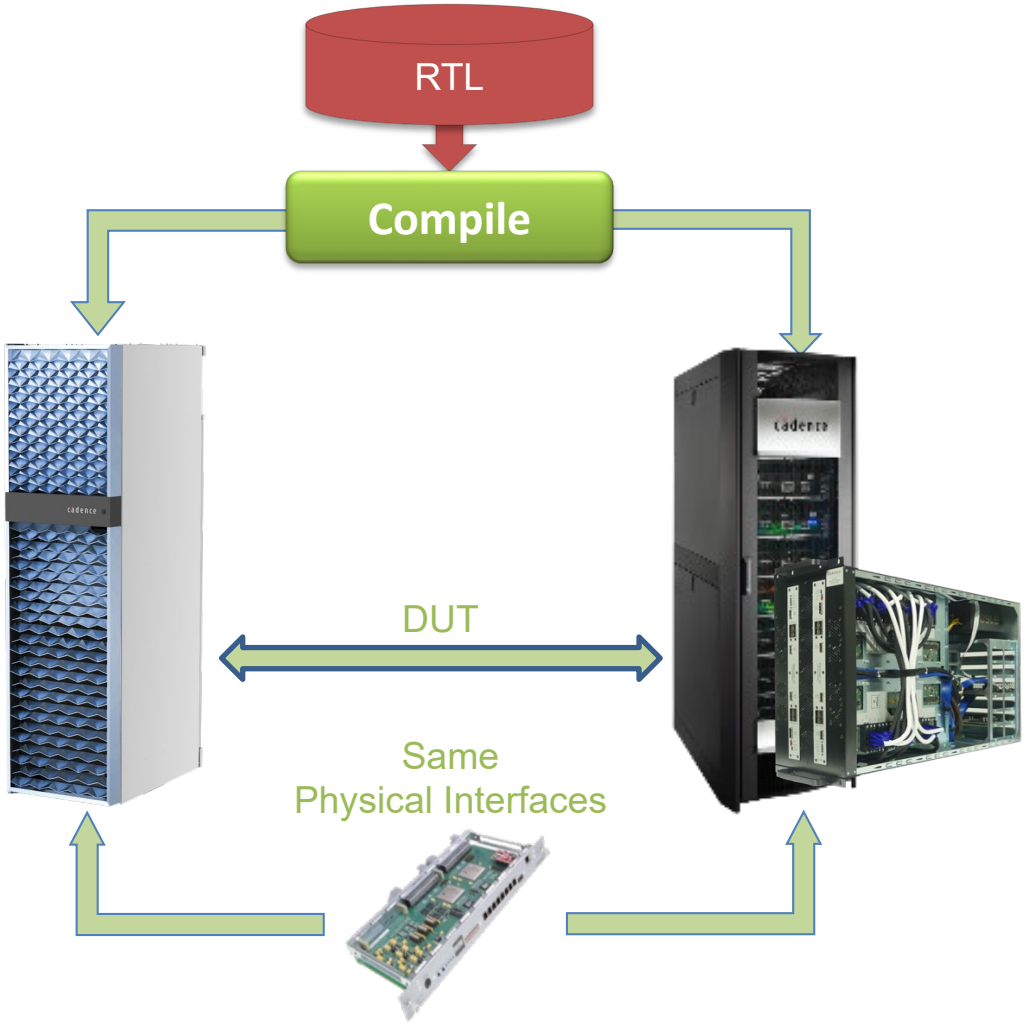
Less Flexible debug

© Cadence Design Systems, Inc - All Rights Reserved

Dynamic Duo: Emulation & Prototyping

Emulation

- Optimized HW/SW debug
- SoC acceleration, HW/SW
- Power & Performance Analysis
- Advanced Use Models



Prototyping

- Automated Bring-Up
- Scalable performance
- SW development
- HW/SW regressions

Congruency and **common environment**

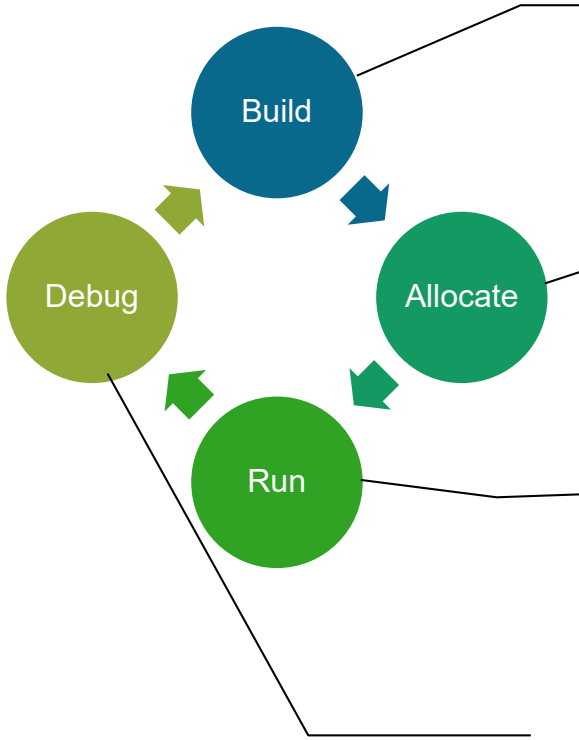
© Cadence Design Systems, Inc - All Rights Reserved

Emulation Capabilities

- Palladium® Z1 **enterprise** emulation platform
  - Up to **5X** greater emulation **throughput**
- **Scalability** from IP blocks to full systems on chip
  - Capacity of up to **9.2 billion gates** with **2304 users**
- Best in class **total cost of ownership** (TCO)
  - **22 use models**
- New era of **datacenter-class** emulation
  - **Proven reliability**

© 2017 Cadence Design Systems, Inc. Cadence confidential – Update for ARM – October 10th 2017
© Cadence Design Systems, Inc - All Rights Reserved

HW-assisted verification productivity loop



Compile databases for different workloads
(Compile: speed, automation, # of workstations)

Allocate as many workloads as possible
(Utilization efficiency: # of parallel jobs, relocation)

Run workloads based on priorities
(Speed: use models, interface solutions)

Debug for both pre & post silicon bugs
(Visibility: trace depth, dynamic trigger)

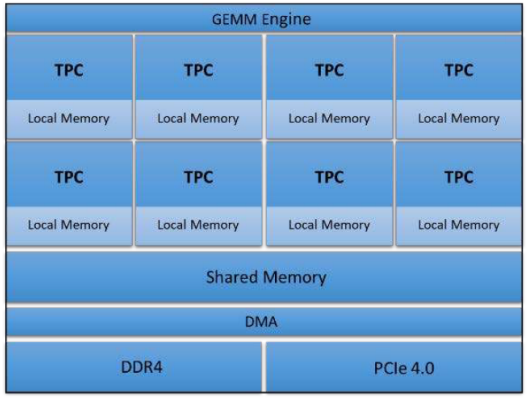
© Cadence Design Systems, Inc - All Rights Reserved

Billion Gate Design Examples

Habana Labs
Billion-gate class inference processor: Goya

Goya Processor Architecture

- Heterogenous compute architecture
  - 3 Engines: TPC, GEMM and DMA
  - Work concurrently using a shared SRAM
- Tensor Processor Core (TPC™)
  - VLIW SIMD vector core
  - C-programmable
- GEMM operations engine
- Tensor addressing
- Robust to any address stride
- Latency hiding capabilities
- PCIe Gen4.0 x16
- 2 DDR4 channels @ 2.667 GT/s, 40GB/s BW, 16GB capacity
- Dedicated HW and TPC ISA for special functions acceleration (e.g. Sigmoid/GeLU, Tanh)
- Mixed-precision data types: FP32, INT32, INT16, INT8, UINT32, UINT16, UINT8



TSMC – 16nm

Habana Labs Hot Chips 2019 6

Source: Hotchip Conference

Fujitsu
Billion-gate class HPC/AI Processor: A64FX

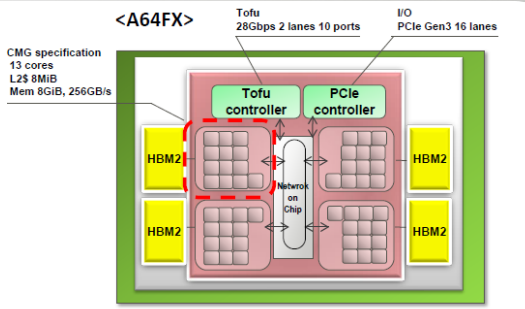
A64FX Chip Overview

FUJITSU

Architecture Features
- Armv8.2-A (AArch64 only)
- SVE 512-bit wide SIMD
- 48 computing cores + 4 assistant cores*
  *All the cores are identical
- HBM2 32GiB
- Tofu 6D Mesh/Torus 28Gbps x 2 lanes x 10 ports
- PCIe Gen3 16 lanes

7nm FinFET
- 8,786M transistors
- 594 package signal pins

Peak Performance (Efficiency)
- >2.7TFLOPS (>90%@DGEMM)
- Memory B/W 1024GB/s (>80%@Stream Triad)



	A64FX (Post-K)	SPARC64 XIfx (PRIMEHPC FX100)
ISA (Base)	Armv8.2-A	SPARC-V9
ISA (Extension)	SVE	HPC-ACE2
Process Node	7nm	20nm
Peak Performance	>2.7TFLOPS	1.1TFLOPS
SIMD	512-bit	256-bit
# of Cores	48+4	32+2
Memory	HBM2	HMC
Memory Peak B/W	1024GB/s	240GB/s x2 (in/out)

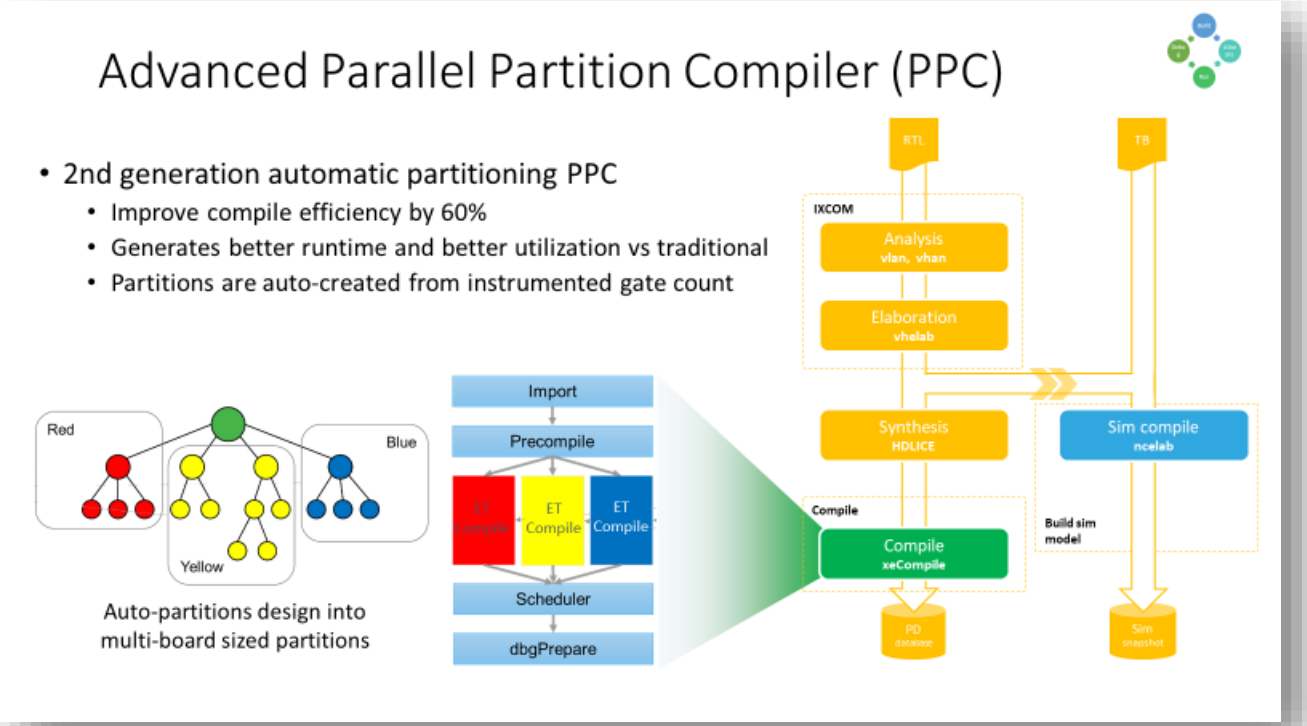
6 All Rights Reserved. Copyright © FUJITSU LIMITED 2018

Source: Hotchip Conference

© Cadence Design Systems, Inc - All Rights Reserved

Emulation Model Scaling: Enabling AI/ML/5G

- Scaling from IP verification to multi-core/die emulation
  - 4 million gates to 7 billion gates
- HW/SW co-verification without performance impact
  - FullVision, waveform streaming, dynamic RTL, physical and virtual JTAG debuggers
  - Supports 3rd party 5G testers
- Emulation in the cloud eases variability of workload sizes and use models



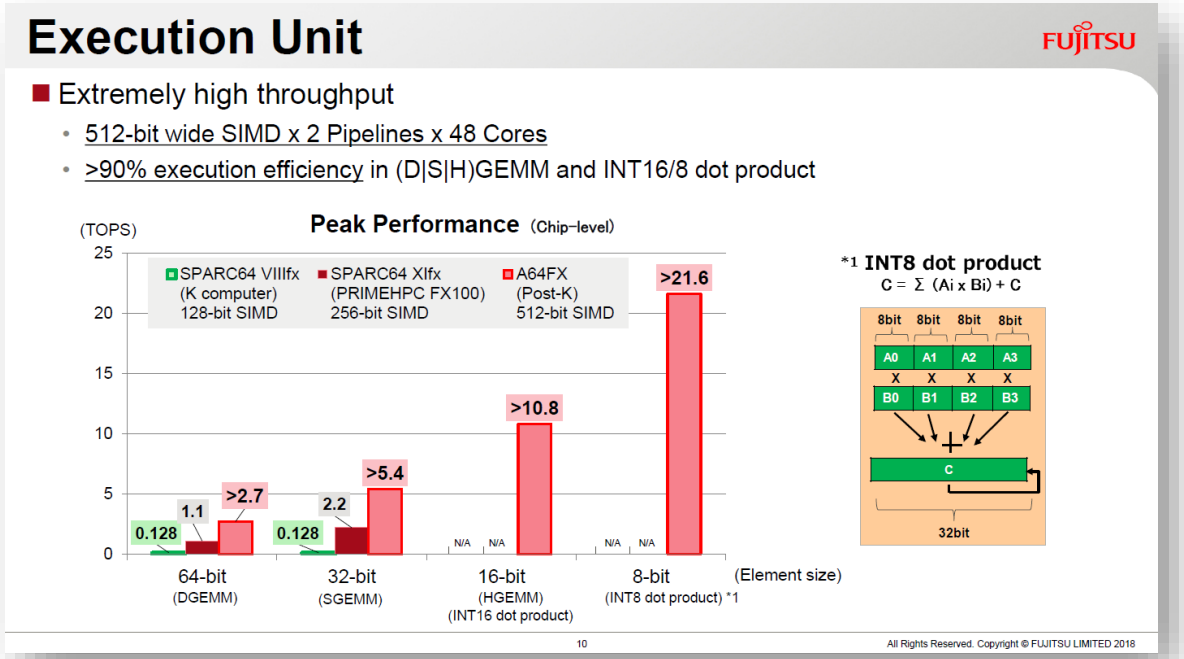
- Parallel Partition Compiler enables one billion gate emulation model to be compiled in about 4 hours
- 2nd generation PPC-based emulation models are fully automated – no need for manual partitioning
- PPC enables practical scaling of emulating billion-gate class AI/ML and 5G designs

© Cadence Design Systems, Inc - All Rights Reserved

INT64 to INT8 Computational Efficiency

Fujitsu

Billion-gate class HPC/AI Processor: A64FX



Source: Hotchip Conference

- For certain ML applications, accuracy may be traded off for faster and more power efficient implementation methods
- Depending on the target applications, design may scale down (INT8) or scale-across (INT 8 to INT 64) architecturally to ensure computational efficiency

© Cadence Design Systems, Inc - All Rights Reserved

Power / Performance Trade-offs

Fujitsu

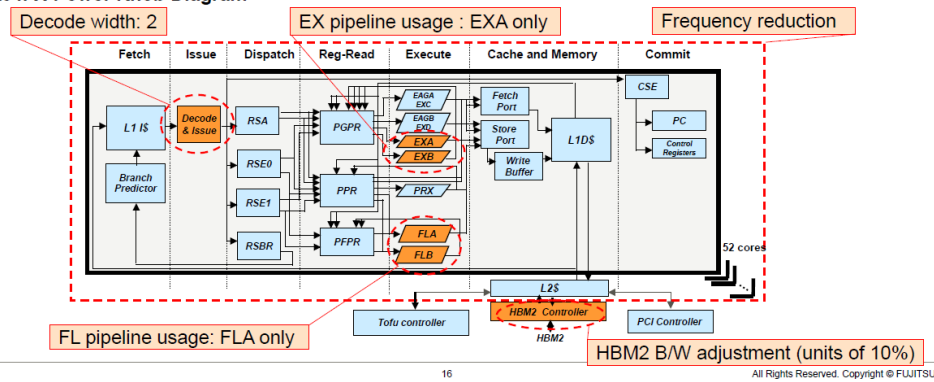
Billion-gate class HPC/AI Processor: A64FX

Power Management (Cont.)

FUJITSU

- "Power knob" for power optimization
  - A64FX provides power management function called "Power Knob"
    - Applications can change hardware configurations for power optimization
  - → Power knobs and Energy monitor/analyzer will help users to optimize power consumption of their applications

<A64FX Power Knob Diagram>



16

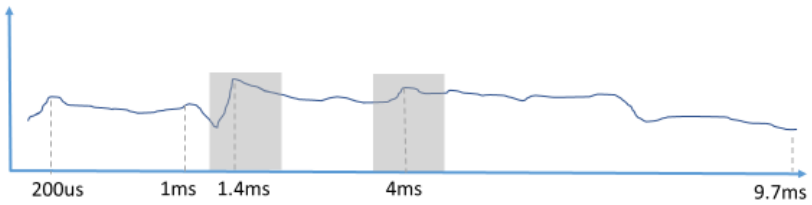
All Rights Reserved. Copyright © FUJITSU LIMITED 2018

Source: Hotchip Conference

Customer Example

HW Toggle Count Case Study

Customer Example 1



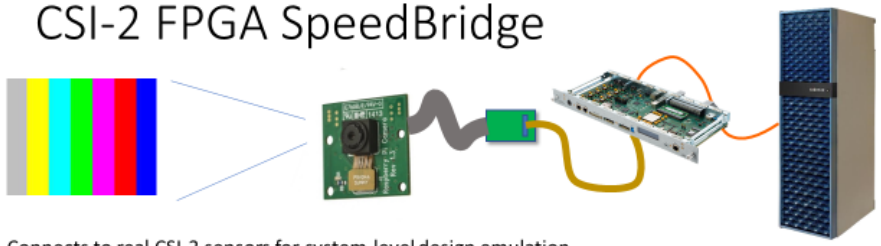
- Pre-silicon peak detection is critical in analyzing and correlating power and performance
- HW-WTC is 304x faster compared to SW-WTC (34.7 minutes vs 7 days) – enabling user to detect peaks faster and perform system analysis and design optimization

Source: Cadence

© Cadence Design Systems, Inc - All Rights Reserved

Senor Models – MIPI CSI-2

CSI-2 FPGA SpeedBridge

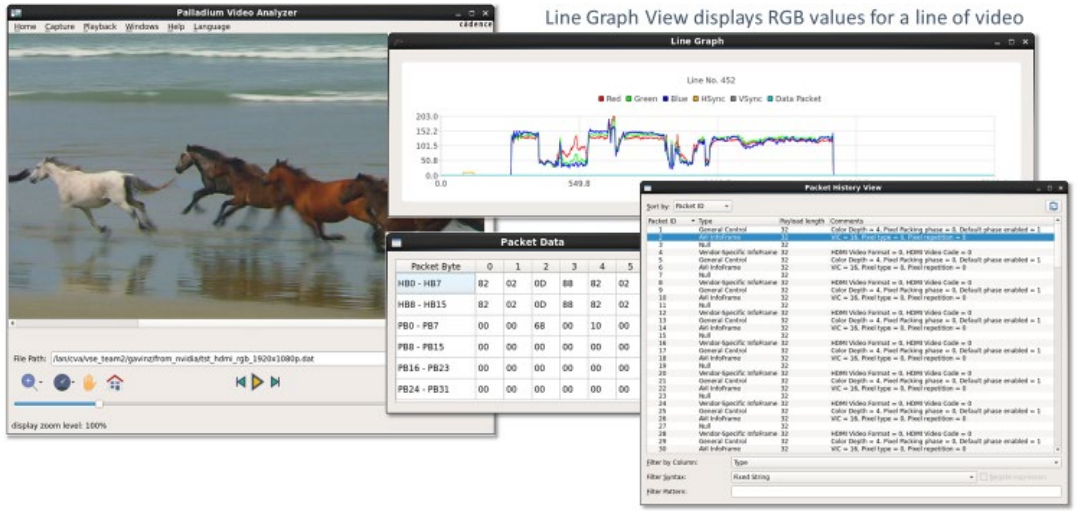


- Connects to real CSI-2 sensors for system-level design emulation
- Sensor video is scaled to emulation speed
- Users can verify their DUT with the actual sensor
  - Vendor specific register spaces, image processing functions, etc
  - Vendor specific embedded data in image packet stream
  - Real sensor behavior
- Daughtercards provide connectivity to different types of camera modules

SpeedBridge Capabilities:
- One sensor per SpeedBridge board
- CSI-2 RX only
- Sensor side up to 4 CSI-2 lanes
- Support max speed of 1.5Gbps
- Emulation side PPI interface up to 4 lanes
- Emulation side CCI interface
- Compatible with Palladium Video Analyzer for CSI stream viewing

Palladium Video Analyzer Views

Line Graph View displays RGB values for a line of video

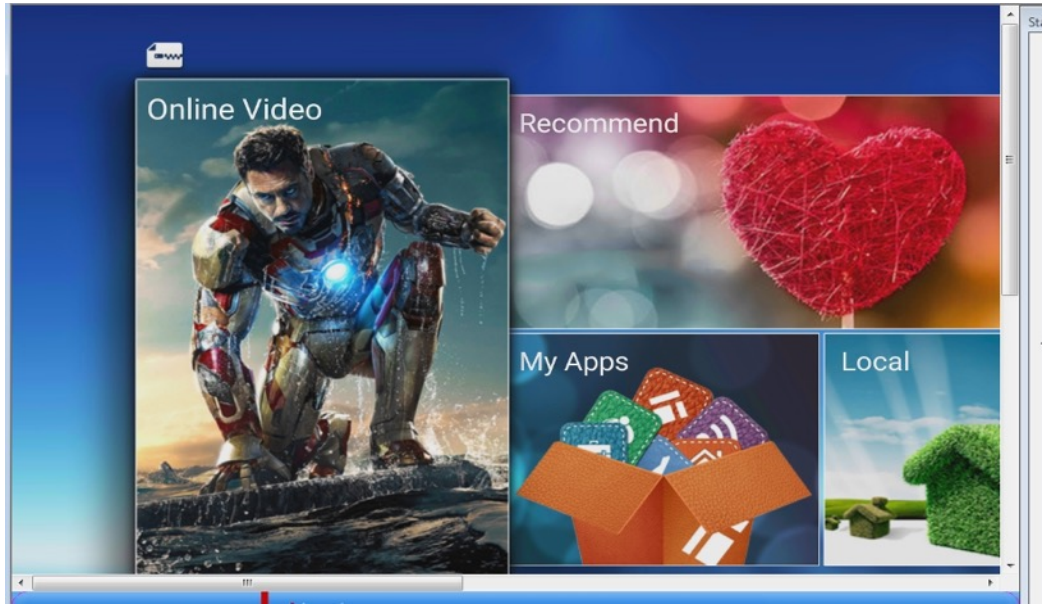


- Emulation with real CSI-2 sensor provides realistic live image capture and processing
- Allows user to conduct visual inspection (e.g. fast forward and replay)
- Video frames can be captured for detailed analysis (e.g. data packet, line graphs, etc.)

© Cadence Design Systems, Inc - All Rights Reserved

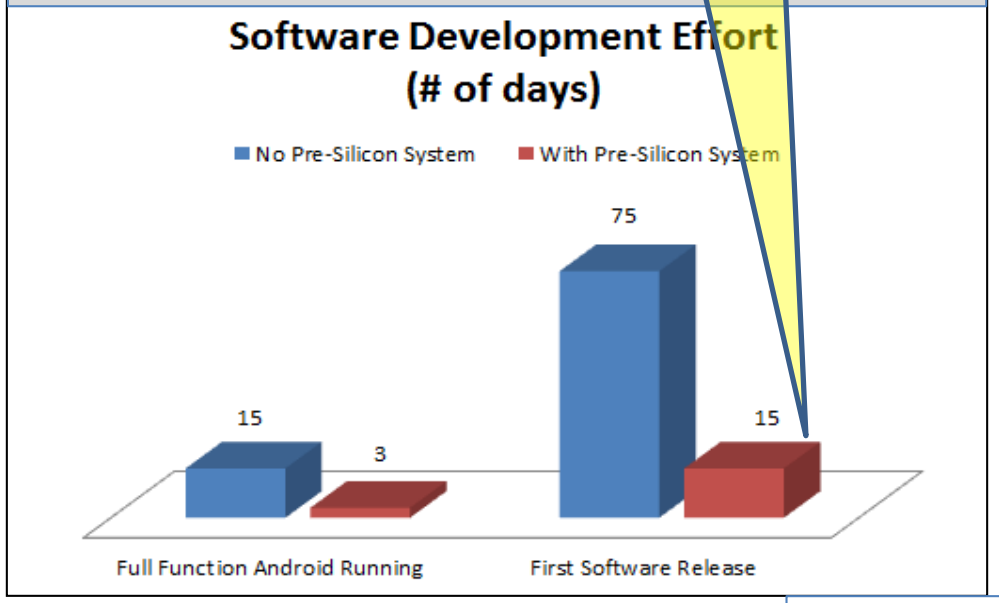
FPGA-based prototyping accelerates time to revenue

- Early, embedded software & firmware development
- Initial systems and/or proof of concept
- Pre-silicon chip (ASIC) verification



2 months faster time to market!

Software is ready when 1st silicon comes in
- Full Android boot 30 min after silicon is back
- Full system demo to customer in 3 days



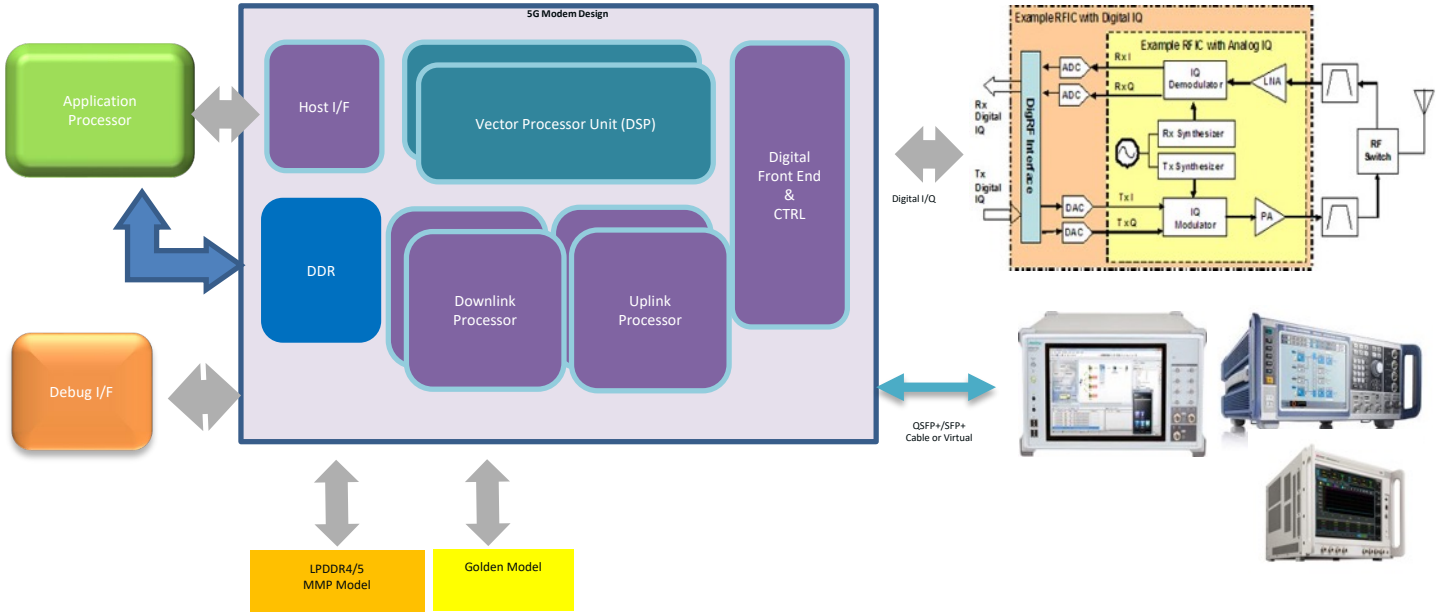
Source: Amlogic

© Cadence Design Systems, Inc - All Rights Reserved

But … your Mom's and Dad's prototyping doesn't cut it any more

- Too many gates
- Too much memory
- Too many peripherals
- Too much software

  …
- And not enough time



© Cadence Design Systems, Inc - All Rights Reserved

Protium X1 Enterprise Prototyping System

NEW

- **Performance**
  - Enabling early firmware and software development, automated bring-up
  - Up to 100MHz for single FPGA; up to 5MHz on billion gate designs
- **Capacity**
  - Advanced blade architecture scales to billions of gates
  - Ideal for AI, ML, 5G, mobile, and graphics applications
- **Fast Bring-up**
  - Unified Palladium® Z1 / Protium™ X1 compile ensures DUT congruency
  - Enables transition from emulation to prototyping in days
- **Multi-user**
  - Single-FPGA granularity assures high utilization and efficiency
  - Ideal for storage, automotive, image, consumer and medical applications

© Cadence Design Systems, Inc - All Rights Reserved

Scalable Capacity

- Blade architecture: scalability and flexibility
  - Each blade (of up to 150M gates) self-contained
  - Can be used as individual desktop system
  - Up to 8 blades mounted into standard 19” rack (1.2BG per rack)
- Racks connected for multi-billion gate prototyping
  - AI, 5G, graphic and mobile designs
- GUI-based, interactive configuration assistant
  - Create optimal configuration for user needs

© Cadence Design Systems, Inc - All Rights Reserved

Scalable Performance

- New fully-automatic partitioning, technology mapping algorithms for best possible performance regardless of design size
  - New Pathfinder multi-FPGA partitioner
  - New multi-strategy, high-speed pin-multiplexing

- Manual optimization capabilities for higher performance up to 100MHz+
  - Black-boxing: native, high-speed interfaces
  - Manual partition guidance: sub-systems
  - Inter-FPGA hardware optimization to customize connectivity

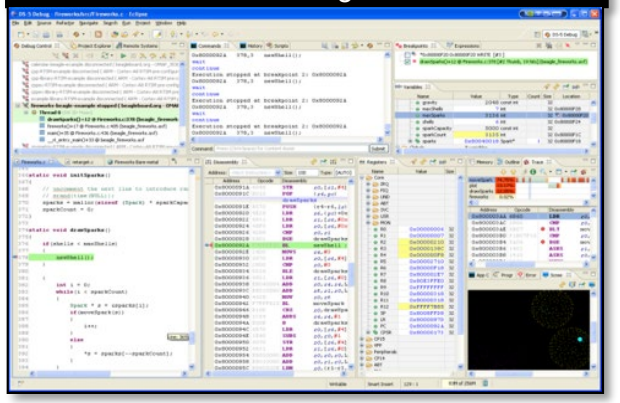
© Cadence Design Systems, Inc - All Rights Reserved

Advanced Debug

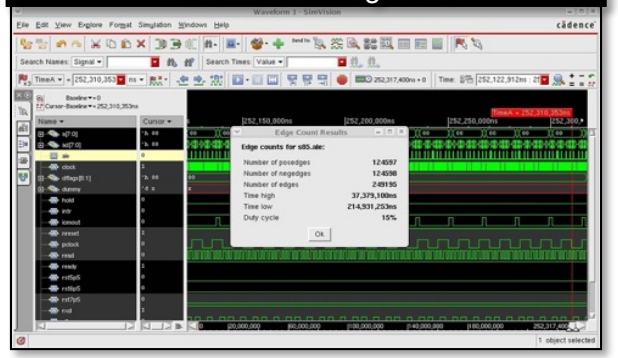
- **Software debug**: early firmware and software development
  - Memory (backdoor) upload and download
  - Clock control to stop and resume the hardware at any time
- Standard interfaces to industry-leading debuggers and software environments - use familiar tools
  - Joint Test Action Group (JTAG) and Universal Asynchronous Receiver/Transmitter (UART) interfaces
- Transaction interface
  - Directly connect to software models and virtual environments

- **Hardware debug:** bring-up design quickly, validate functionality
  - Force and release for internal notes
  - Prototyping full visibility
- Data capture card (DCC): 1000's of signals, millions of cycles
- Assertion checkers: automation of remote test execution

Software Debug: C Code



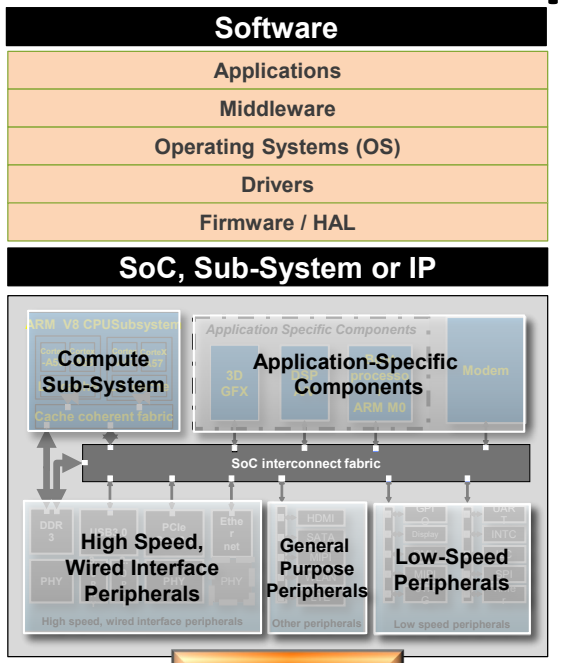
Hardware Debug: RTL



© Cadence Design Systems, Inc - All Rights Reserved

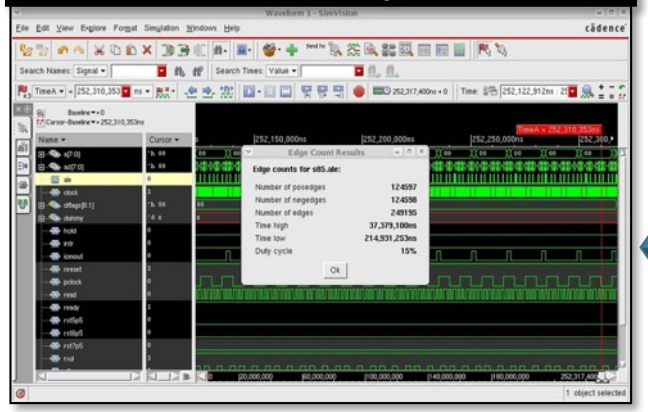
Advanced Prototping Debug

- Waveforms across partitions - Design-centric view
- Force/Release - Predefined signals to “0” or “1” during runtime
- Real-time monitoring of predefined (at compile time) signals
- Data Capture - Thousands of signals for millions of cycles
- Prototyping Full Visibility Probe without recompile
- Assertion Checkers - Non-intrusive hardware monitors



- Backdoor memory access - change boot code, software, etc.
- Clock control - Start/Stop the clock on demand
- Fully scriptable runtime environment
- Remote access - Network resource anytime from anywhere
- High-performance link to software model

Hardware Debug: RTL

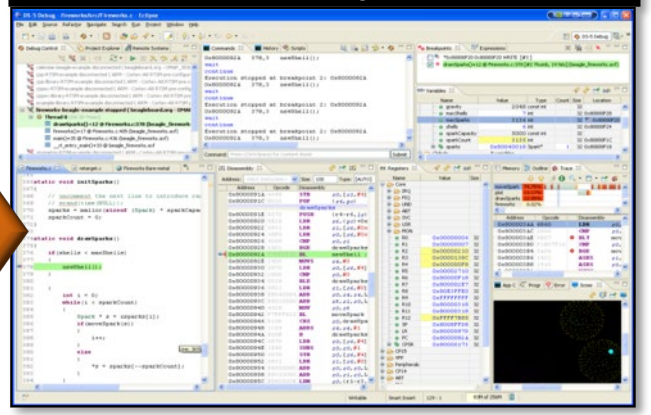


Probes

Daughter Cards & Peripherals

JTAG

Software Debug: C Code



ved

DeepChip – “Best of 2019”

The users have spoken!

- The Protium users also gushed a lot about the fact that they could *go back to Palladium* for fast debug and waves if needed.
  - "We run our design on Protium then go back to Palladium for debug. With Palladium, we can capture waves up and down the hierarchy of every net in our chip. It's a really big advantage of Protium.”
  - "With Protium, we get the speed of an FPGA-based system, with the fast ramp-up, debug, and signal traces of Palladiun. It only takes seconds for us to see all the waveforms.”
- Protium took 1.2 to 2.1 days to recompile vs. Synopsys HAPS taking 3.9 to 6.1 days to recompile.

And that's why Protium (actually the crazy fast incremental Protium compiles with FPGA 8.3 Mhz simulation speeds) wins the #1 Best of EDA in 2019 award from the end users this year.

What the EDA users REALLY think
DEEPCHIP

cadence Spectre X Simulator

(DAC'19 Item 1a) ---------- [12/19/19]

Subject: CDNS Protium crazy fast "Palladium-compiles" #1a for Best of 2019

FAST COMPILES ROCK!: My quick-and-dirty summary of the emulator/prototyper world. Say you have two designs to simulate. One design is 200 million gates, the other design is 1 Billion gates.

	Initial Ramp Up Time / Incremental Compile Time	Operating Speed
Palladium	initial ramp 2-4 weeks	
200 M gates	1.0 hour	1.2 Mhz
1.0 B gates	5.0 hours	800 Khz
Zebu Server 4	initial ramp 4-6 weeks	
200 M gates	25.8 hours (1.1 days)	2.0 Mhz
1.0 B gates	41.2 hours (1.7 days)	750 Khz
HAPS-80	initial ramp 2-3 months	
200 M gates	93.6 hours (3.9 days)	20.0 Mhz
1.0 B gates	146.4 hours (6.1 days)	5.0 Mhz
Veloce Strato	initial ramp 3-5 weeks	
200 M gates	5.1 hours	1.6 Mhz
1.0 B gates	12.5 hours	750 Khz
	ramp 24 hours w/Palladium	
Protium	ramp 4-6 weeks w/o Palladium	
200 M gates	28.8 hours (1.2 days)	8.3 Mhz
1.0 B gates	50.4 hours (2.1 days)	4.5 Mhz

© Cadence Design Systems, Inc - All Rights Reserved

Test Solutions for Emulation / Prototyping

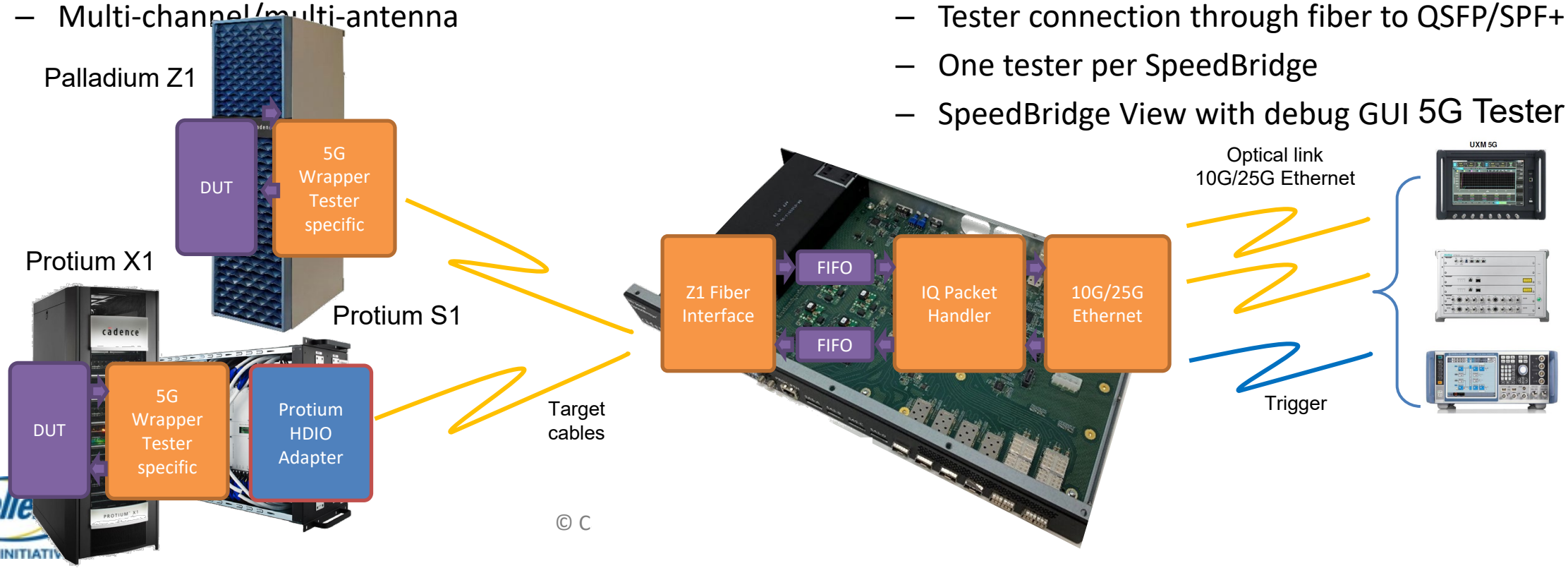
- Next Gen 5G testers: serial high speed link
  - Due to higher 5G bandwidth requirements
  - Better physical constraints, standards based
  - 25G/40G Ethernet through QSFP/SFP+ fiber

- Off the shelf solution needed: rate adapt, convert to IQ data
  - Testers no longer slowdown parallel IQ data
  - Need to packetize/de-packetize Ethernet traffic
  - Need triggering method to rate adapt traffic flow

© Cadence Design Systems, Inc - All Rights Reserved

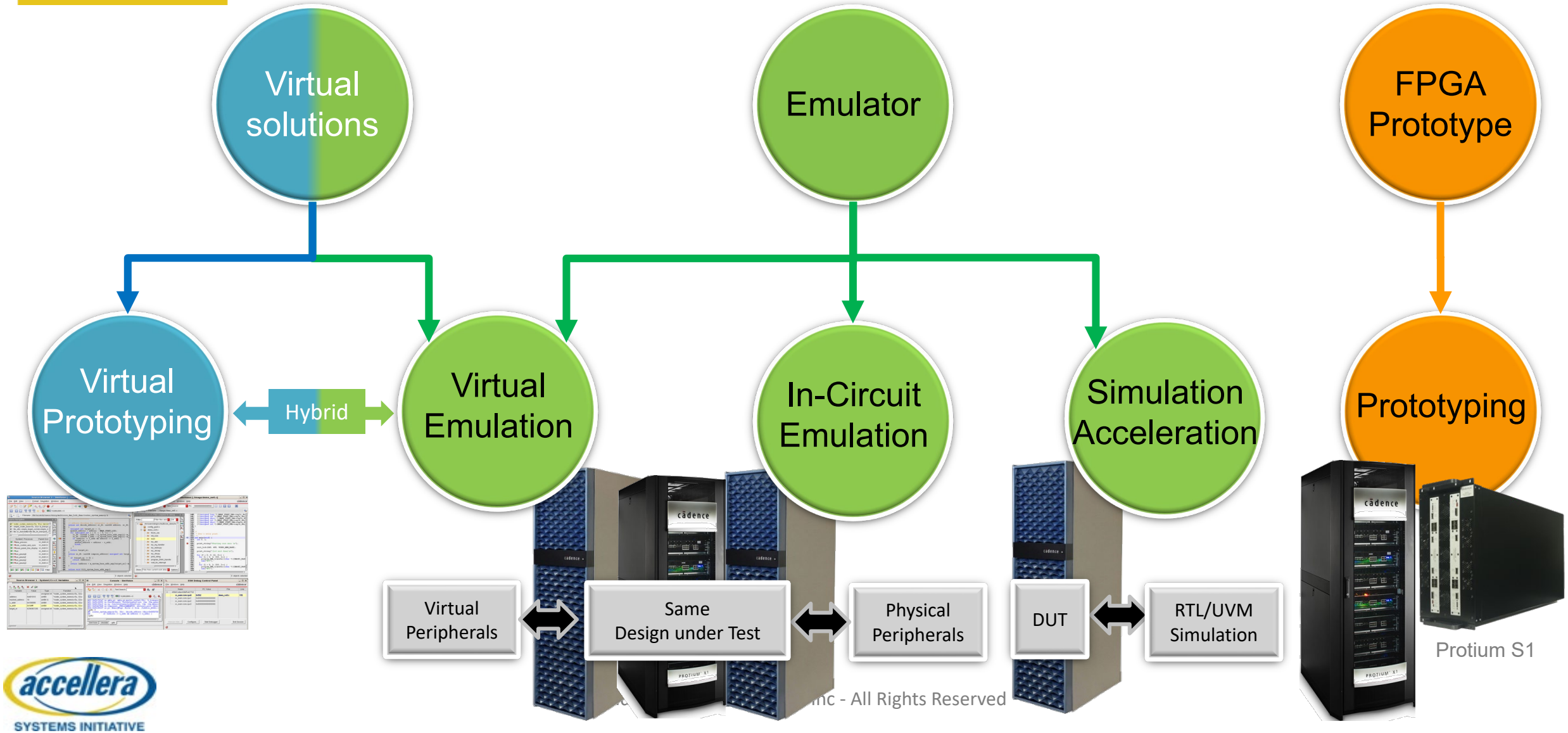
HD 5G Rate Adapter

- Capabilities
  - R&S and Keysight
  - Palladium Z1, Protium S1/X1
  - Slow 5G IQ data to 5G Handset DUT
  - Multi-channel/multi-antenna
- Features
  - Direct fiber connection (Z1)
  - HDIO adapter (Protium S1/X1)
  - 1U 19” Rack mountable
  - Tester connection through fiber to QSFP/SPF+
  - One tester per SpeedBridge
  - SpeedBridge View with debug GUI 5G Tester



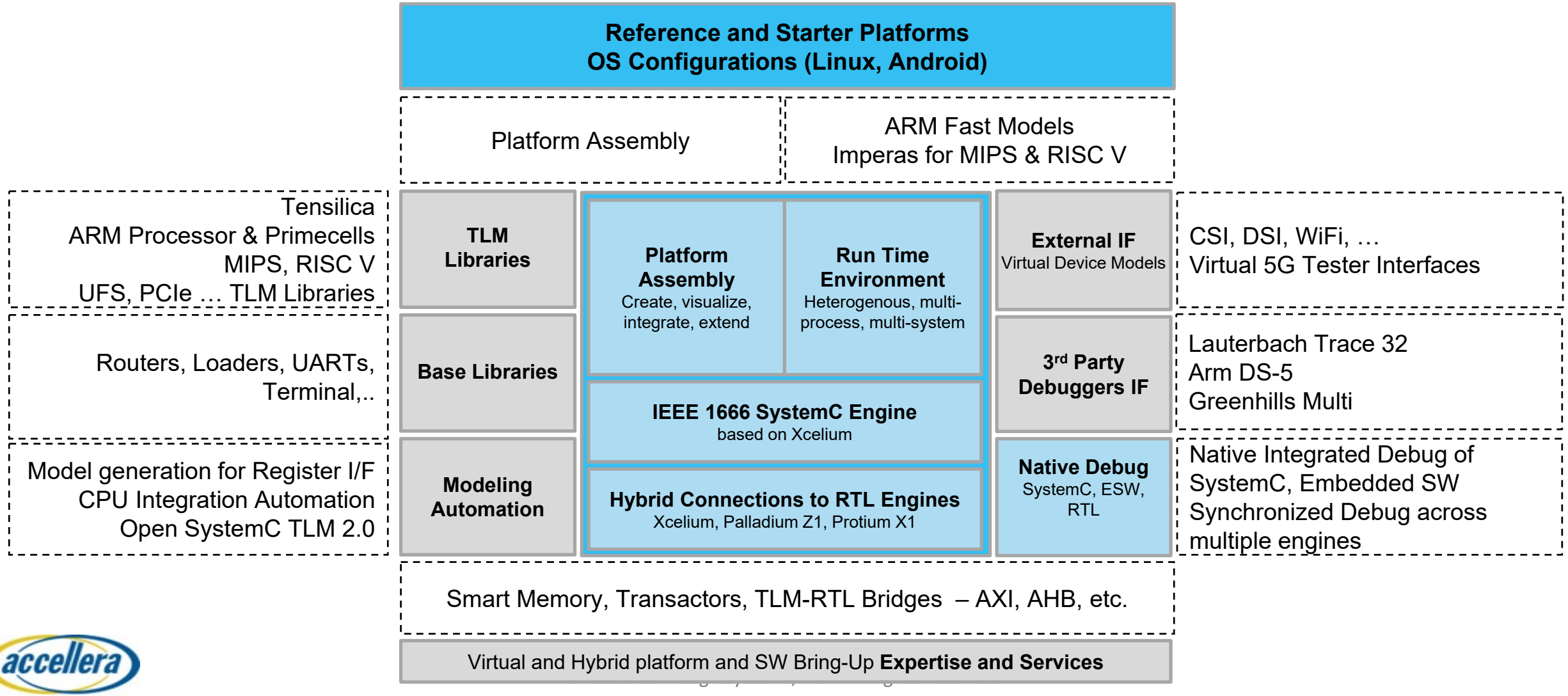
© C

Virtual, Emulation & Prototyping

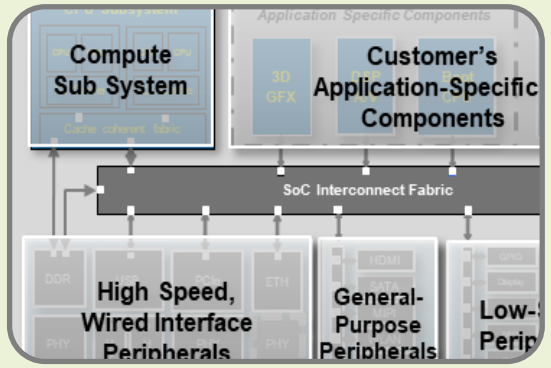


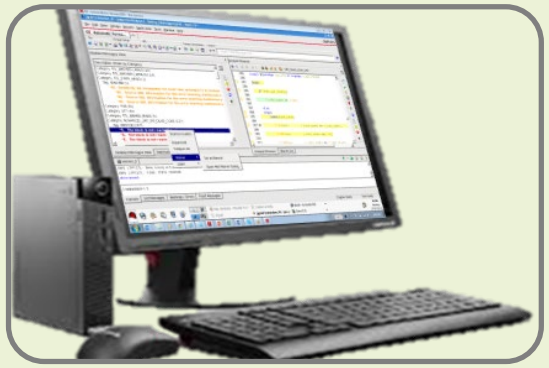
inc - All Rights Reserved

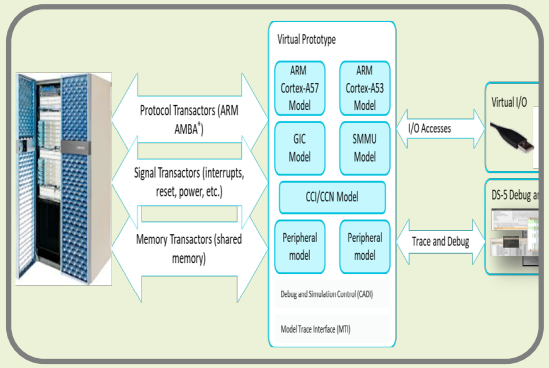
Virtual Prototyping and Hybrid Environment

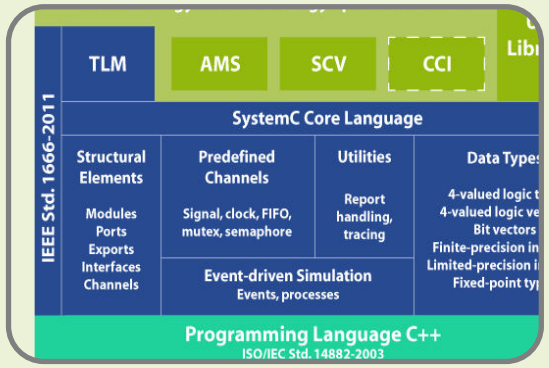


Use Models









Architecture Analysis

Accuracy, Accuracy

Best achieved "RTL-up", auto generation for interconnect

AT modeling requires careful effort-accuracy-performance considerations

Software Development

Speed, Speed, Speed

Loosely timed

"just enough" accuracy is sufficient

Mixed Fidelity Hybrids

Speed & Fidelity

Details in hardware

Keep processor sub-system and other peripherals virtual

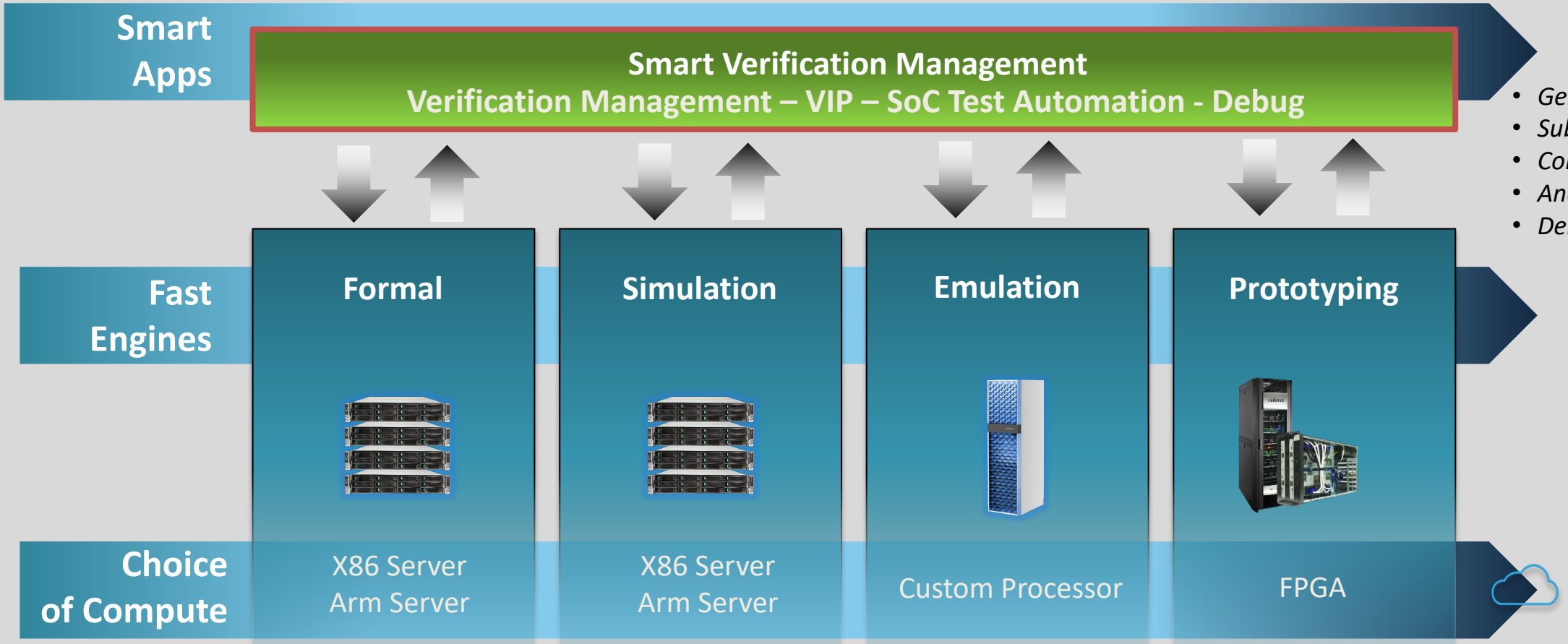
Smart synchronization between TLM and RTL

Hardware Verification

SystemC to drive hardware DUT

© Cadence Design Systems, Inc - All Rights Reserved

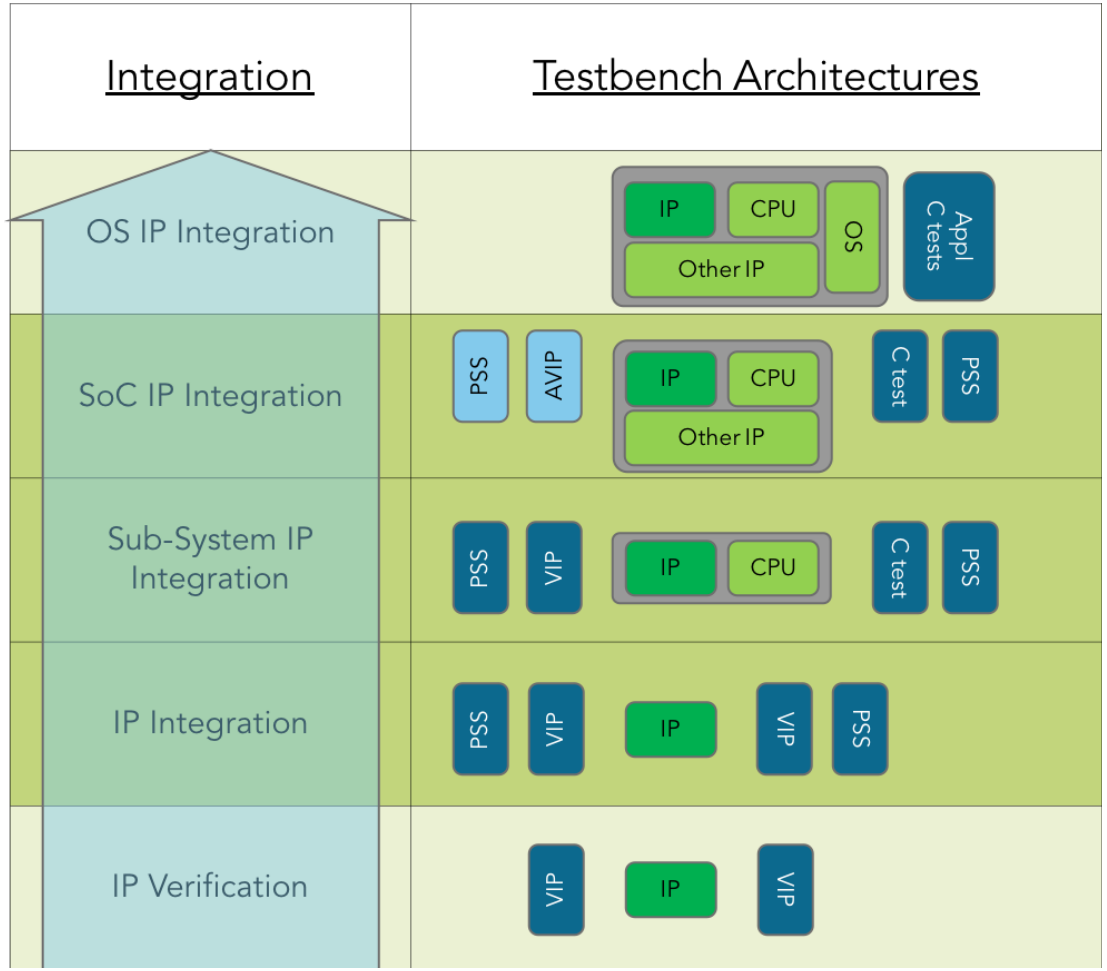
Smart Verification Management

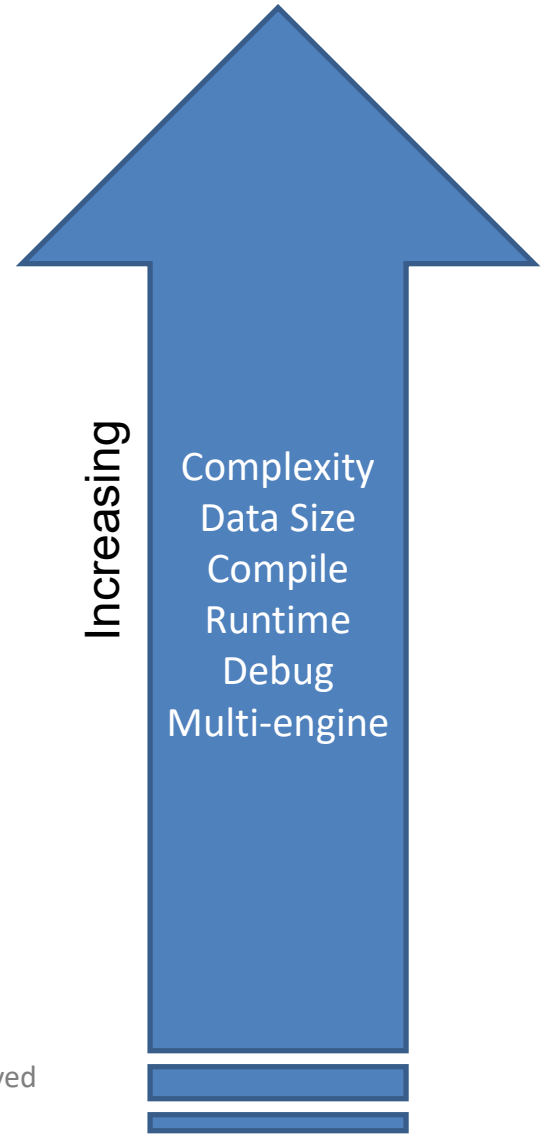


- *Generate tests*
- *Submit jobs*
- *Collect coverage*
- *Analyze coverage*
- *Debug failures*

© Cadence Design Systems, Inc - All Rights Reserved

The Problem





Formal **Simulation** **Emulation** **Prototyping**

© Cadence Design Systems, Inc - All Rights Reserved

Verification Management: Data-Driven

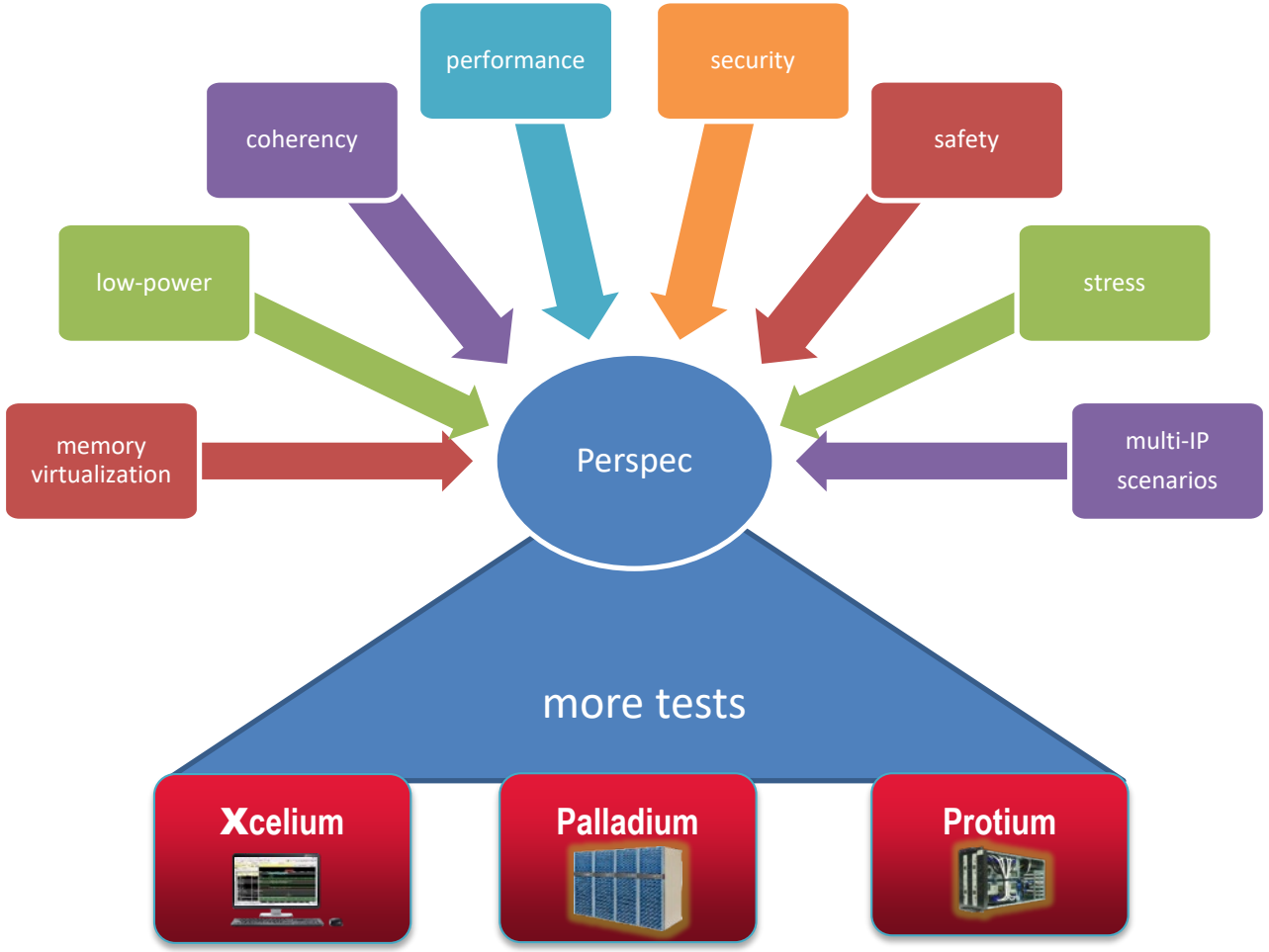
Use-case-based	• Define legal operations • Workload matters: must represent real operation
Data Collection	• Non-intrusive data collection • Use the right execution platform
Analysis	• Correlate, filter, learn, predict • Anomaly Detection
Goal-based	• Verification throughput • Smarter bug hunting

© Cadence Design Systems, Inc - All Rights Reserved

Use-case Based Test Generation

Accellera Portable Stimulus Standard

- Describe Test Intent and Design Behaviors
  - Use-cases
  - Legal scenario space
- Deliver Test Portability
  - Vertical reuse: From IP to SoC
  - Horizontal reuse: from Simulation to Emulation to Post Silicon
- Across Users
  - Abstract modeling
  - Actions, inputs, outputs, resources



© Cadence Design Systems, Inc - All Rights Reserved

PSS Impact on Stimulus

EXISTING STIMULUS	STIMULUS WITH PSS
Silicon • Generally OS-based tests • Failures difficult to bring to emulation or simulation for debug • Longer debug time	**Silicon** • Efficient deterministic bare-metal tests • Compose complex scenarios • Easily bring debug to Emulation • Generate large set of scenarios for regression
Pre-silicon SoC • Simple directed feature tests • Difficult to manually create complex scenarios • Prohibitively long digital simulation restricts plans for end-to-end complex scenarios	**Pre-silicon SoC** • Describe test intent with PSS • Automation helps with complex scenario composition • Reuse tests post-silicon
Pre-silicon IP • Excellent UVM-based constrained random testbench • IP initialization sequences not easily portable to FW or post-silicon tests	**Pre-silicon IP** • Keep existing UVM infrastructure • Export initialization sequences to firmware and post-silicon • Export IP specific scenarios to SoC

SHIFTLEFT

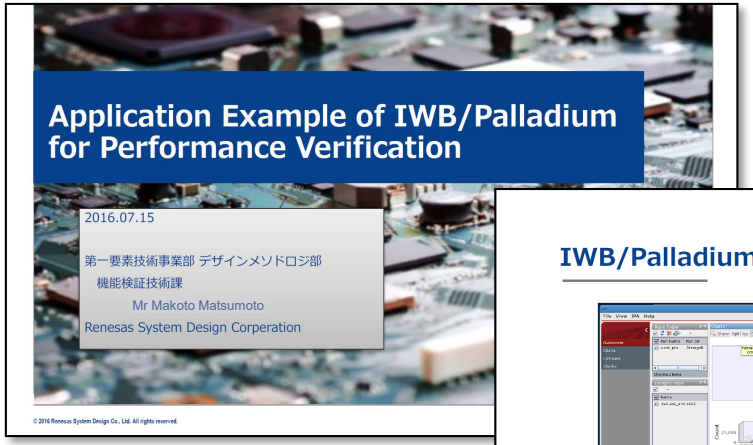
4 | AMD Shift Left with PSS | 2019 | AMD confidential - shared under NDA

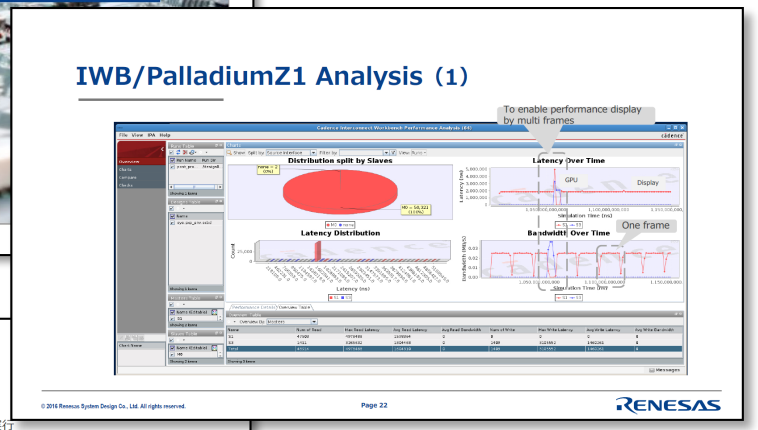
AMD

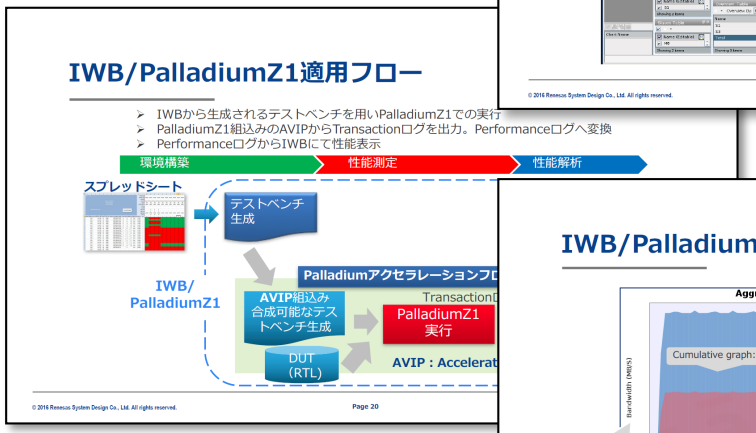
© Cadence Design Systems, Inc. All Rights Reserved.

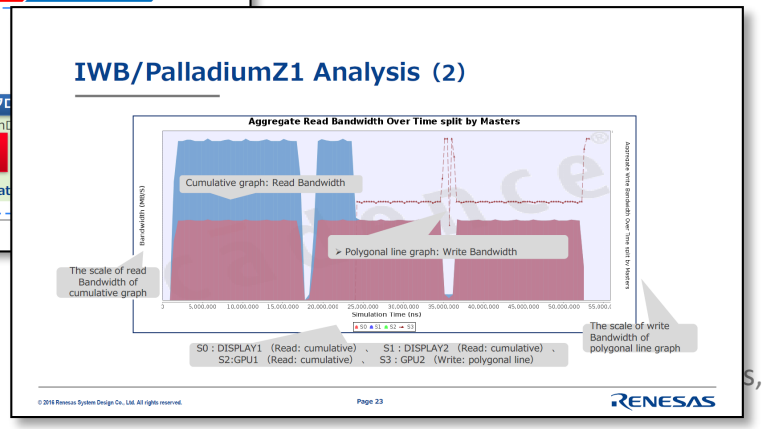
Excerpt from AMD DAC 2019 presentation

Renesas Performance Verification with Perspec Generated Use-cases









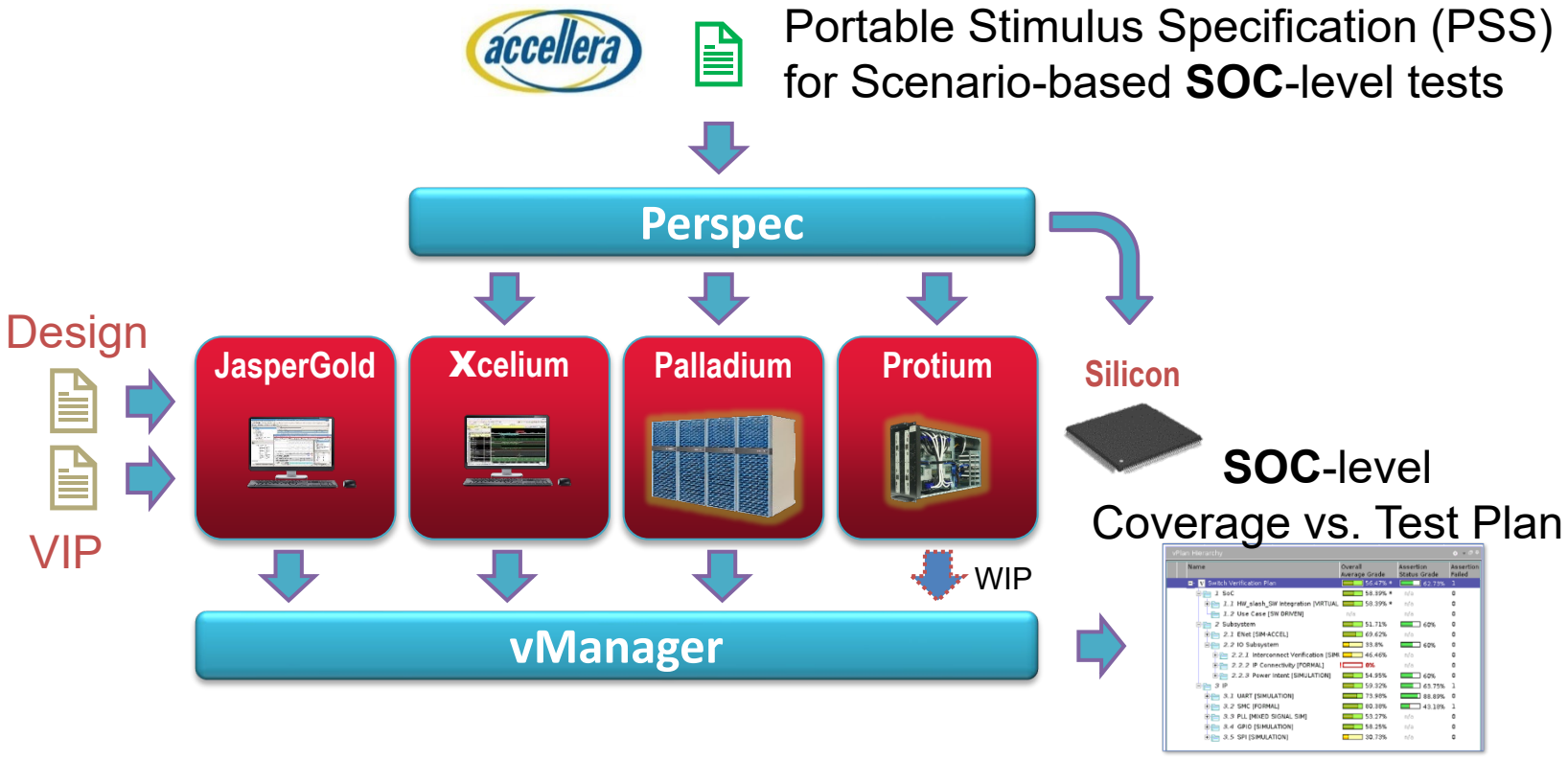
- Leading industrial and automotive MCUs
  - Number of integrated IPs is increasing
  - Switched interconnect
  - Configuration has big impact on performance
- Interconnect Workbench performance analysis
  - Early performance characterization
  - Interconnect tuning to optimize performance
  - Use case performance validation
- Palladium Z1 with Perspec use cases
  - Bring-up the entire design and software
  - Perspec generating use case tests
  - Reduce from 50 hour simulations to 12 minutes

s, Inc - All Rights Reserved

Perspec: Automated SoC Test Generation

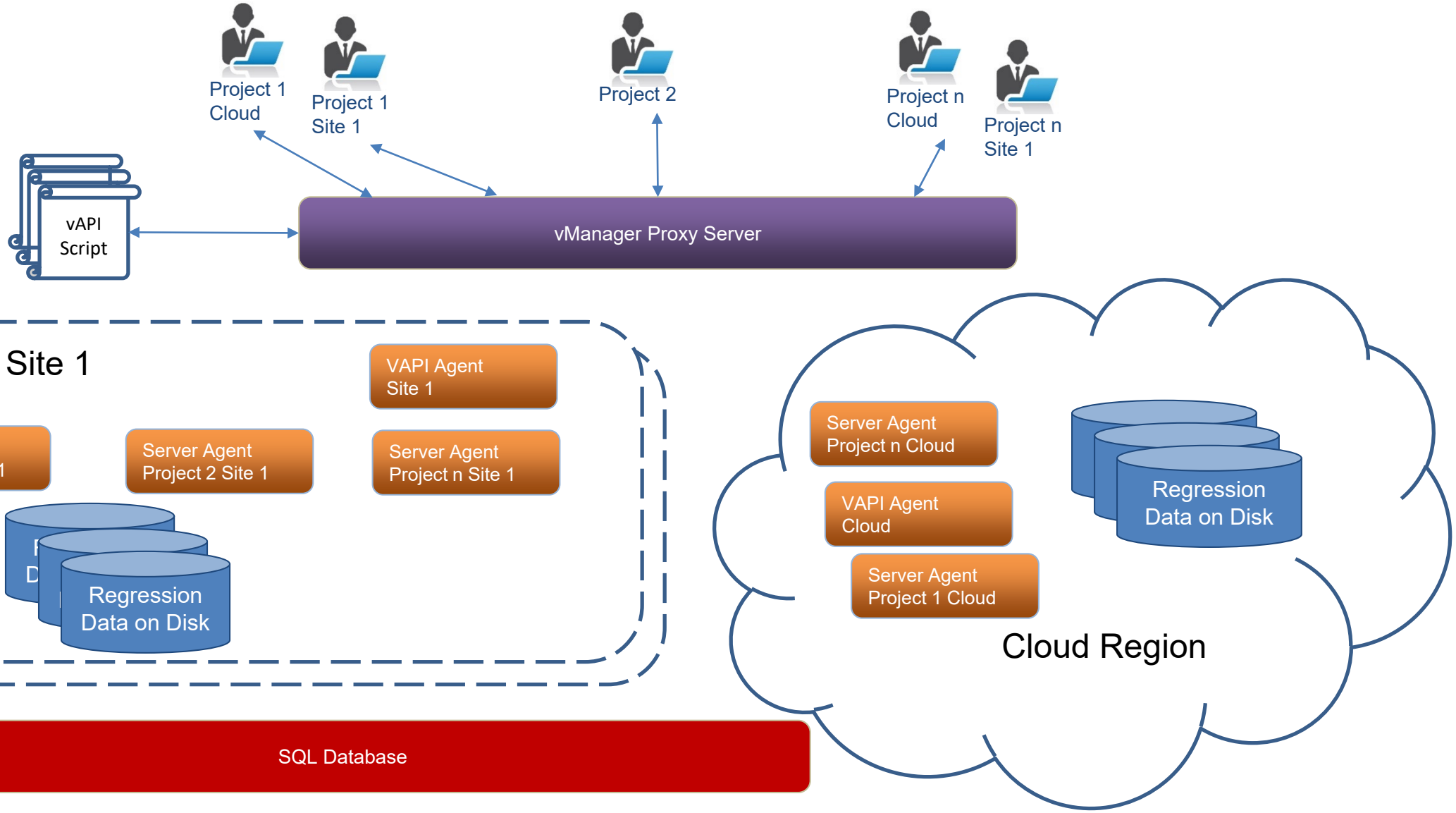
10x productivity gain creating regression suites

- Generates Scenario-based randomized tests
- Auto-generates C-based tests
- Portable across all engines
- Supports Accellera PSS 1.0



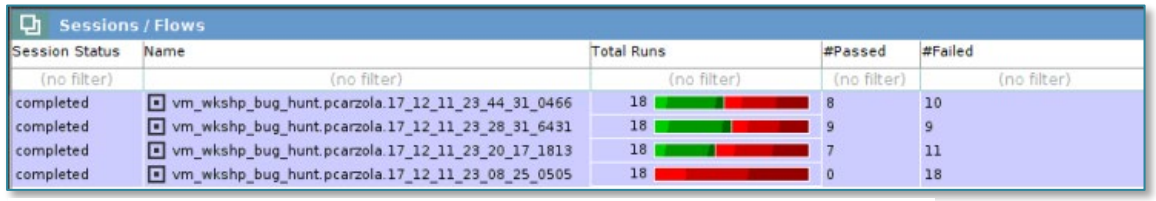
© Cadence Design Systems, Inc - All Rights Reserved

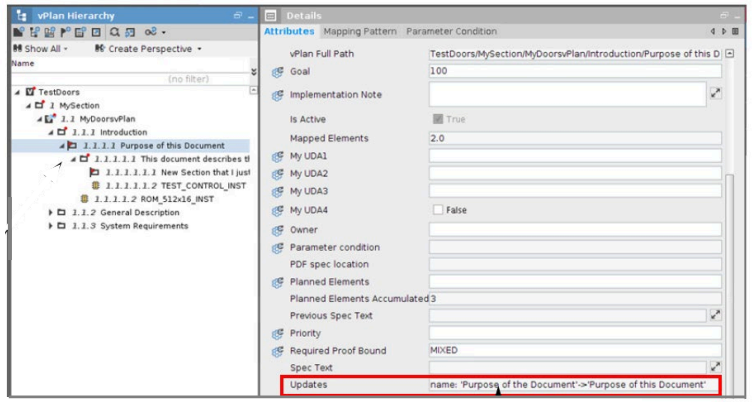
Distributed Data: Centralized Management

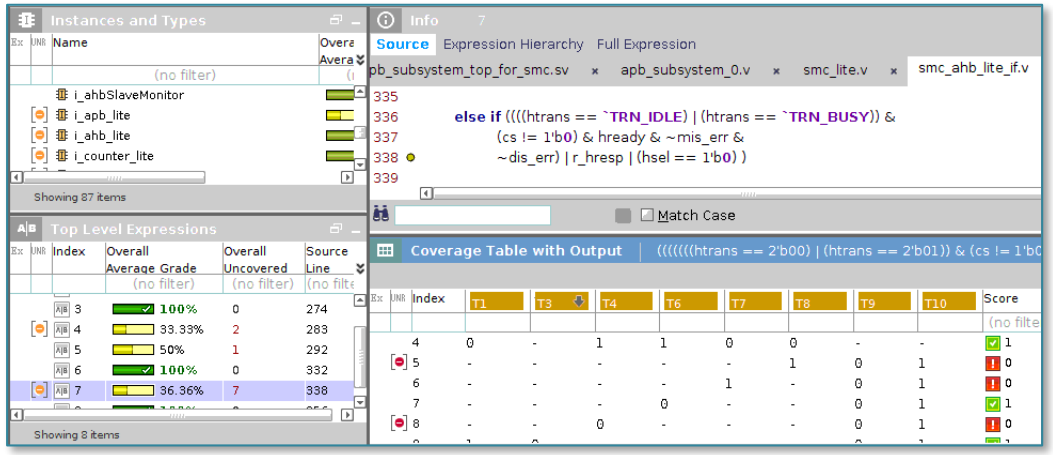


vManager Analytics

- Regression Productivity
  - Manage massive parallel execution
  - Analyze: rank by failures
  - Automate rerun of failures to collect debug data
- Requirements Traceability
  - Requirement import to seed verification plan creation
  - Link requirements and verification plan
  - Visibility of changes to requirements
- Coverage Analysis
  - Multi-engine coverage merge/combine
  - Refinement (Unreachability (UNR), UNR crosses)
  - Analysis of coverage vs. plan







© Cadence Design Systems, Inc - All Rights Reserved

vManager: Multi-engine Coverage

Combined Metrics | Xcellium | JasperGold

Verification Hierarchy

default

Name	Combined Average Grade	Combined Status Grade	Overall Average Grade	Assertion Status Grade	Formal Average Grade	Formal Status Grade	Valid Metrics
(no filter)	(no filter)	(no filter)	(no filter)	(no filter)	(no filter)	(no filter)	[b,s,a]
Verification Metrics	86.84%	44.34%	59.33%	45.31%	64.66%	19.38%	- - - - - - -
Types	67.03%	45.31%	64.63%	45.31%	0%	0%	- - - - - - -
Instances	95.77%	43.36%	54.02%	45.31%	90.96%	38.76%	- - - - - - -
top_tgen_pvmem	95.77%	43.36%	54.02%	45.31%	90.96%	38.76%	- - t - - - -
u_tgen_dut	97.86%	43.36%	63.76%	45.31%	95.6%	38.76%	b - t s - - -
u_apb_biu	100%	n/a	73.05%	n/a	100%	n/a	b - t s - - -
u_core_blk	99.25%	n/a	58.41%	n/a	99.25%	n/a	b e t s - - -
u_axi_biu	98.29%	n/a	66.34%	n/a	97.07%	n/a	b - t s - - -
u_chk_tgen	96.67%	36.92%	42.27%	44.62%	16.47%	22.22%	b - t - - a c
u_axi_checker_tgen	83.84%	45.55%	82.96%	45.55%	0%	41.44%	b - t - - a -
u_pvmem	90.03%	n/a	59.5%	n/a	79.52%	n/a	- - t - - - -

vPlan Hierarchy

Name	Overall Average Grade	User Entered Grade	User Calc Grade
APB_UART	74.51%	n/a	72.22%
1 APB_UART	67.89%	n/a	63.3%
1.1 Interfaces	69.79%	n/a	69.79%
1.2 Functional Features - B	n/a	n/a	49.5%
1.2.1 blah	n/a	n/a	60.5%
1.2.1.1 grade 1	n/a	77%	77%
1.2.1.2 grade 2	n/a	44%	44%
1.2.2 blah blah	n/a	n/a	38.5%
1.2.2.1 grade 3	n/a	53%	53%
1.2.2.2 grade 4	n/a	24%	24%
1.2.2.3 grade 5	n/a	n/a	n/a

- Single click merge across regressions
- Multi-engine coverage
- User-defined grade calculation

© Cadence Design Systems, Inc - All Rights Reserved

VIP Catalog – Protocol Support

2020 DESIGN AND VERIFICATION™ DVCON CONFERENCE AND EXHIBITION UNITED S

CCIX SEAMLESS ACCELERATION

nvm 1.3 EXPRESS

DisplayPort 1.4

SUPERSPEED USB 4.0 CERTIFIED — NEW 2019

HyperFlash

PCI EXPRESS 5 — NEW 2019

25G Ethernet 50G 200G 400G

UFS 3

mipi I3C

TSN Industrial Ethernet

ARM AMBA CHI-D — NEW 2019

CoaxPress®

DDR5

HBM2E

LPDDR5

GDDR6 — NEW 2019

Cloud/Data Center

USB 3.1 w OTG

PCI Express 4.0

ETHN 400G

DDR4, **DDR5**

SAS 24G

HBM 2.0 HMC

NVM Express

CCIX

CHI.B

Mobile/Consumer

ARM AMBA4/5 AXI/AHB/ACE

Display Port 8K

WiFi®

LPDDR4 **LPDDR5**

Bluetooth®

MIPI DSI 2

Wide I/O 2

MIPI CSI-2

USB 3.2

UFS 2.1

USB Type C

eMMC 5.0

MIPI I3C

MIPI SoundWire

SD Card 4.0

Automotive

ENET 10/100/ 1G/10G

CAN FD LIN

ENET AVB

Flash ONFI

DDR 2,3,4

ENET TSN

NOR FLASH

SPI

OctaRam

I2C

MIPI SoundWire

MIPI I3C

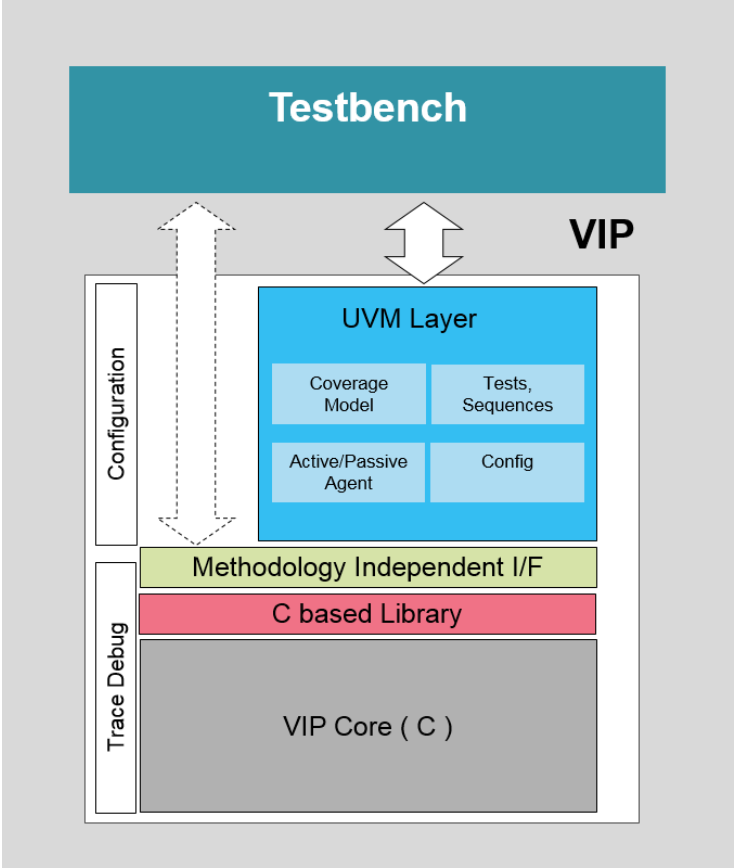
MIPI DSI

accellera SYSTEMS INITIATIVE

© Cadence Design Systems, Inc. All Rights Reserved

Cadence VIP Architecture

- Fast VIP
  - All VIPs implemented in C for the most optimized performance!
  - Uses dynamic memory allocation for all internal memories/registers/data structures
- Portable and Scalable architecture
  - Seamless IP=> SoC and Project => Project transition
- Support all languages, simulators, methodologies
  - Native SV, Verilog, OVM, UVM, VMM, C, SystemC, etc.
- Consistent user experience across multiple protocols
  - Common UVM library for all VIPs and Memory models



© Cadence Design Systems, Inc - All Rights Reserved

AI/ML INSIDE/OUTSIDE FOR VERIFICATION

© Cadence Design Systems, Inc - All Rights Reserved

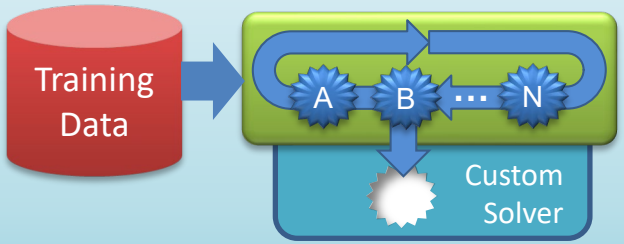
Glossary

- Machine learning
  - The study of algorithms and statistical models that computer systems use to perform a specific task effectively without using explicit instructions
- Machine learning model
  - An approximative model for a function, automatically generated from training data, that allows inference over new data
- Reinforcement learning
  - A machine learning task that allows computer systems to automatically and dynamically determine the ideal behavior within a specific context to maximize its performance, based on observation, reward and action
- Supervised learning
  - A machine learning task in which the algorithm builds a model of an unknown function from a subset of its inputs and the desired outputs
  - In this type of learning, one needs to supply labels to the output data

© Cadence Design Systems, Inc - All Rights Reserved

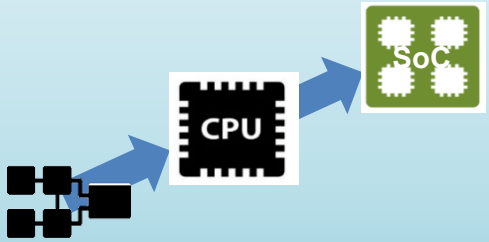
ML-enabled Formal Verification

Smart Proof Technology



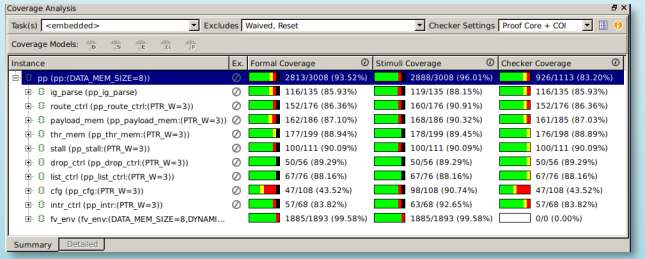
ML for solver inference and multi-advisor orchestration

Advanced Design Scalability



2x design capacity increase and **50%** memory footprint reduction

Signoff-quality Formal Coverage



Signoff-accurate formal coverage with new intuitive analysis GUI

Third-Generation JasperGold® Formal Verification Platform

"We measured 2x faster proofs out-of-the-box, 5x faster regressions and non-converged properties reduced by 50%"

-Mirella Negro Marcigaglia, digital design verification manager, STMicroelectronics

© Cadence Design Systems, Inc - All Rights Reserved

Smart Proof automation framework

Component & Data Management

Proof Profiling Data

- Regular read/write

Proof Caching

- Cache storage in single file
- Automatic cleanup of old cache data

Multi Advisor Proof Orchestration

- Forced `on`
- No overwrite on engine mode

Engine Algorithm Selection

- Automatic training and inference

Learning

Machine Learning

Optimizes subsequent runs/regressions

Optimizes out-of-the-box proofs

Find more bugs

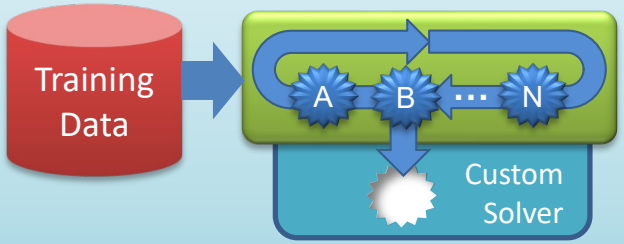
Better convergence

Faster proofs

© Cadence Design Systems, Inc - All Rights Reserved

Smart Proof technology: ML-enabled optimizations

Smart Proof Technology



ML for algorithm selection and multi-advisor orchestration

Algorithm Selection

Supervised learning: ML uses training data from 500+ customer designs
Supervised inference: Best-fit core engine selected and configured to create custom solver
=> Up to 4X (2X avg.) better out-of-the-box proofs

Multi-Advisor Orchestration

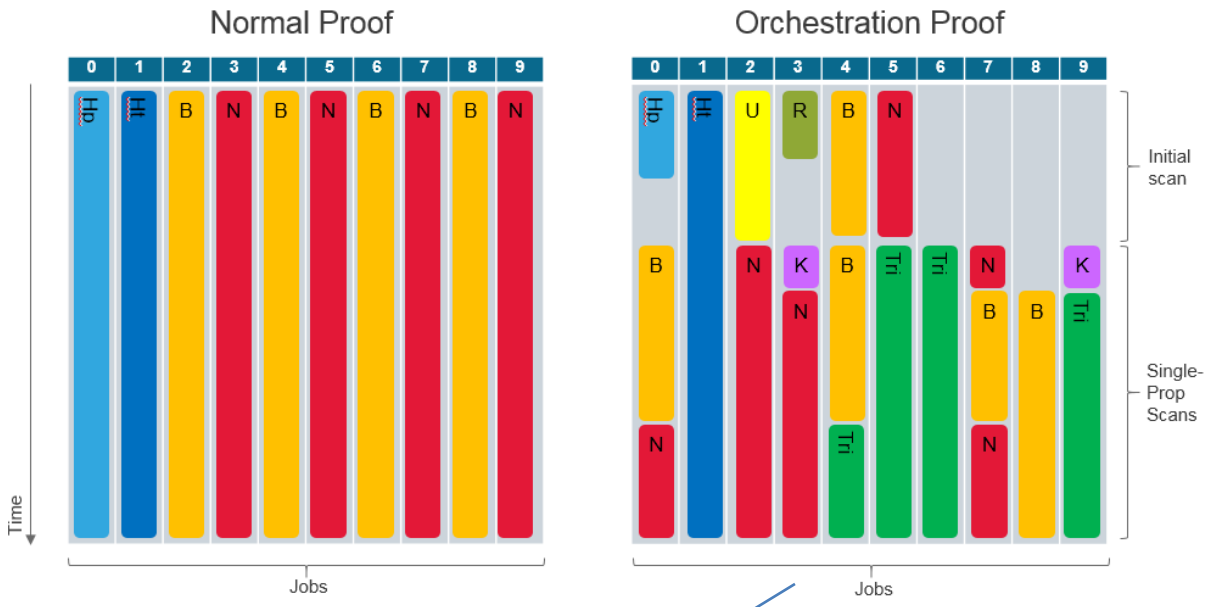
Multiple proof advisors use reinforcement learning to improve proof efficiency

Uses training data for **better out-of-the-box proofs**
Adjusts training data using proof profiling for **up to 6X (5X avg.) better subsequent proofs and regressions**

© Cadence Design Systems, Inc - All Rights Reserved

What is multi advisor proof orchestration?

- Algorithm that dynamically adjusts engine selection, time limits, license usage, etc, during a proof, enabled by reinforcement learning
  - Hides complexity from the user: hand technology control over to the tool
  - Focuses on resources (max licenses, max jobs, global time limit), which are still respected in orchestration
  - Uses time slicing to switch engines during run, allowing engine mode diversity
  - Opens the door to future automatic optimizations (partitioning, bug hunting, etc)
  - No need to deploy every new strategy
- Continuously expanded to include new technology

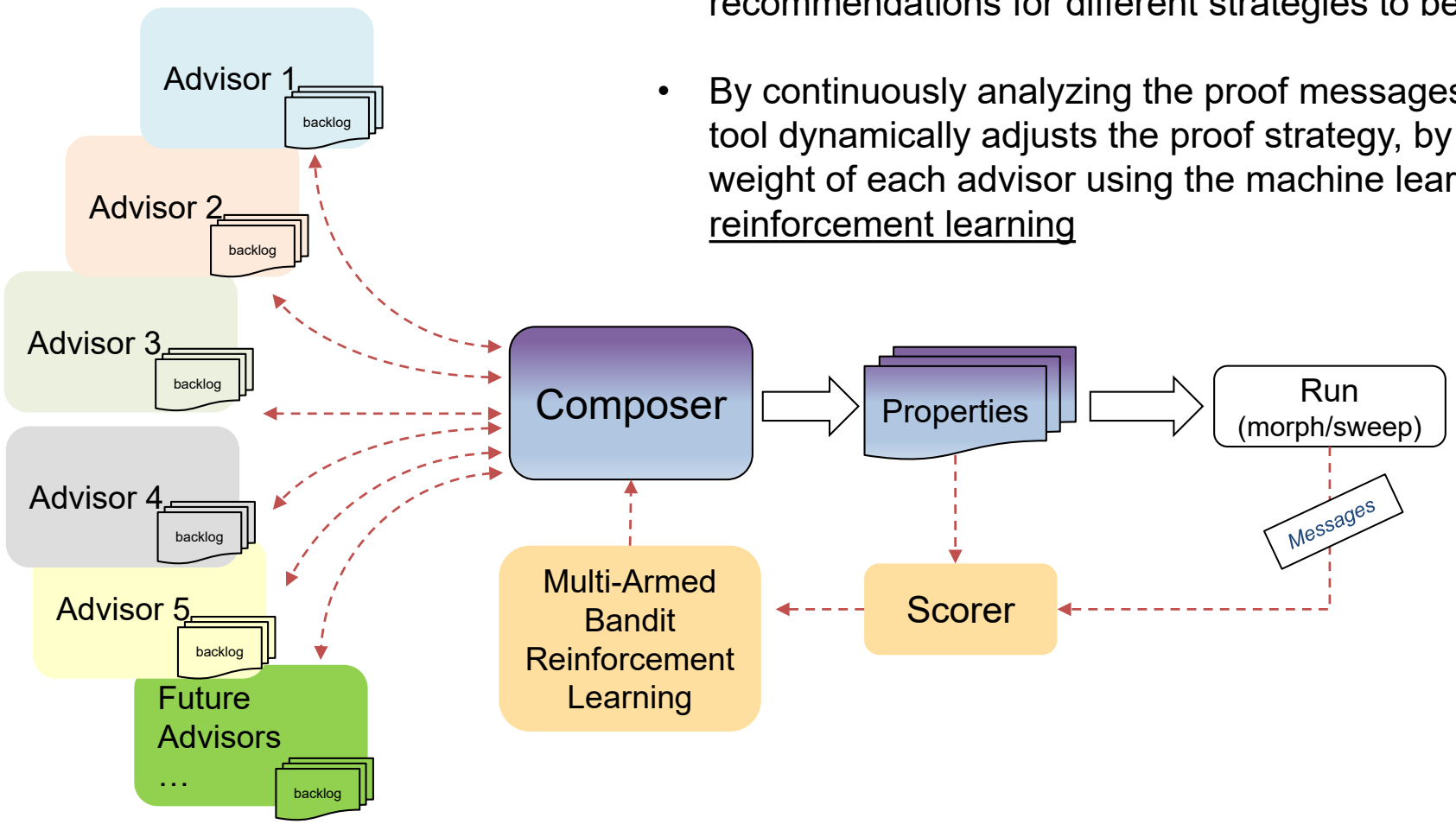


+ Custom engines dynamically created during proof, to further explore non-default engine settings

© Cadence Design Systems, Inc - All Rights Reserved

Multi advisor proof orchestration overview

- A group of advisors with different weight provide recommendations for different strategies to be run during a proof
- By continuously analyzing the proof messages and results, the tool dynamically adjusts the proof strategy, by updating the weight of each advisor using the machine learning strategy of <u>reinforcement learning</u>



© Cadence Design Systems, Inc - All Rights Reserved

How to enable multi advisor orchestration?

- Orchestration is on by default whenever the user doesn't change the engine mode for a proof, but can be forced to be turned on with configuration commands

```
[<embedded>] % prove -task <embedded>
INFO (IPF036): Starting proof on task: "<embedded>"
INFO (IPF031): Settings used for proof thread 1:
    orchestration                  = on (auto)
    time_limit                     = 86400s
    per_property_time_limit        = 1s * 10 ^ scan
    engine_mode                    = auto
    proofgrid_per_engine_max_jobs  = 1
    max engine jobs                = auto
    proofgrid_mode                 = local
    proofgrid_restarts             = 10
```

```
[<embedded>] % prove -task <embedded> -engine mode {B R D}
INFO (IPF036): Starting proof on task: "<embedded>", 53 properties
INFO (IPF031): Settings used for proof thread 3:
    orchestration                  = off (auto)
    time_limit                     = 86400s
    per_property_time_limit        = 1s * 10 ^ scan
    engine_mode                    = B R D
    proofgrid_per_engine_max_jobs  = 1
    max engine jobs                = B R D, total 3
    proofgrid_mode                 = local
    proofgrid_restarts             = 10
```

```
1558735989: INFO (IPF152): Orchestration is evaluating 3 properties using the following settings:
    per_property_time_limit        = 2s
    engine_mode                    = M AD AMcustom2*
```

...

```
1558736021: INFO (IPF152): Orchestration is evaluating 3 properties using the following settings:
    per_property_time_limit        = 8s
    engine_mode                    = D I Ncustom3*
```

...

```
1558739909: INFO (IPF152): Orchestration is evaluating 3 properties using the following settings:
    per_property_time_limit        = 512s
    engine_mode                    = Tri AMcustom2 Ncustom4*
```

© Cadence Design Systems, Inc - All Rights Reserved

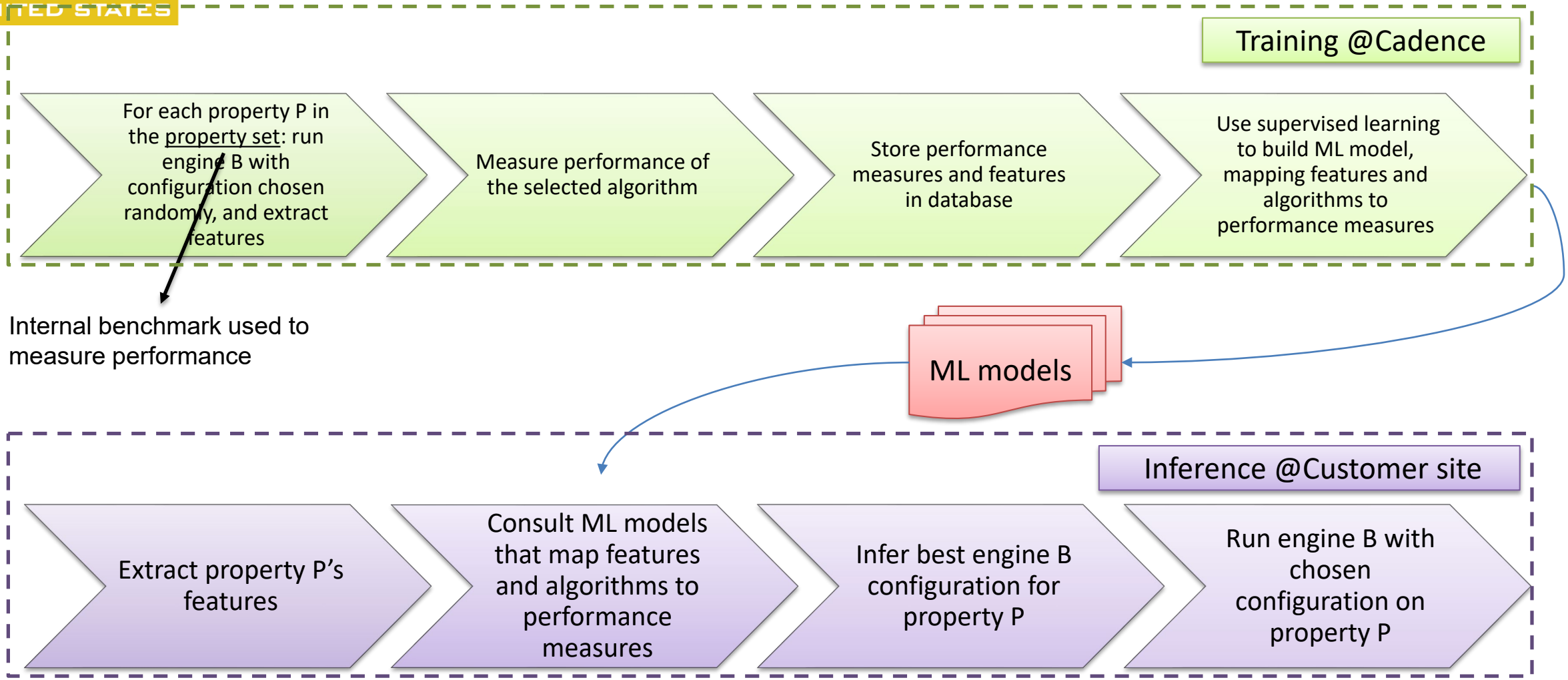
*Custom engines created dynamically in the proof to further explore non default settings

Algorithm selection concept: example engine B

- Challenge
  - JasperGold has different technologies which are used internally by engine B
  - The challenge is to choose the best engine B parameter configuration to run on each given property
- Approach
  - Use supervised learning to create a machine learning model trained on our internal benchmarks across multiple customer designs, to try to infer on the fly better-than-default engine B configuration on each given property
  - Wrap solution into a new engine called B4, which can be added to the engine mode using `set_engine_mode` command like any other engine

© Cadence Design Systems, Inc - All Rights Reserved

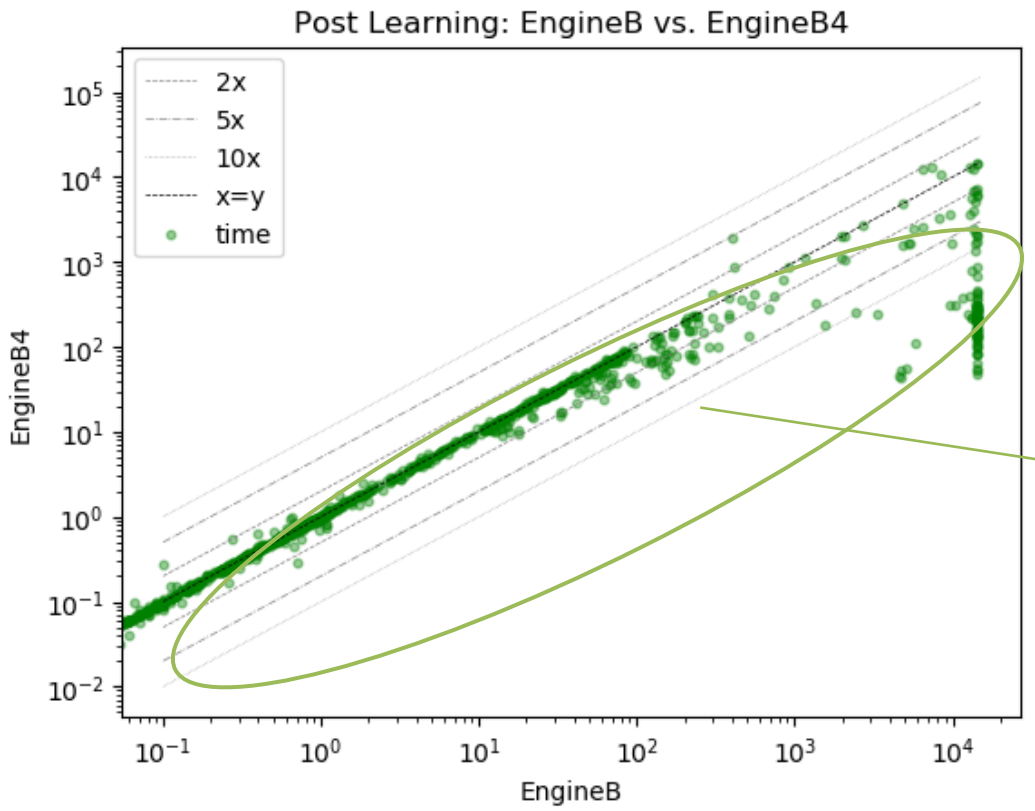
Algorithm selection for engine B overview



© Cadence Design Systems, Inc - All Rights Reserved

Algorithm selection for engine B results

- Inference evaluation against engine B showed overall performance boost when using engine B4



Dots below the diagonal line correspond to properties which were proven faster with engine B4, when compared to engine B

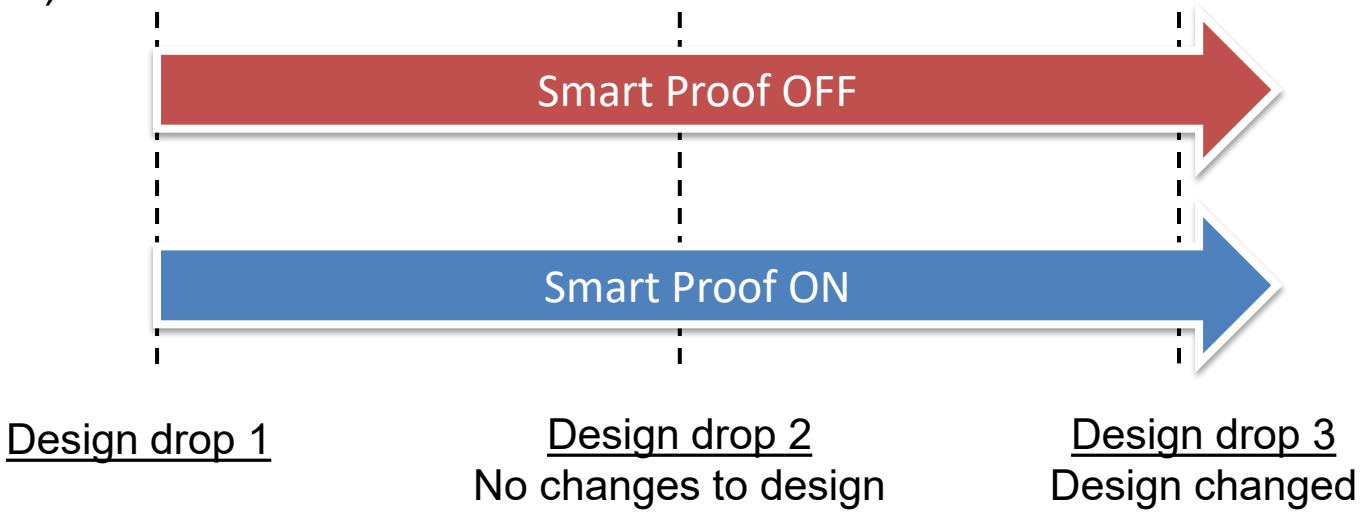
© Cadence Design Systems, Inc - All Rights Reserved

Smart Proof case study

A customer scenario

- Three versions of a customer design (processor core)
- Goal: compare performance of proof using Smart Proof *vs.* proof without it over time
- Resources
  - Dedicated machine
  - 20 jobs, 20 licenses, 24h

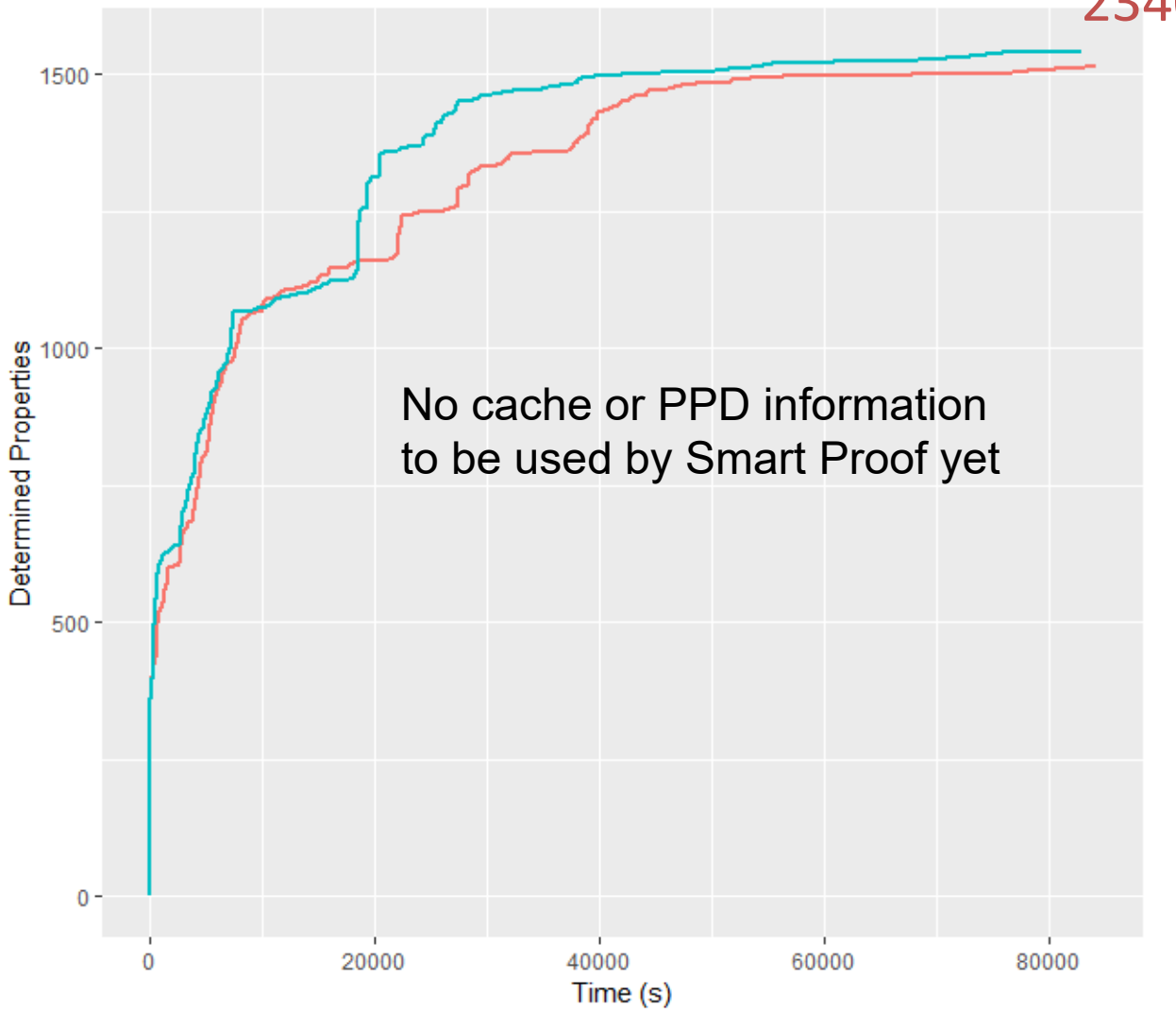
Characteristics for design drop 1	
# Properties	2340
# Gates	290k
Convergence after 24h with 20 jobs (Smart Proof off)	65%



© Cadence Design Systems, Inc - All Rights Reserved

Design drop 1

2340 properties

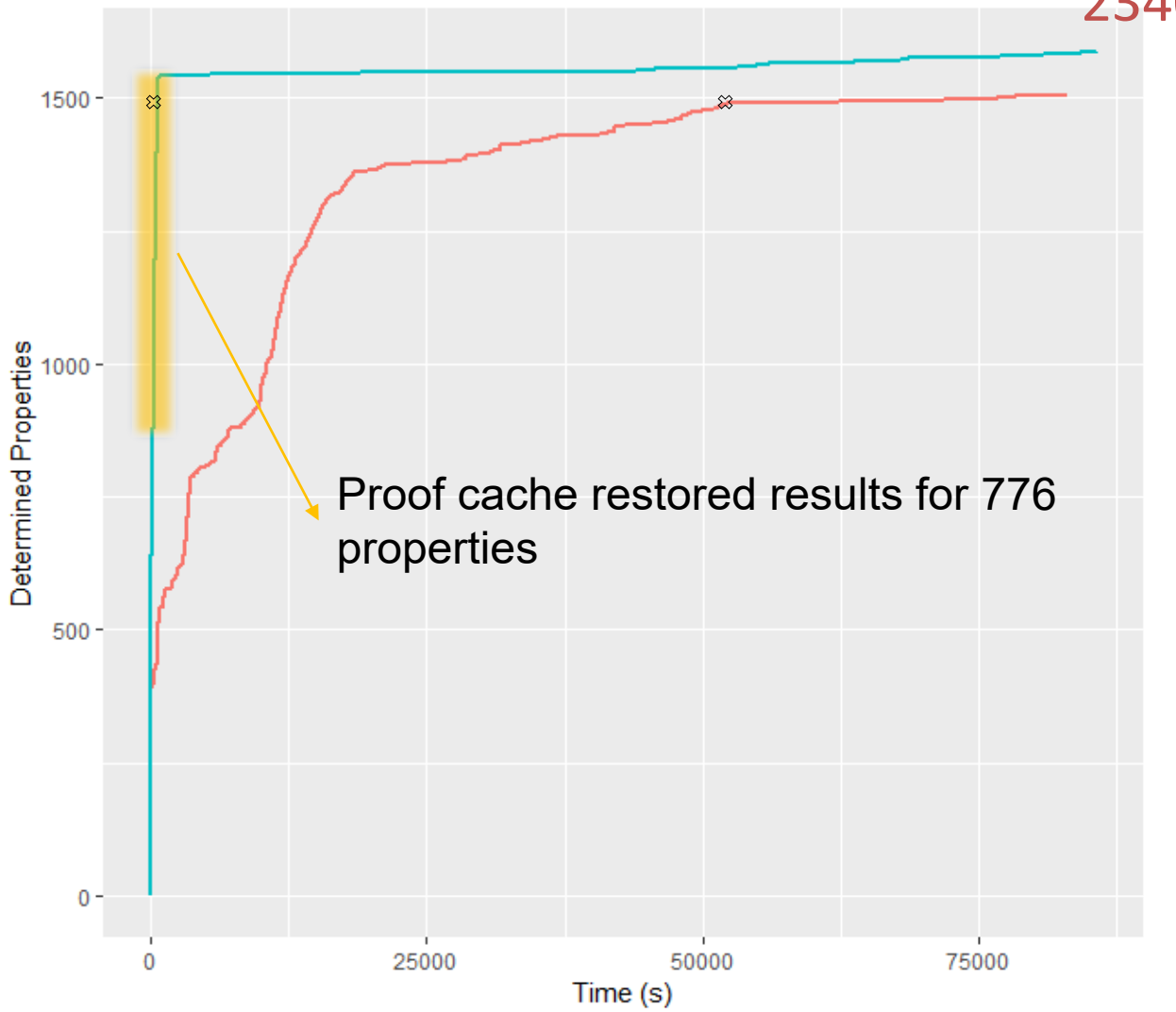


Run	# Determined
Smart Proof OFF	1516
Smart Proof ON	1543

© Cadence Design Systems, Inc - All Rights Reserved

Design drop 2: no changes to design

2340 properties



Run	# Determined
Smart Proof OFF	1506
Smart Proof ON	1590

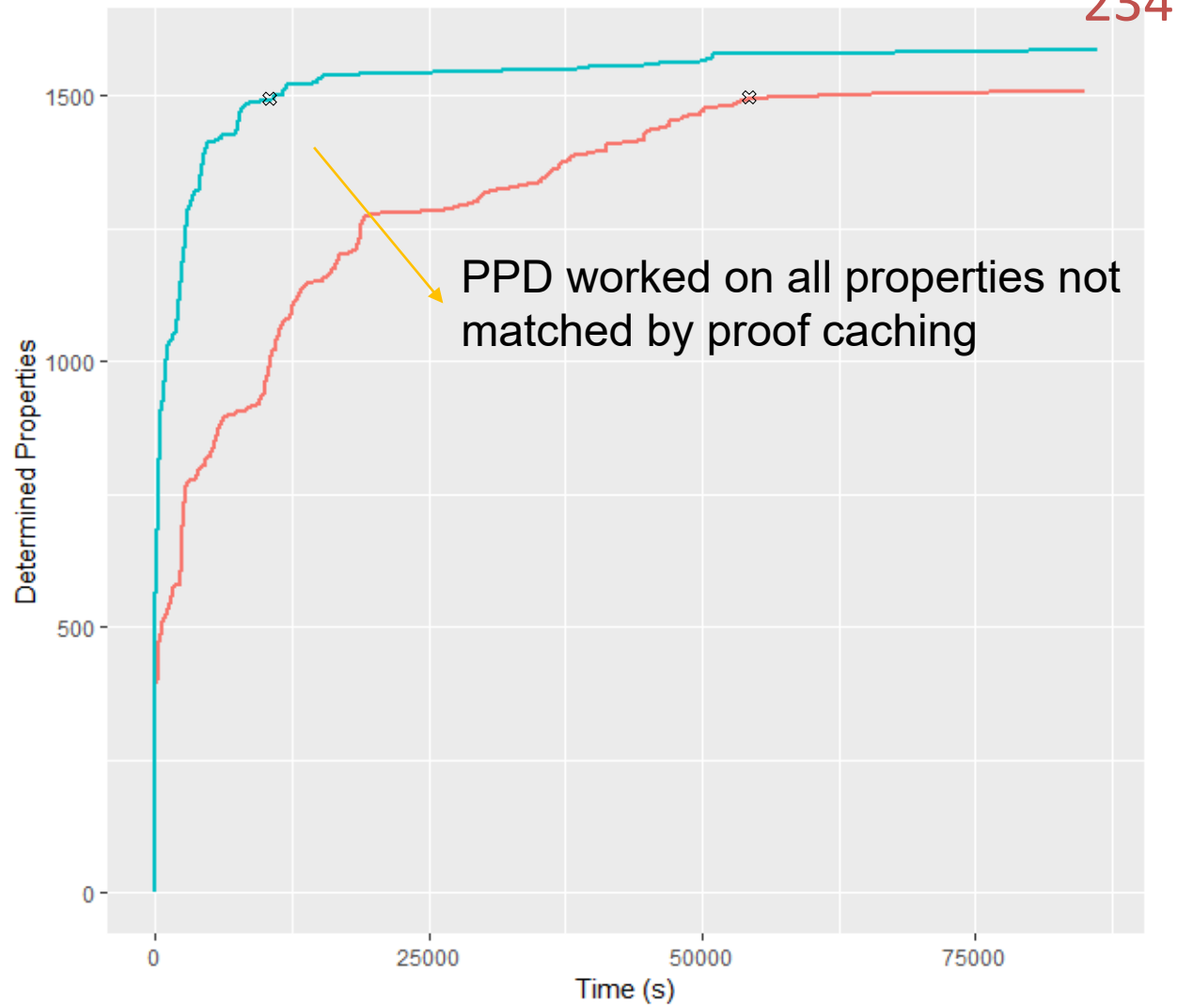
Speedup: run with Smart Proof ON is 32x faster than proof without it

All properties determined in drop 1 using Smart Proof were reproduced in drop 2 (the same is not true for the runs with Smart Proof off in this case study)

© Cadence Design Systems, Inc - All Rights Reserved

Design drop 3: design changed

2342 properties

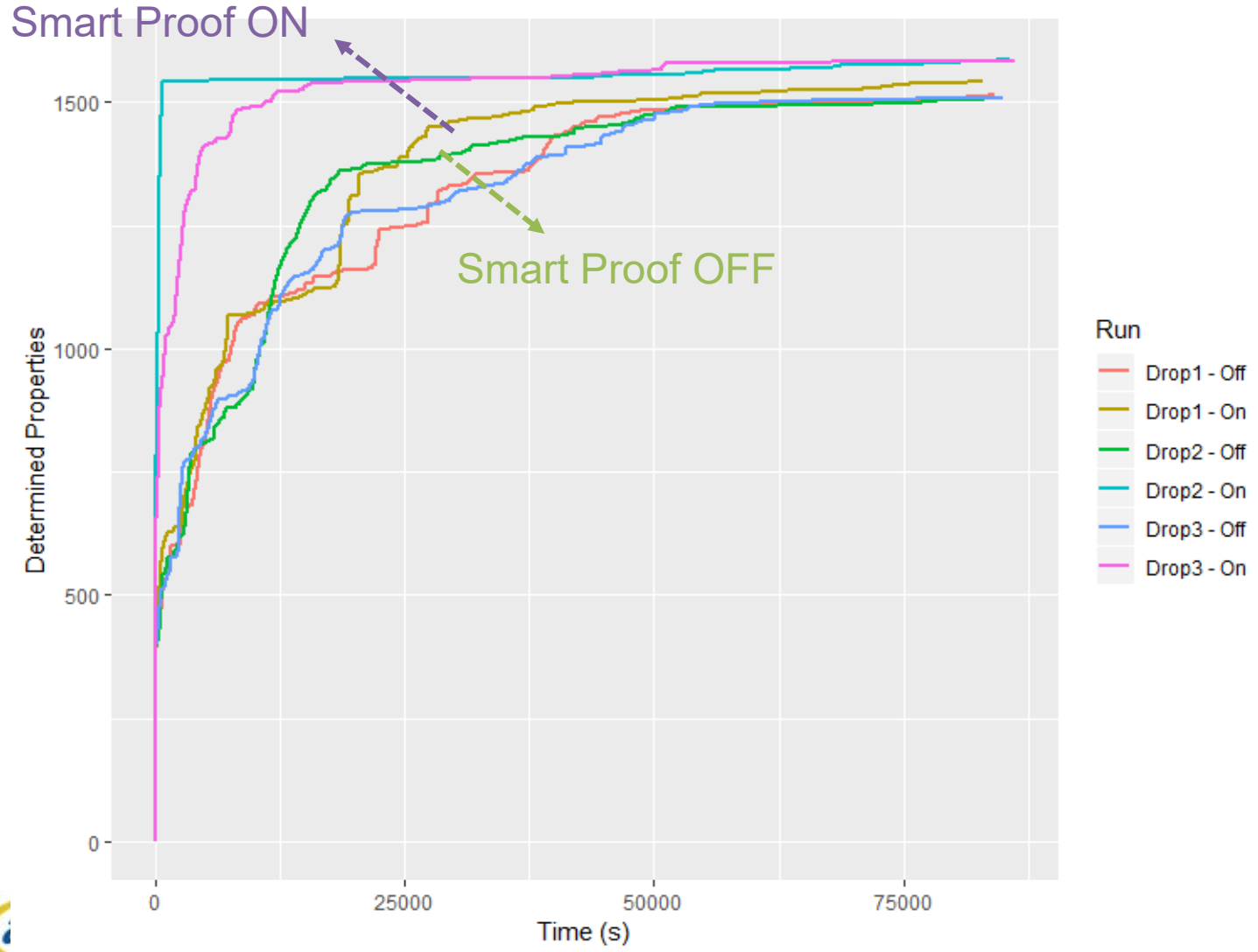


Run	# Determined
Smart Proof OFF	1510
Smart Proof ON	1586

Speedup: run with Smart Proof ON is 5x faster than proof without it

© Cadence Design Systems, Inc - All Rights Reserved

Case study take away



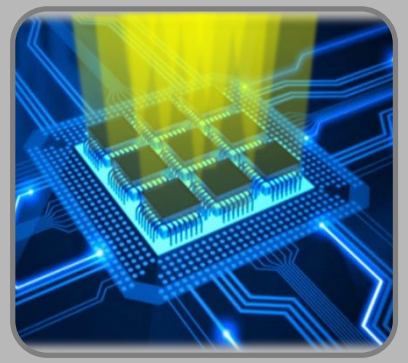
- By learning from previous runs, Smart Proof frees up resources, which are better utilized to solve hard properties
- When the design is stable, Smart Proof is a great tool to help reproduce previous proof results

Reserved

SUMMARY - OUTLOOK

© Cadence Design Systems, Inc - All Rights Reserved

Enabling Verification of AI/ML Designs



Formal

- High **Capacity**
- Regression improvements
- SAT Solver Inferencing

Simulation

- **Fast simulation** for high-activity designs
- UVM **randomization**
- Fast elaboration for **replicated structures**
- Coverage **metrics**

Emulation

- **Billion-gate designs**
- Parallel Partition Compiler
- INT8 to INT64
- **Power / Performance**
- Memory **models**: HBMx
- **Senor model**: MIPI CSI

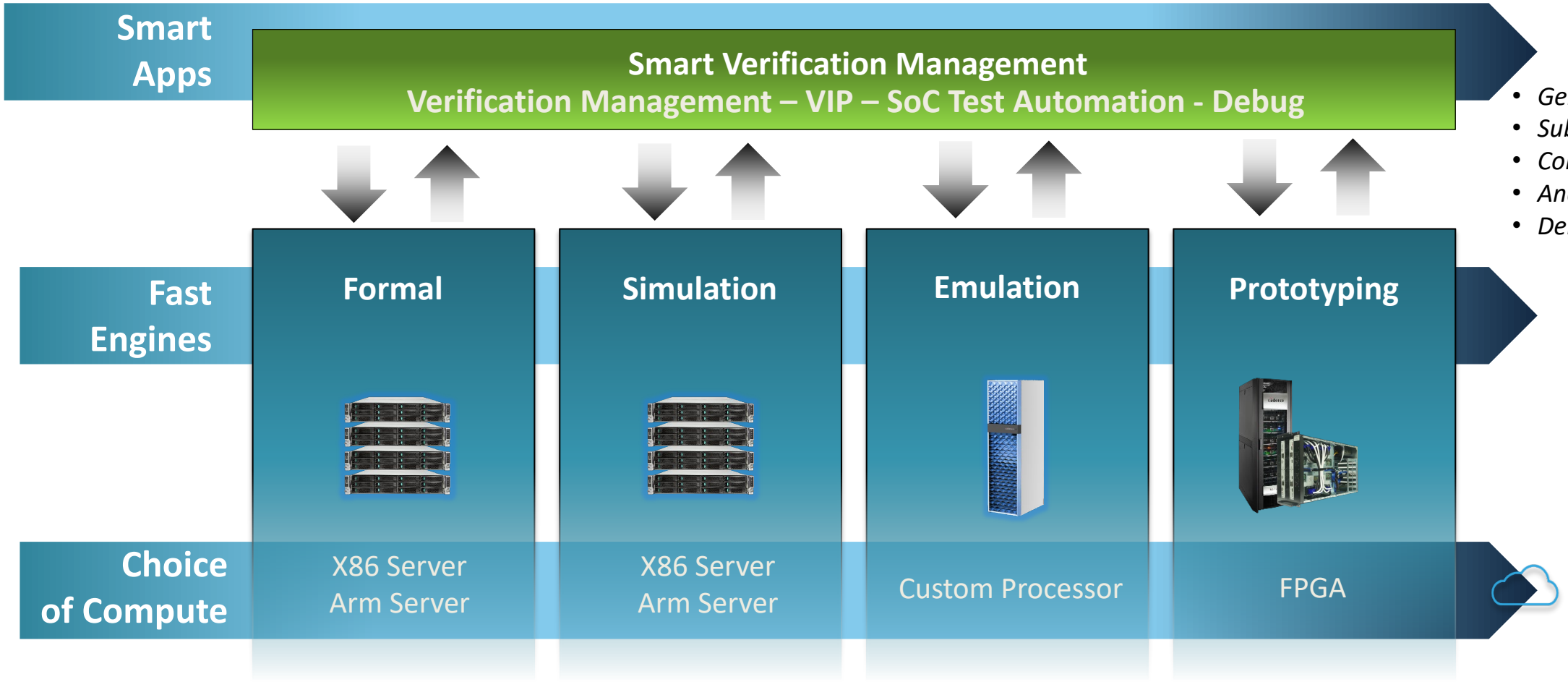
Prototyping

- **Scaling** to large designs
- Unified frontend
- ICE test suite: **Faster, data-extended (DL) regression**
- **SW driven validation** – deep learning data training refinement

© Cadence Design Systems, Inc - All Rights Reserved

Verification Throughput!

Find and fix the most bugs per $ invested in bare metal compute



- *Generate tests*
- *Submit jobs*
- *Collect coverage*
- *Analyze coverage*
- *Debug failures*

© Cadence Design Systems, Inc - All Rights Reserved

Some Examples

- Datacenter / Edge limited by compute power
- Accelerators
  - Custom to the ML application. No one size fits all.
  - HW/SW Co-validation at architectural level - performance using up to date user models
- Training/Inference in datacenter:
  - Massive # of systems - every milliwatt counts
  - HW/SW Co-validation with power analysis, user models
- Inference at Edge:
  - Very limited power budget
  - Validating power spec required before fabrication

habana	**Palladium Z1** "instrumental for Gaudi" Source: Cadence Earnings Call Q1'19
SambaNova Systems	**Palladium Z1** Source: Cadence Earnings Call Q1'19
FUNGIBLE	**HW Portfolio** Source: Cadence Earnings Call Q2'19
thinci blaize	**Full Verification Suite** Source: Press Release
NVIDIA	**HW Portfolio** – Capacity, Debug Source: Video @ DVCON Keynote

© Cadence Design Systems, Inc - All Rights Reserved