THE ELECTRONICS RESURGENCE INITIATIVE

Distribution Statement A- Approved for Public Release, Distribution Unlimited

LINTON SALMON

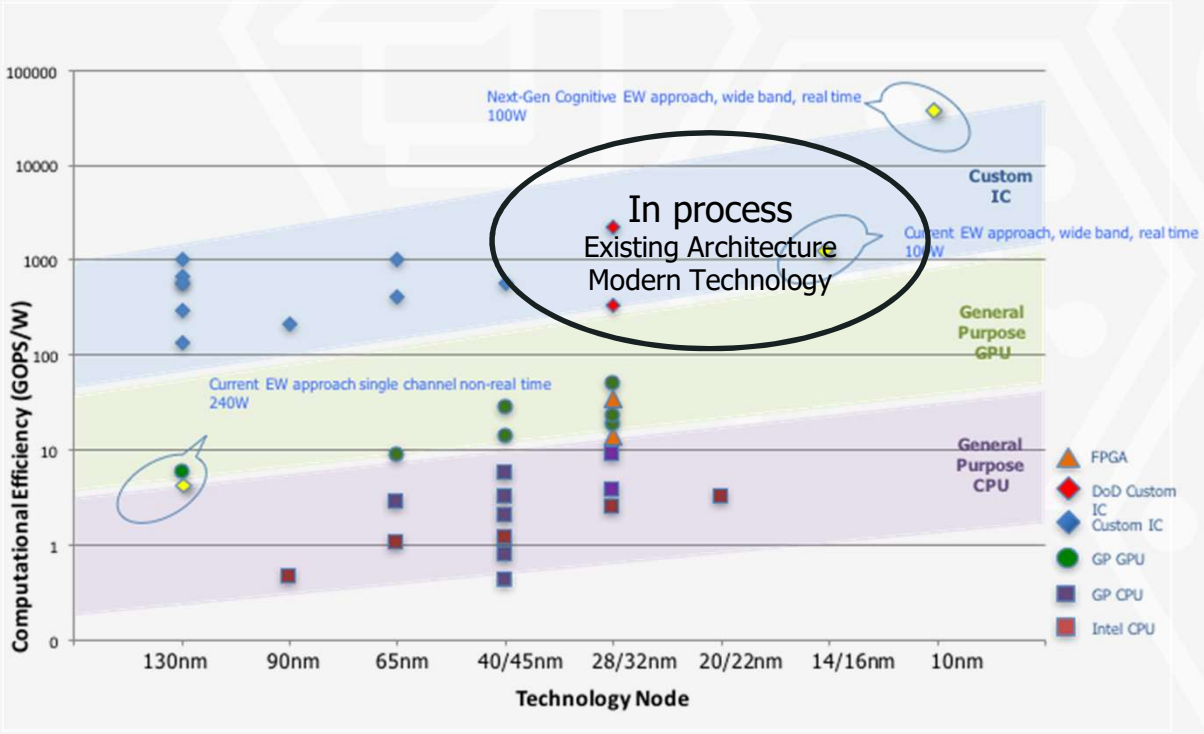
PROGRAM MANAGER
DARPA/MTO

Distribution Statement A – Approved for Public Release, Distribution Unlimited

CIRCUIT REALIZATION AT FASTER TIMESCALES (CRAFT)

Distribution Statement A- approved for Public Release, Distribution Unlimited

LEADING-EDGE CMOS PERFORMANCE IS CRITICAL FOR DOD



Distribution Statement A- Approved for Public Release, Distribution Unlimited

THE IMPROVEMENT ISN'T JUST REFLECTED IN GRAPHS

Circa 2001
130nm
https://www.pinterest.com

Circa 2016
16nm
https://www.verizonwireless.com

Cell phones have gotten much more capable

Would you take that brick phone for free?

Circa 2001
https://www.halloweencostumes.com/

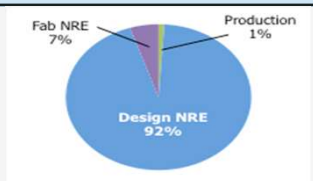
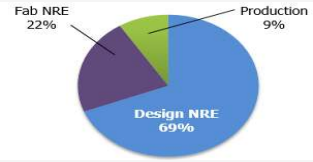
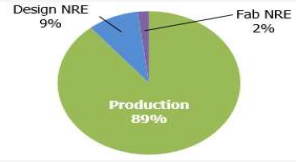
Circa 2014

Image processing and all types of processing have gotten better

Most people wouldn't even think about playing that 2001 game.

Distribution Statement A- Approved for Public Release, Distribution Unlimited

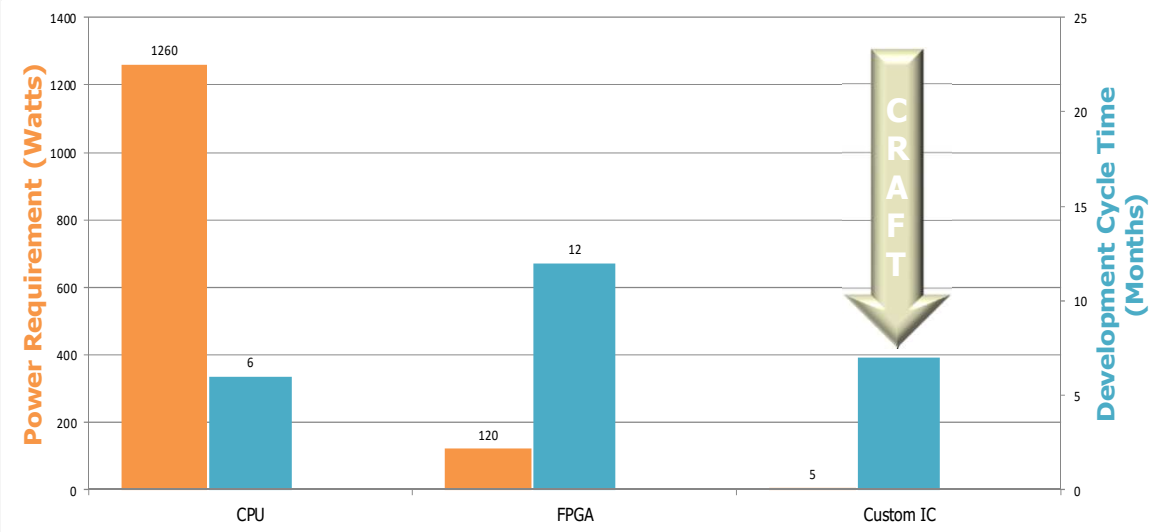
THE PROBLEM FOR DOD IS DESIGN COSTS, NOT FAB COSTS

	Low-Volume (DoD)	Moderate Volume Commercial	High Volume Commercial
			
Design Cost	Major contributor to total SoC cost	Major contributor to total SoC cost	Minor portion of total SoC cost
Fabrication Cost	Small contributor to total SoC cost	Significant contributor to total SoC cost	Major contributor to SoC cost
Volume	1k parts	1,000k parts	100,000k parts
Area	Not important	Relatively unimportant	Critical
Design Schedule/Risk	Critical	Critical	Critical
Performance at Power	Required	Required	Required

Distribution Statement A- Approved for Public Release, Distribution Unlimited

CRAFT (CIRCUIT REALIZATION AT FASTER TIMESCALES)

Today DoD have to choose between performance and schedule/cost.



Example Data from representative DoD design

To sharply reduce the barriers to DoD use of custom integrated circuits built using leading-edge CMOS technology while maintaining the high level of performance at power promised by this technology.

Distribution Statement A- Approved for Public Release, Distribution Unlimited

CRAFT: ENABLING USE OF THE BEST COMMERCIAL TECHNOLOGY

CRAFT provides solutions to the four major obstacles restricting custom IC design and fabrication for DoD systems.

DESIGN

- Design requires 18-24 months of effort
- Design verification takes far too much effort

CRAFT creates new design flows that will reduce custom IC design cycle time by ***10X***

PORT/ MIGRATE

- Designers are limited to one foundry
- Migration of designs from one foundry to another is difficult/ expensive

CRAFT uses new design flows to reduce porting effort by 5X and migrating effort by 50%

Leading-Edge CMOS Access

- DoD systems generally use lagging node CMOS technology
- Only 4 sources of leading-edge CMOS available world-wide

CRAFT provides access to 16/14nm technology through 4 dedicated, DARPA rapid turnaround time MPW runs

REPOSITORY

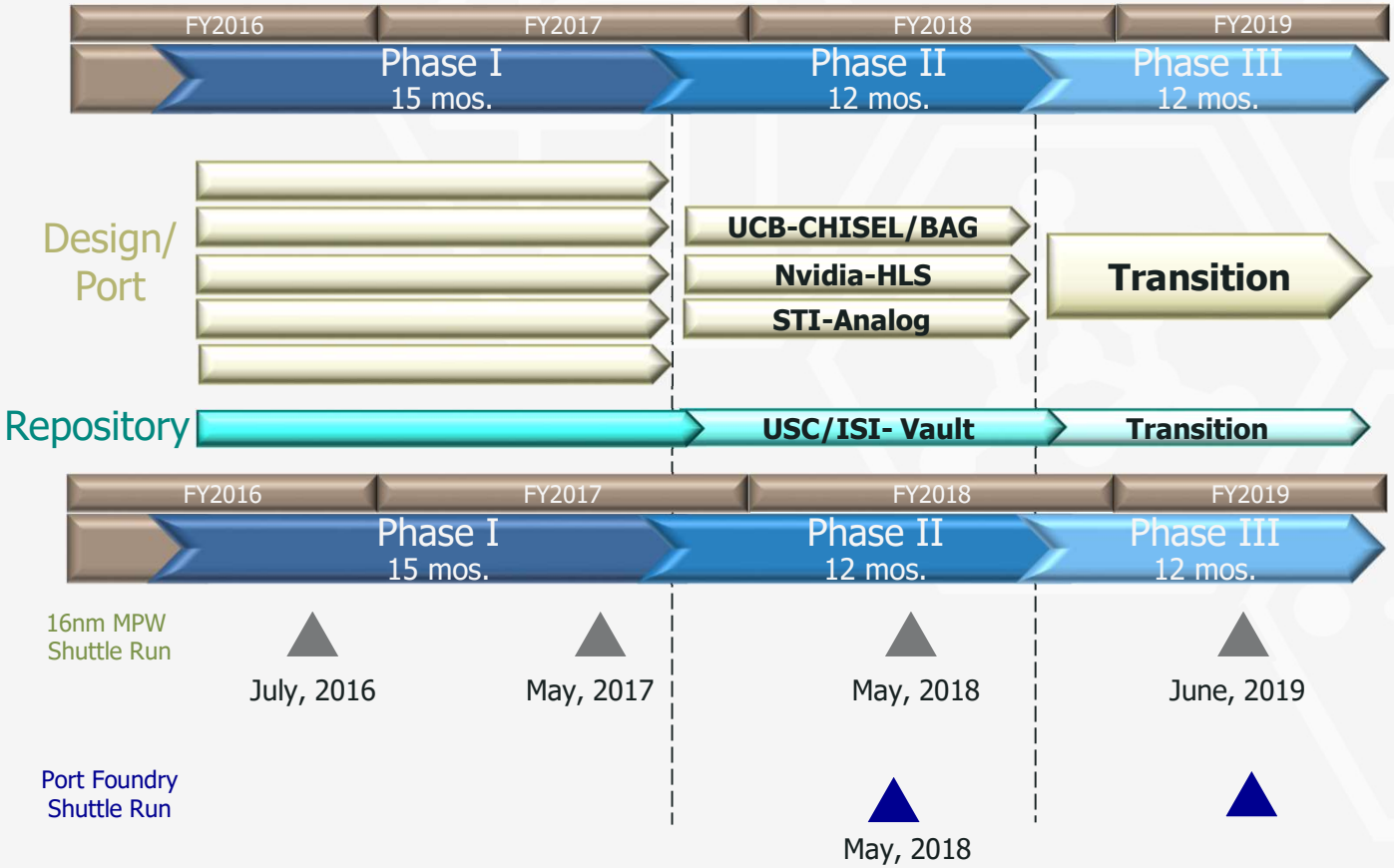
- Severe lack of IP reusability for DoD designs
- Current model for custom IC design/hardware security is broken

CRAFT establishes a data location and methodology to ensure 50% IP reuse by DoD performers

CRAFT will enable more efficient custom IC design/fabrication to enable HIGH performance electronic solutions FASTER and with more FLEXIBILITY

Distribution Statement A- Approved for Public Release, Distribution Unlimited

CRAFT PROGRAM PLAN



Distribution Statement A- Approved for Public Release, Distribution Unlimited

RESULTS TO DATE - DESIGN

Design Efficiency

Metric	Current Best in Class	End of Program Goal	Phase 2 Results		
			Nvidia	UCB	STI
Average Digital	3.5 kgates/eng-day	100 kgates/eng-day	41 kgates/eng-day	53 kgates/eng-day	NA
Peak Digital			120 kgates/eng-day	180 kgates/eng-day	NA
Average Analog	.2 blocks/eng-week	1.5 blocks/eng-week	NA	2.6 blocks/eng-week	6.3 blocks/eng-week
ADC Design	40 eng-weeks	4 eng-weeks	NA	4 eng-weeks	5.8 eng-weeks
Overall Design	100%	10%	9%	12%	11%

Representative SoC Complexity

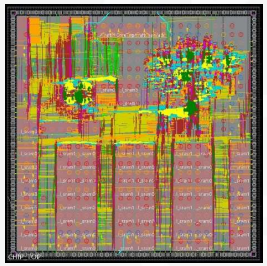
SoC Design	Digital Modules	Logic Size	Memory Size	Mixed-Signal	Die Size
UCB	8- Core Rocket 8-Vector Proessor	24.6M gates	29Mb	26Gb/s SERDES 8-bit ADC	25mm^2
Nvidia	DNN PE NoP Router	7.6M gates	6.4Mb	NA	6mm^2

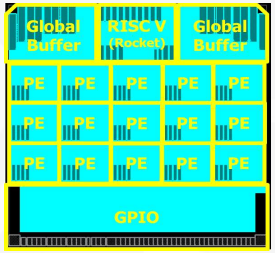
Large increases in digital and analog design efficiency demonstrated on moderately complex SoCs

Distribution Statement A- Approved for Public Release, Distribution Unlimited

RESULTS TO DATE - PORTING

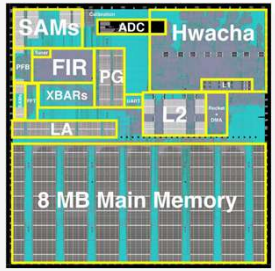
GF
14LPPXL

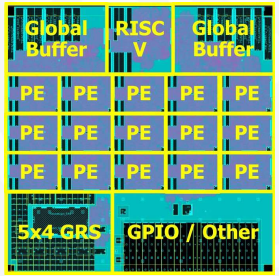




GF
14LPPXL

TSMC
16FF





TSMC
16FF

UC-Berkeley CRAFT P1 SoC

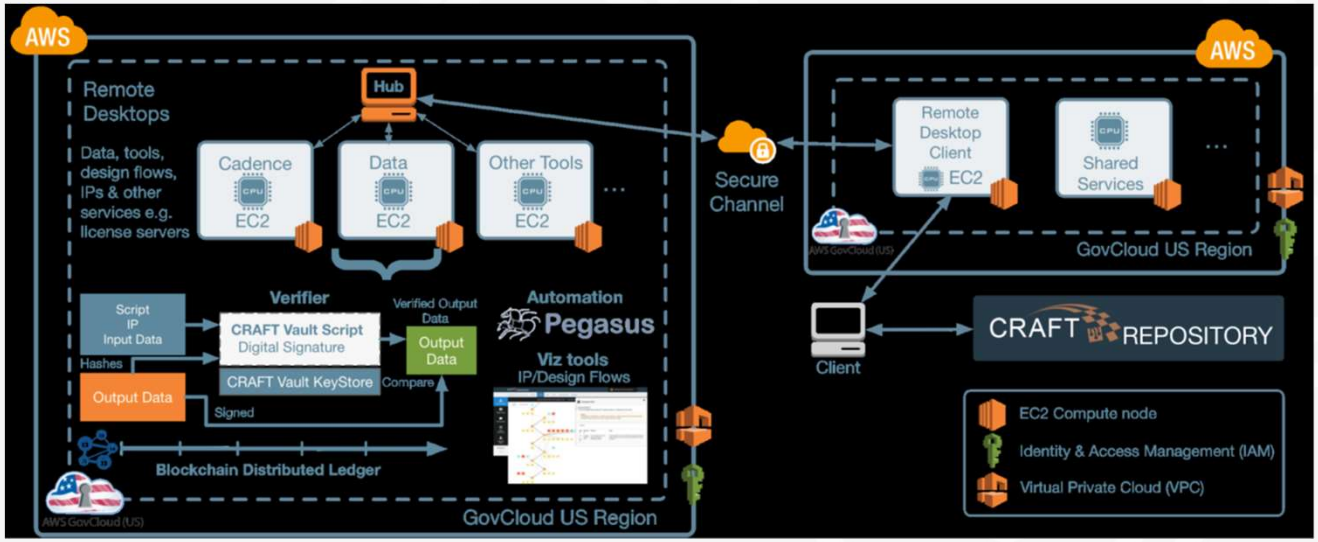
- Phase 1 hours: 14,000
- Phase 2 achievement: 2,663 hours
- Percentage: 19%

Nvidia RC17 SoC

- Phase 1 hours: 3216 hours
- Phase 2 achievement: 754 hours
- Percentage: 23%

Distribution Statement A- Approved for Public Release, Distribution Unlimited

RESULTS TO DATE – REPOSITORY/VAULT



From Ewa Deelman: USC-ISI

- The "Vault" will be a fully configured design environment for CRAFT flows
  - Will provide an environment with EDA tools, IP, and scripts set up for CRAFT flows
  - Currently in development by the USC/ISI and Notre Dame teams
  - Beta version available October, 2018
- NDAs and financial arrangements will be made external to the Vault

Distribution Statement A- Approved for Public Release, Distribution Unlimited

ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2018 | SAN FRANCISCO, CA | **JULY 23-25**

CRAFT: GENERATOR-BASED AGILE HARDWARE DESIGN

Distribution Statement A- approved for Public Release, Distribution Unlimited

ELAD ALON

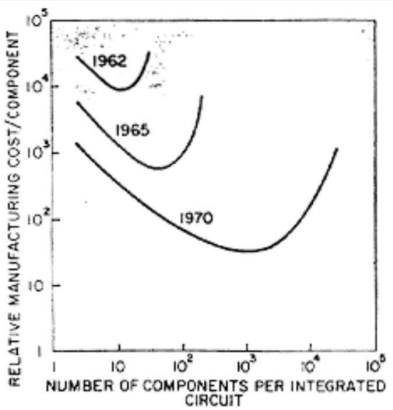
UC BERKELEY

- This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)
- The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government

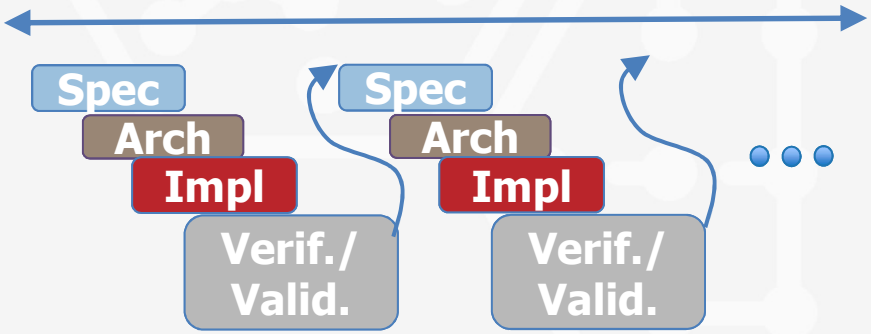
Distribution Statement A- Approved for Public Release, Distribution Unlimited

DESIGN COST AND AGILE DESIGN

- Development cost for a state-of-the art SoC is ~$300M
  - Need to sell 60M $5 chips for manuf. cost to equal design cost



- Software faced a similar challenge with ballooning NRE (complexity = $$$)
  - "Agile" approach developed in response, realized order-of-magnitude improvement



Distribution Statement A- Approved for Public Release, Distribution Unlimited

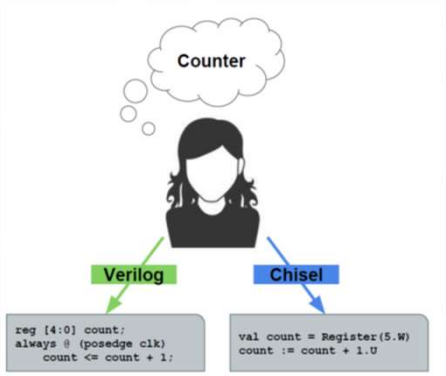
TRADITIONAL HARDWARE DESIGN ISN'T AGILE

- Dominant problem is dearth of re-use
  - Hardware IP of course is re-used, and expected to continue
  - But often building an SoC precisely because you want to specialize something…
- Our approach: **re-use the design process** itself, not the results of it
  - I.e., capture designers' methodologies in the form of executable **generators**
- Under CRAFT, our UC Berkeley, Cadence, Northrop-Grumman, and BAE team has been developing this approach and quantifying its benefits

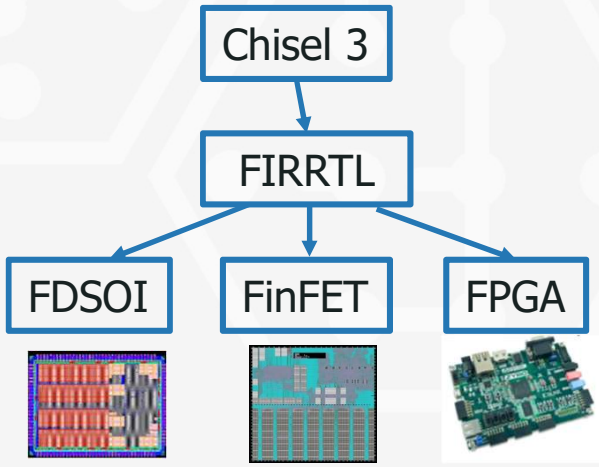
Distribution Statement A- Approved for Public Release, Distribution Unlimited

PLATFORM FOR DIGITAL GENERATORS

- **Chisel** (embedded within Scala) provides same level of designer control as RTL
  - I.e., not "high-level synthesis"
  - But does provide improved **software** abstractions for capturing methodologies



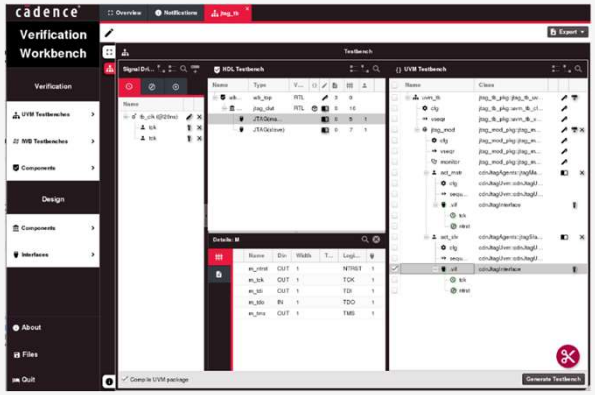
- Decouple generic vs. implementation-specific (process tech./ASIC vs. FPGA) optimizations to maximize re-use
  - Borrow from software again and use an intermediate representation (**FIRRTL**)

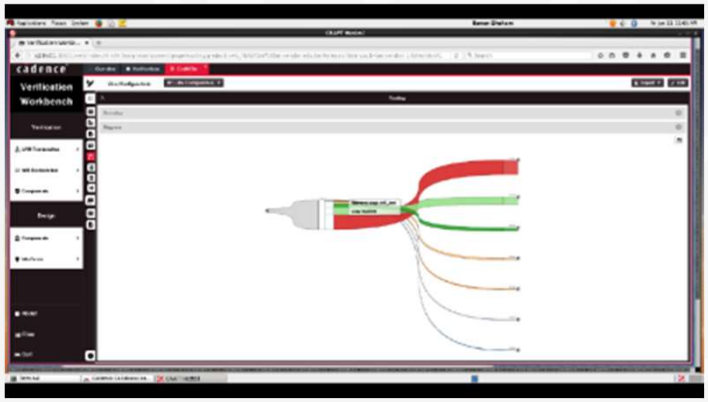


Distribution Statement A- Approved for Public Release, Distribution Unlimited

GENERATORS FOR VERIFICATION

- No one will bank their $300M SoC on “correct by construction”
  - So need to (agilely) verify any instance that you generate
- Cadence’s Verification Workbench (VWB) eliminates manual labor by automatically generating the test environment
  - VWB integrated into the CRAFT Chisel-based flow via IP-XACT metadata



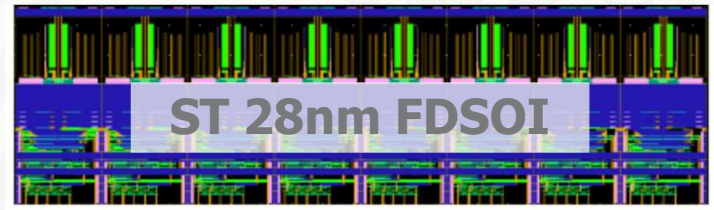


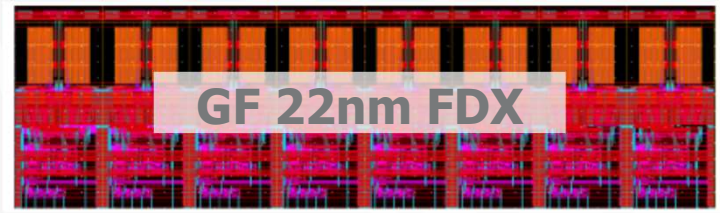
Distribution Statement A- Approved for Public Release, Distribution Unlimited

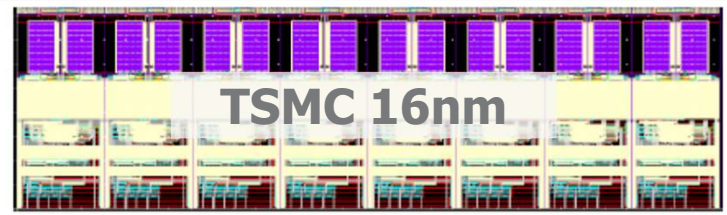
WHAT ABOUT ANALOG/MIXED-SIGNAL?

- Analog has traditionally been most resistant to automation
  - But requirements imposed by sub-20nm processes and analog complexity within SoCs have aligned to make generators very appealing
- Developed the **Berkeley Analog Generator (BAG)** as a Python-based process-portable framework enabling such generators

ADC Generator:





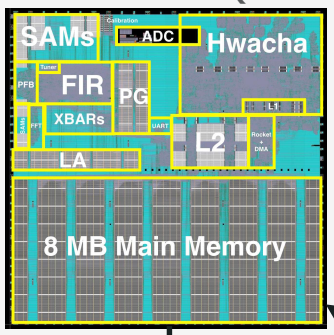


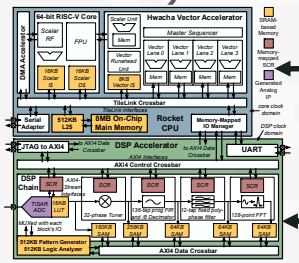
Distribution Statement A- Approved for Public Release, Distribution Unlimited

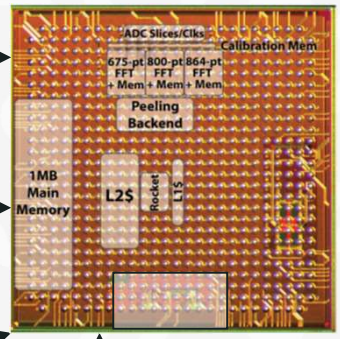
GENERATORS AND CHIPS (PHASE 1)

10.3M Gate Signal Analysis SoC
(TSMC 16nm)

4.2M Gate Sparse Recon. SoC
(TSMC 16nm)







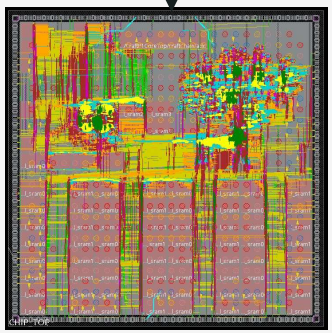
RISC-V Generator

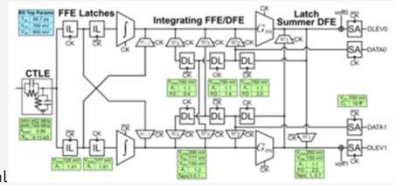
DSP Generator

Accel. Generators

ADC Generator

SerDes Frontend Generator

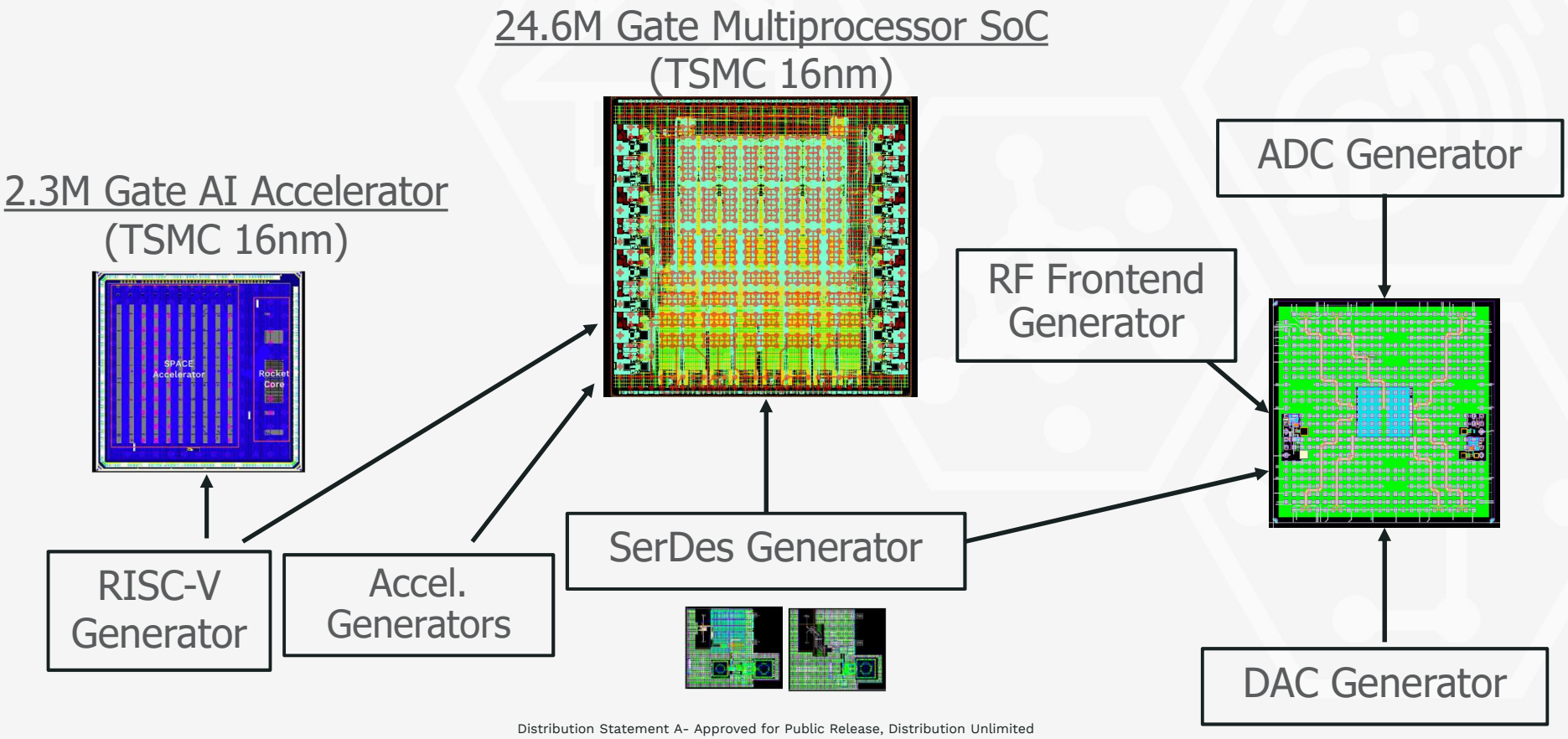




Ported to GF 14nm in <20% of original effort

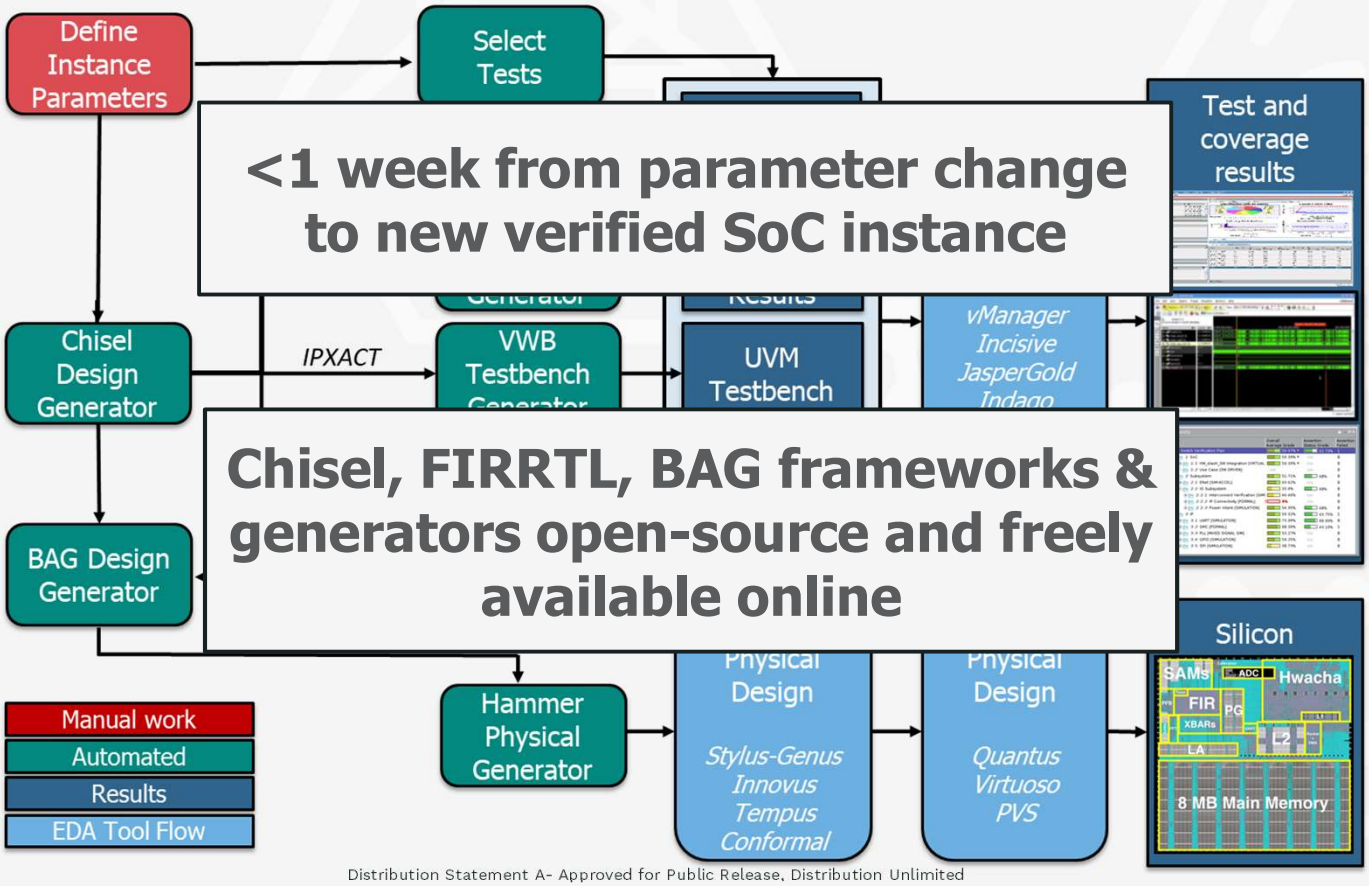
Distribution Statement A- Approved for Public Release, Distribution Unl

GENERATORS AND CHIPS (PHASE 2)



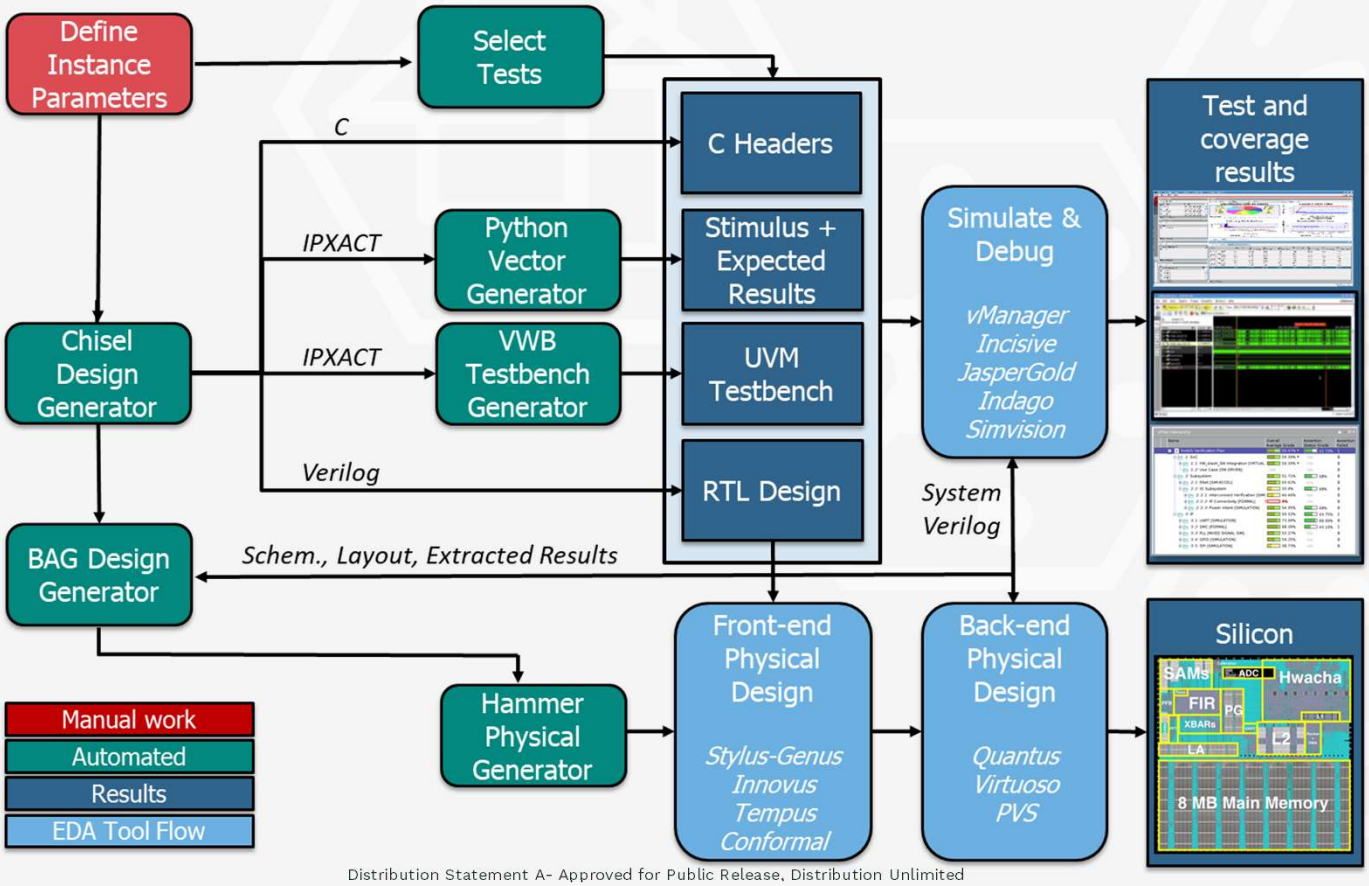
Distribution Statement A- Approved for Public Release, Distribution Unlimited

PUTTING IT ALL TOGETHER



Distribution Statement A- Approved for Public Release, Distribution Unlimited

PUTTING IT ALL TOGETHER



Distribution Statement A- Approved for Public Release, Distribution Unlimited

ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

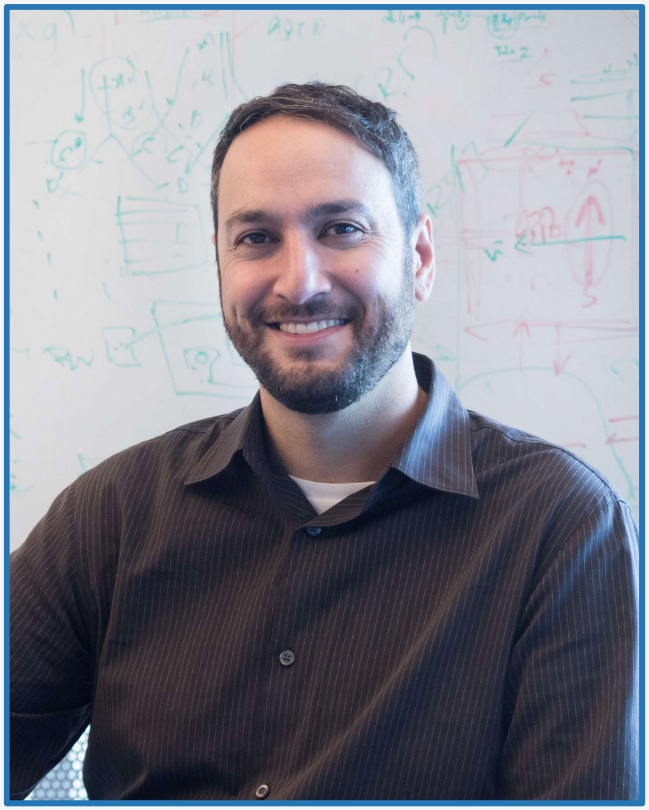
2018 | SAN FRANCISCO, CA | **JULY 23-25**

A MODULAR DIGITAL VLSI FLOW FOR HIGH-PRODUCTIVITY SOC DESIGN

NVIDIA

HARVARD UNIVERSITY

Distribution Statement A- approved for Public Release, Distribution Unlimited



BRUCEK KHAILANY

DIRECTOR OF RESEARCH,
ASIC & VLSI
NVIDIA CORPORATION

© NVIDIA 2018

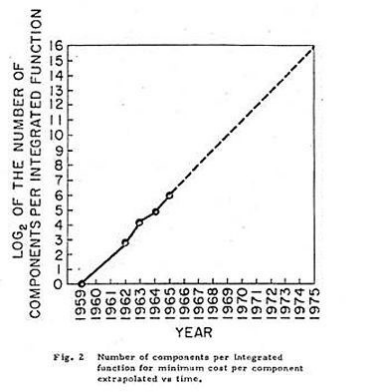
- This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)
- The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government

Distribution Statement A- Approved for Public Release, Distribution Unlimited

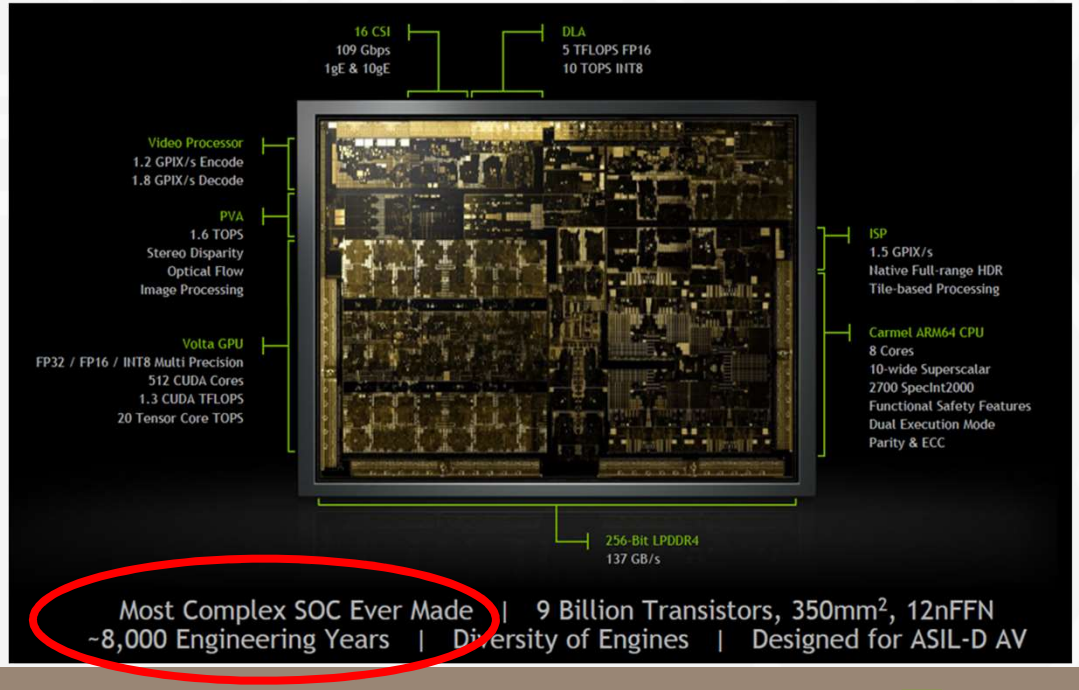
DESIGN COMPLEXITY

MORE TRANSISTORS: IMPROVED CAPABILITY AND DESIGN COSTS

[Electronics 1965]



NVIDIA Xavier Self-Driving Car SoC [CES 2018]

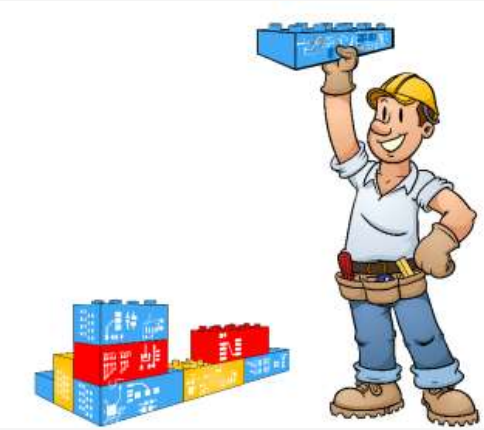


© NVIDIA 2018

RESEARCH AREAS

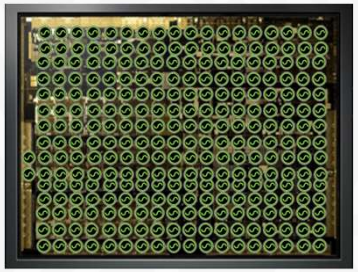
RTL DESIGN AND VERIFICATION

- Raise the level of design abstraction to C++ with High Level Synthesis (HLS) tools
- Libraries of commonly used hardware components in C++



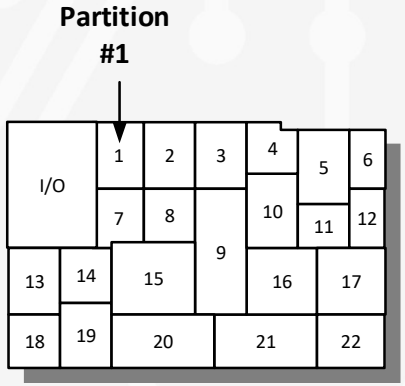
CLOCKING AND TIMING CLOSURE

- 1000s of distributed clock generators
- Correct-by-construction communication



FLOORPLANNING

- Small partitions for place-and-route tools with auto-generated floorplans



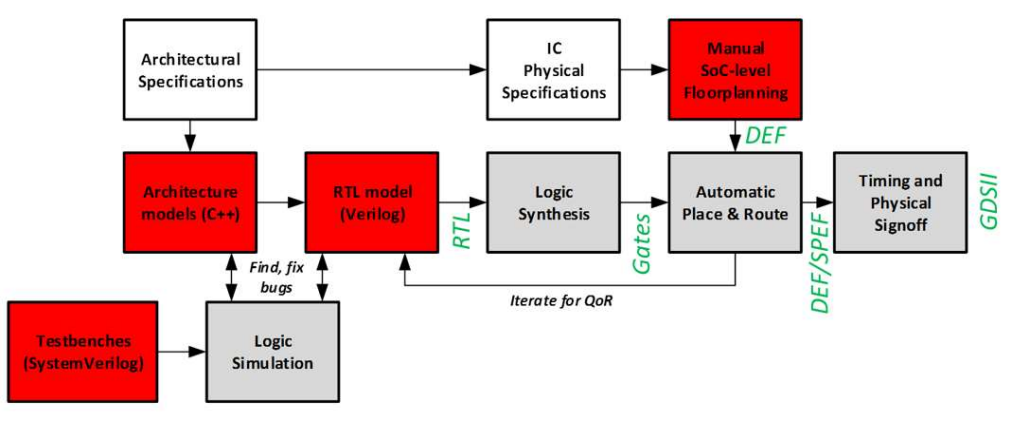
Distribution Statement A- Approved for Public Release, Distribution Unlimited

NVIDIA
© NVIDIA 2018

KEY PROBLEM: RTL DESIGN AND VERIFICATION

TODAY'S DESIGN ABSTRACTIONS ARE TOO LOW-LEVEL

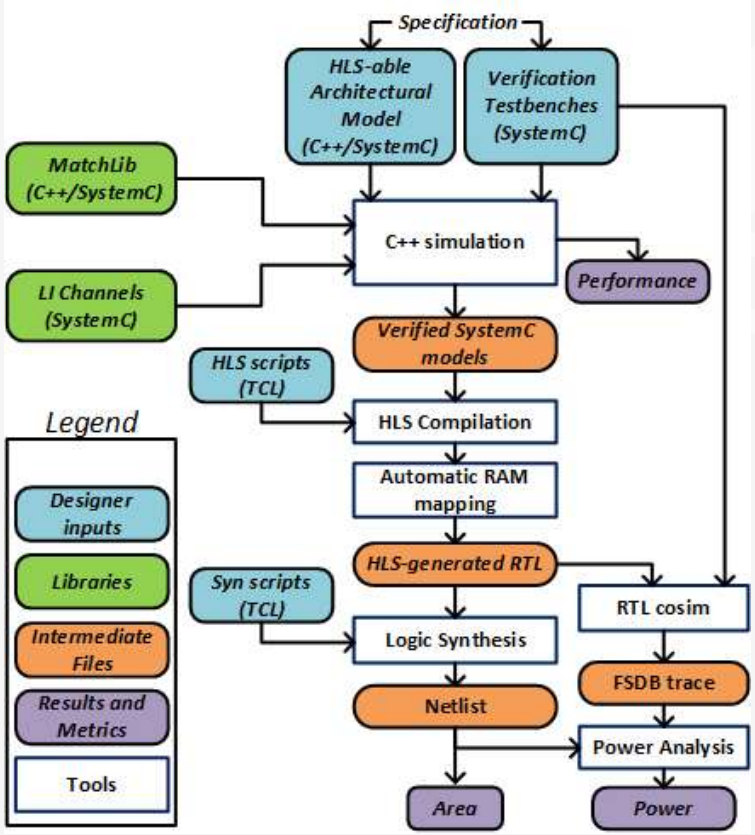
- Architects write simulators in C++, verify performance
- An ASIC design engineer rewrites the design in Verilog but can only test code with a simulator that runs 10^6 times slower than the hardware.



- An ASIC verification engineer makes sure that two models are 100% functionally equivalent before tapeout using SystemVerilog code only understood by ASIC verification experts

© NVIDIA 2018

PROPOSAL: OBJECT-ORIENTED HLS-BASED DESIGN



- Leverage HLS tools to design with C++ and SystemC models
- ***MatchLib:*** Library of commonly-used micro-architectural components in HLS-able C++
- All communication between SystemC modules through ***Latency-Insensitive Channels*** and/or on-chip networks
- Target: 10x productivity of manual RTL coding

Distribution Statement A- Approved for Public Release, Distribution Unlimited

© NVIDIA 2018

MATCHLIB – MODULAR APPROACH TO CIRCUITS AND HARDWARE LIBRARY

GOAL: "STL/BOOST" FOR HARDWARE

- Untimed C++ and loosely timed SystemC models
- Compatible with HLS for mapping to RTL
- Addresses a usability and Quality-of-Results gap with existing HLS toolflows
- Open source release planned in 2H 2018

	Component	Description
functions	nv_float	Floating-point arithmetic functions (mul, add, mul-add)
	Crossbar	N-to-N configurable bit width cross-bar switch
	Encoder/Decoder	1-hot encoders and decoders
classes	FIFO	Configurable FIFO
	Arbiter	1-out-of-N round-robin selector
	Mem_array	Abstract memory class
	nvhls_vector	Vector helper container w/ vector operations
	nvhls_array	Array helper container to work around Catapult issues
	Connections	Modular IO supporting latency-insensitive channels
	Arbitrated Crossbar	Crossbar with conflict arbitration and queuing
	Arbitrated Scratchpad	Scratchpad memories with arbitration and queueing
	Reorder Buffer	Out-of-order writes into queue, in-order reads
modules	Serializer/Deserializer	N-bit packets to/from M cycles of (N/M)-bit packets
	Cache	Parameterized by linesize, capacity, associativity
	Scratchpad	Banked memory array with crossbar
	DataStreamer	DMA-like functionality
	SFRouter	Store-and-Forward NoC router
	WHVCRouter	Wormhole NoC router with virtual channels
	AXI Components	Master/Slave Interfaces & bridges for AXI interconnect

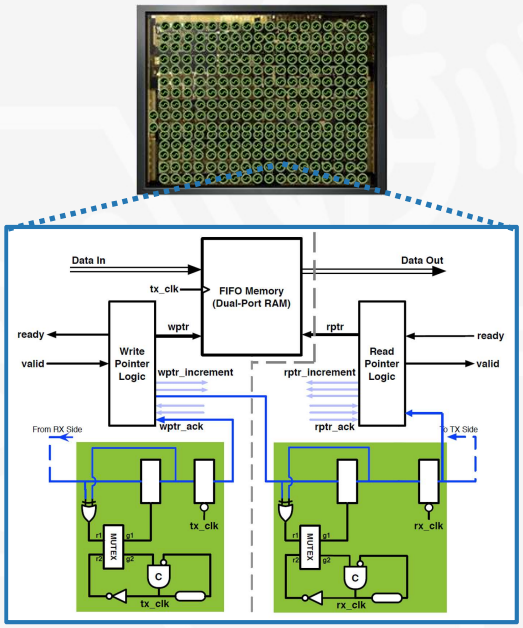
Distribution Statement A- Approved for Public Release, Distribution Unlimited

NVIDIA
© NVIDIA 2018

HIGH-PRODUCTIVITY CLOCKING METHODOLOGY

GLOBALLY ASYNCHRONOUS LOCALLY SYNCHRONOUS (GALS) ADAPTIVE CLOCKING

- Research Goals
  - "Correct by construction" top-level timing closure
  - Reduce voltage margin needed for power-supply noise
- Approach
  - 100s-1000s of on-chip per-partition adaptive clock generators
  - Low-latency error-free clock domain crossings



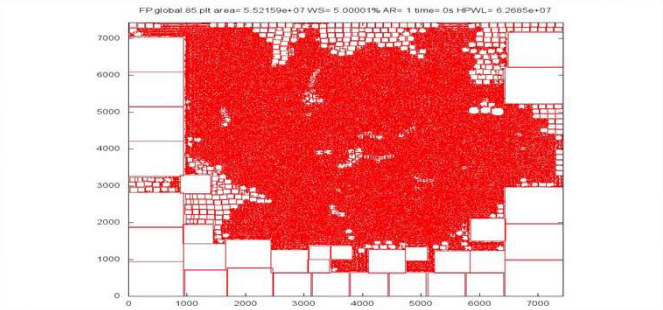
Pausible Bisynchronous FIFO for low-latency error-free interfaces
[Keller et al., ASYNC 2015]

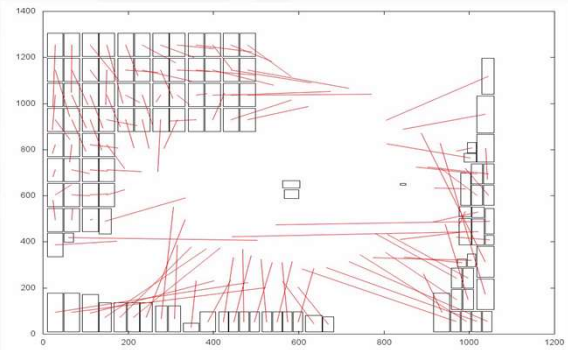
© NVIDIA 2018

PHYSICAL DESIGN AND FLOORPLANNING

SMALL AUTO-FLOORPLANNED PARTITIONS

- Research Goals
  - Demonstrate <24 hour iteration time through place-and-route tools
  - Develop auto-floorplanning algorithms for bounding box, pin placement, macro placement
- Approach
  - Minimize wirelengths and overlaps via gradient descent and multilevel clustering
  - RAM legalization via simulated annealing





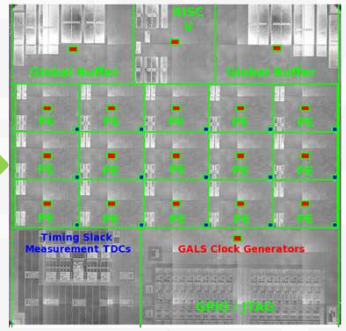
© NVIDIA 2018

SOC TESTCHIP METHODOLOGY DEMONSTRATIONS

GOALS: DEMONSTRATE PRODUCTIVITY AND PERF/AREA/POWER

RC17

Programmable ML Inference Accelerator in TSMC 16FF+



RC18

Optimized DL Inference Accelerator

June 2018 Tapeout

2016 2017 2018

RC17b

Port to GF 14nm

Project management approach

- Agile hardware design
- Daily "C++-to-layout" spins

NVIDIA
© NVIDIA 2018

SUMMARY AND FUTURE WORK

DESIGN METHODOLOGY R&D CAN SOLVE THE COMPLEXITY PROBLEM

- Summary
  - RTL and verification effort reduction from raising abstraction levels and leveraging reusable libraries
  - Rethink design best practices (e.g. fine-grained GALS clocking)
  - EDA tool and algorithms research
- Future work
  - Release MatchLib open source, refine and expand libraries
  - Continue EDA algorithms research
    (e.g. GPU-accelerated EDA, ML-based techniques)

Distribution Statement A- Approved for Public Release, Distribution Unlimited

© NVIDIA 2018

ACKNOWLEDGEMENTS

- **Research sponsored by DARPA under the CRAFT program (PM: Linton Salmon).**
- **NVIDIA Collaborators:** Evgeni Krimer, Rangharajan Venkatesan, Jason Clemons, Ben Keller, Matthew Fojtik, Alicia Klinefelter, Angshuman Parashar, Michael Pellauer, Nathaniel Pinckney, Mark Ren, Yakun Sophia Shao, Stephen Tell, Yanqing Zhang, Brian Zimmer, Bill Dally, Joel S. Emer, Stephen Keckler
- **Harvard Collaborators:** David Brooks, Gu-Yeon Wei, and team
- **Former NVIDIA interns/postdocs:** Christopher Fletcher, Ziyun Li, Antonio Puglielli, Shreesha Srinath, Gopal Srinivasan, Chris Torng, Sam (Likun) Xi
- *Thanks to the **Mentor Graphics Catapult HLS team** for discussions and support:* Bryan Bowyer, Stuart Clubb, Moises Garcia, and Khalid Islam.

This research was, in part, funded by the U.S. Government, under the DARPA CRAFT program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

Distribution Statement A- Approved for Public Release, Distribution Unlimited

© NVIDIA 2018

ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2018 | SAN FRANCISCO, CA | **JULY 23-25**