AI Division
Advanced Computing Lab

Updated September 2022

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM22-0851

Advanced Computing Lab Team

Dr. John Wohlbier
Senior Research Scientist
& Principal Investigator

Dr. Scott McMillan
Principal Research Engineer

Annika Horgan
Associate Software Developer

Dr. Franz Franchetti
Associate Dean for Research, CIT

Dr. Tze Meng Low
Assistant Research Professor

Dr. Drew Dolgert
Senior Research Scientist

Collin Abidi
Associate Software Developer

David Graham
Senior Project Lead

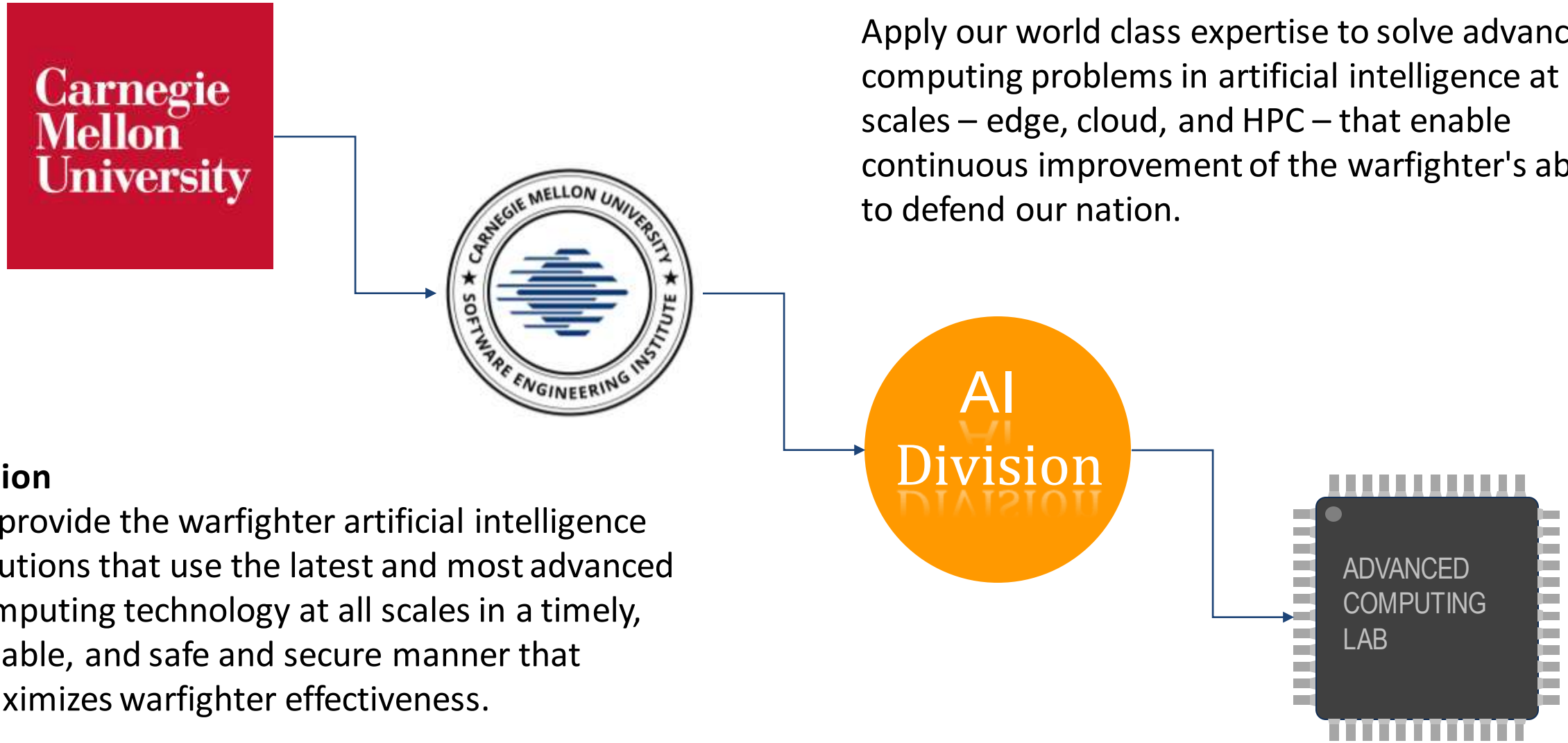
Dr. Ken Mai
Principal Systems Scientist

Mission

Apply our world class expertise to solve advanced computing problems in artificial intelligence at all scales – edge, cloud, and HPC – that enable continuous improvement of the warfighter's ability to defend our nation.

Vision

To provide the warfighter artificial intelligence solutions that use the latest and most advanced computing technology at all scales in a timely, reliable, and safe and secure manner that maximizes warfighter effectiveness.

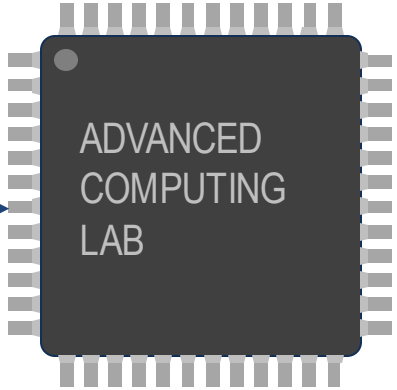


Carnegie Mellon University

Mission

- DARPA T&E
- Domain specific SoCs
- Performant code at the edge
- Homomorphic encryption
- Co-design of edge software-hardware

AI Division



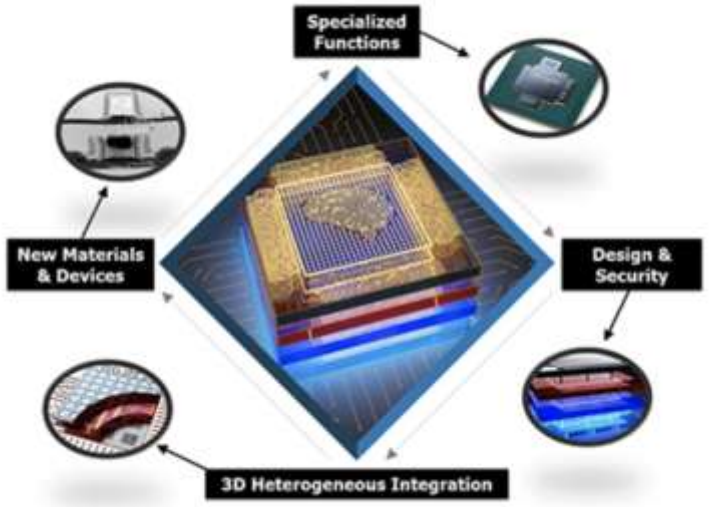
Vision

- Cognitive radio
- RF situational awareness
- JADC2
- Satellite communications
- Data analytics at scale

Portfolio

Test and Evaluation

- DARPA Domain-Specific System on Chip (DSSoC)
- DARPA Data Protection in Virtual Environments (DPRIVE)
- DARPA Software Defined Hardware (SDH)



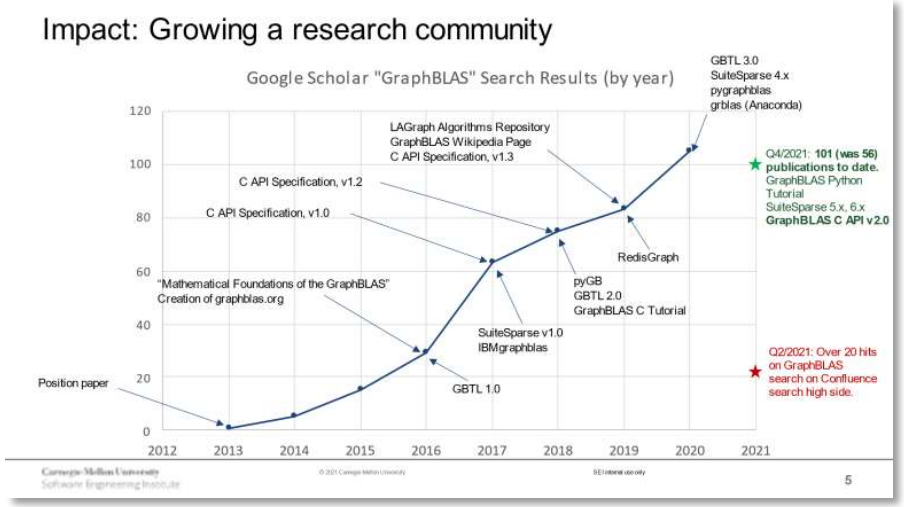
Internal Research

- Portable, High-performance Inference at the Tactical Edge (PHITE)
- Co-design for Edge AI
- Spiral AI/ML
- Spiral Graph

UAS with Prototype STORM Payload

Technical Advisory

- Test, evaluate, monitor, and provide expertise for entity developing a GraphBLAS implementation
- Advising GraphBLAS implementations on new hardware



Portfolio

Test and Evaluation

- DARPA Domain-Specific System on Chip (DSSoC)
- DARPA D… Virtual En…
- DARPA S… Hardware…

Internal Research

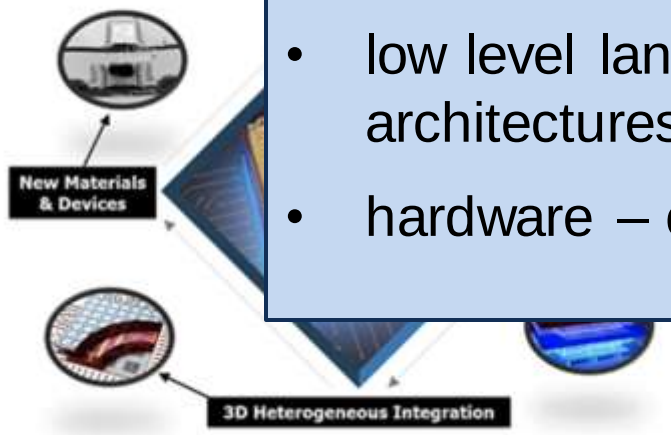
- Portable, High-performance Inference at the Tactical Edge

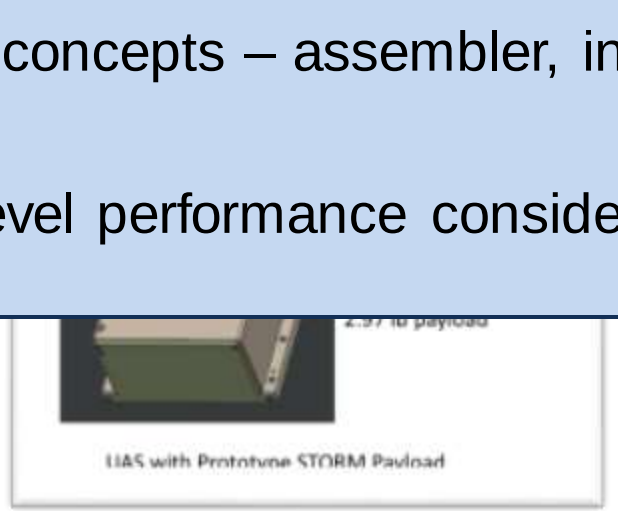
Technical Advisory

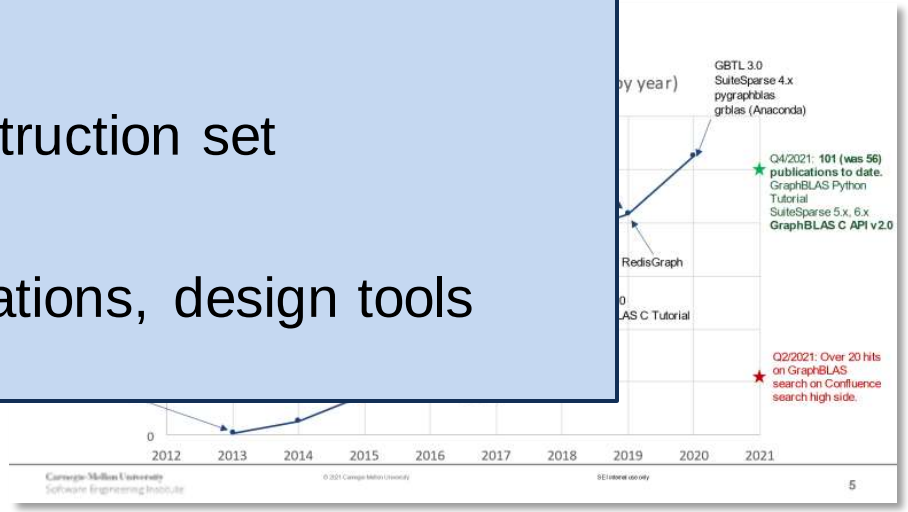
- Test, evaluate, monitor, and provide expertise for …ntity …LAS …ew

Full stack expertise

- algorithms – graphs, AI/ML, etc
- high level programming languages – Python, C, C++
- compiler technologies – abstract syntax trees, intermediate representations
- low level language concepts – assembler, instruction set architectures
- hardware – cycle level performance considerations, design tools

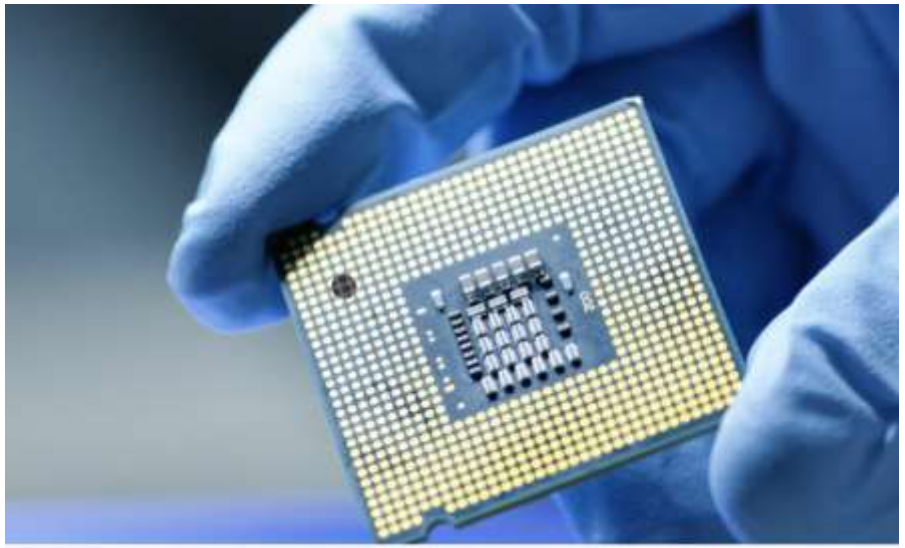


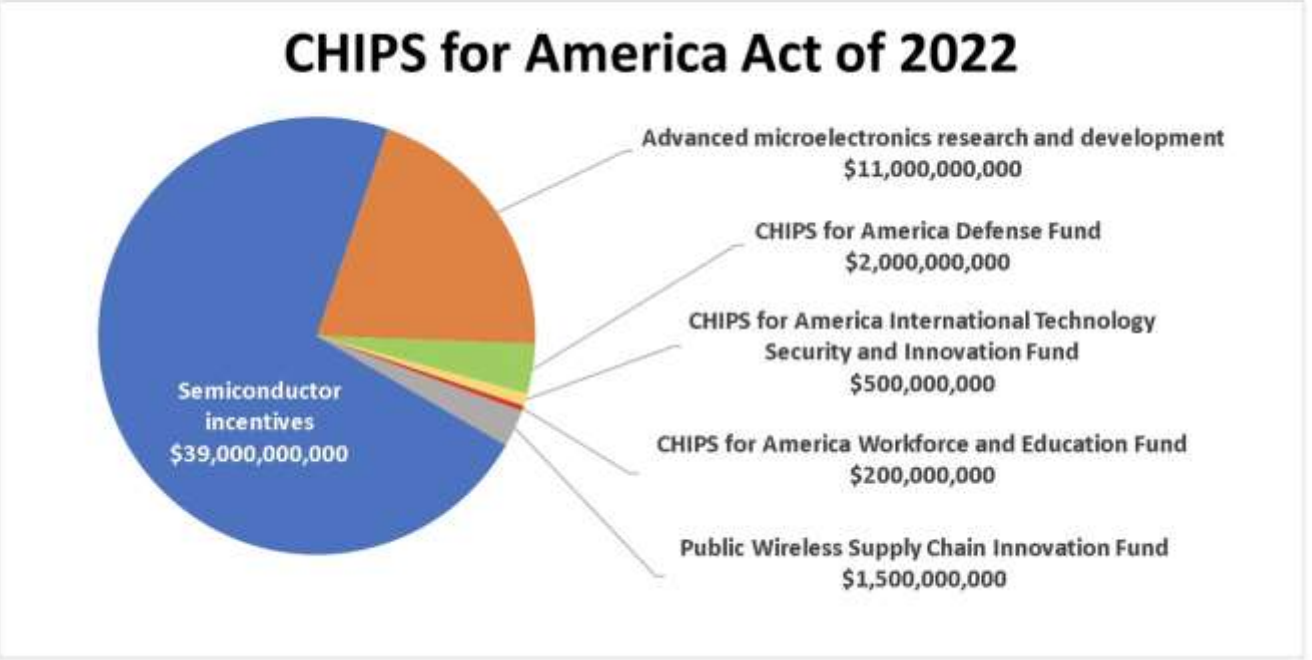




A focus on microelectronics

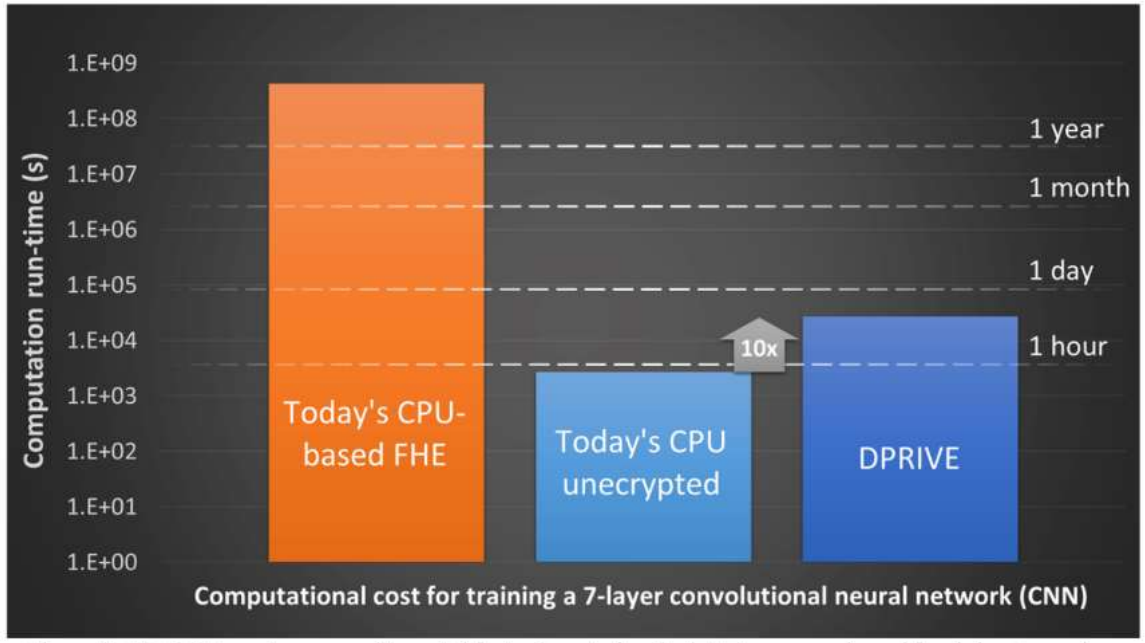
- A national security priority
- AI runs on advanced microelectronics
- TSMC accounts for 92% of advanced chip manufacturing
- US leads in design, but China is investing heavily
- $52B CHIPS act
  - $39B for “incentives” (i.e., fabs)
  - $11B for R&D
  - $200M for Workforce and education
  - $2B for DoD

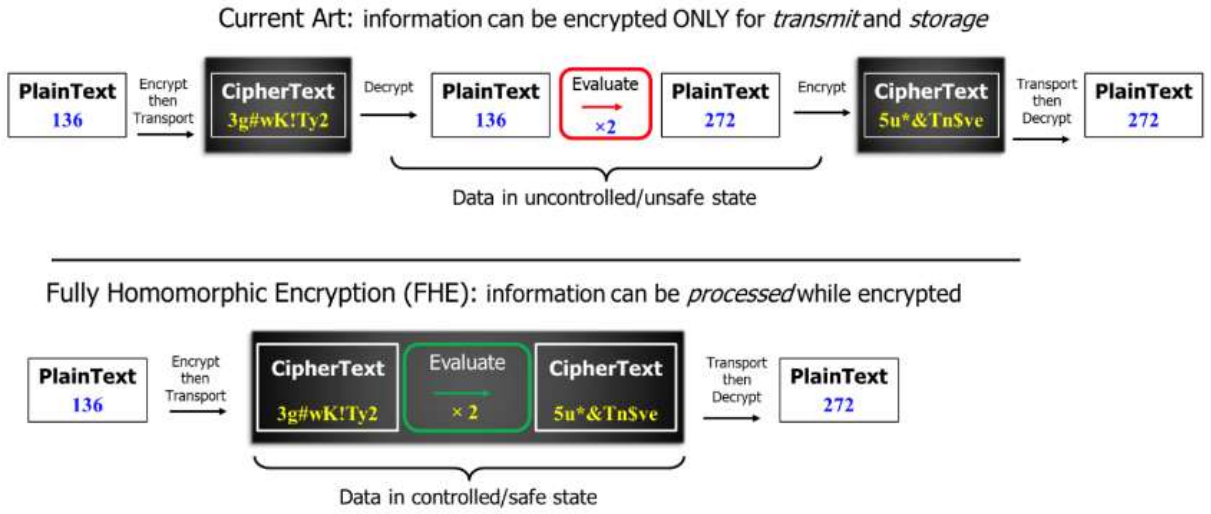




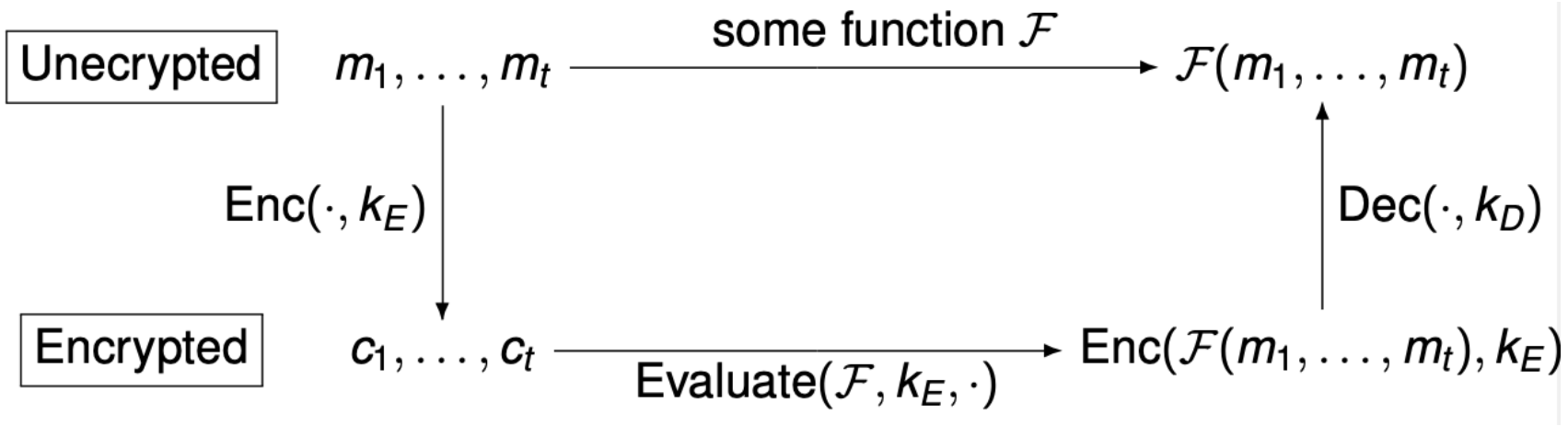
Data PRotection In Virtual Environments (DPRIVE)

- Hardware accelerators for fully homomorphic encryption
  - Computing on encrypted data
- Develop benchmark code for performer evaluation
  - Map benchmarks to performer instruction set architectures
  - Evaluate performer designs and report to program manager





Introduction to Fully Homomorphic Encryption

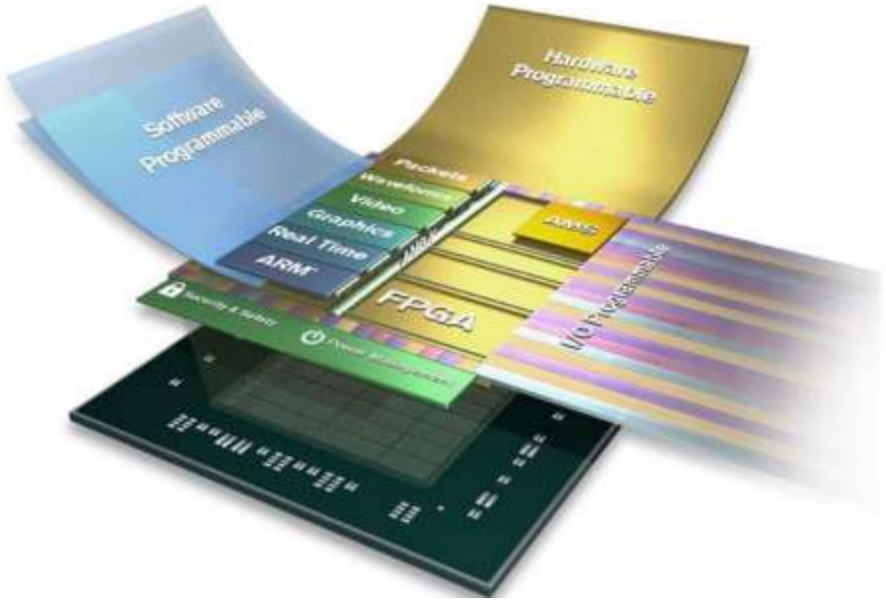


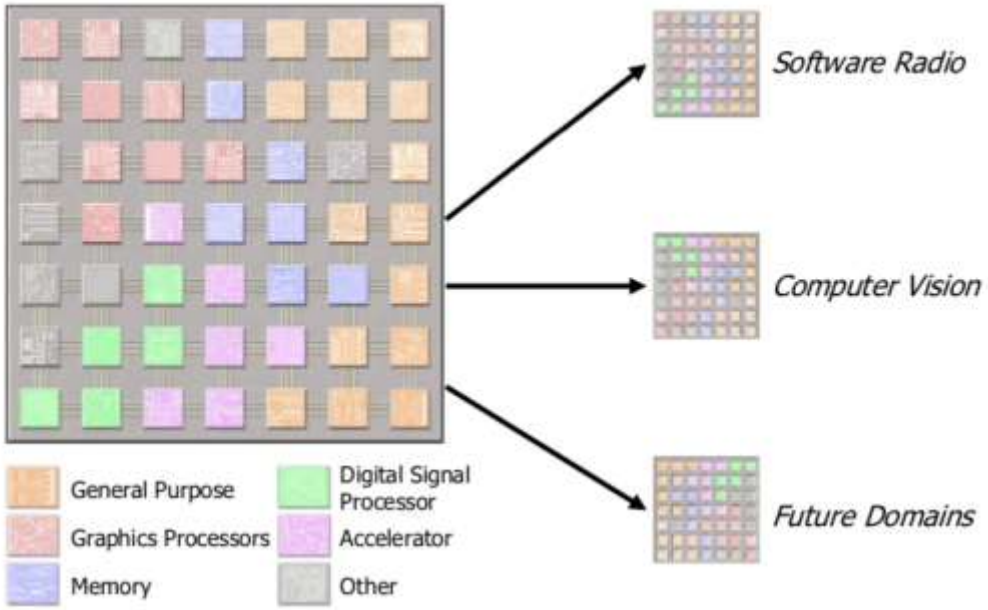
- Cryptography based on underlying problem that is hard to solve unless have the key
  - RSA – based on hardness of factoring two large prime numbers
  - (BGV) FHE – based on hardness of Ring Learning With Errors (RLWE) problem
  - Such lattice based constructions are considered "post-quantum" in the sense that they will not be breakable by a quantum computer

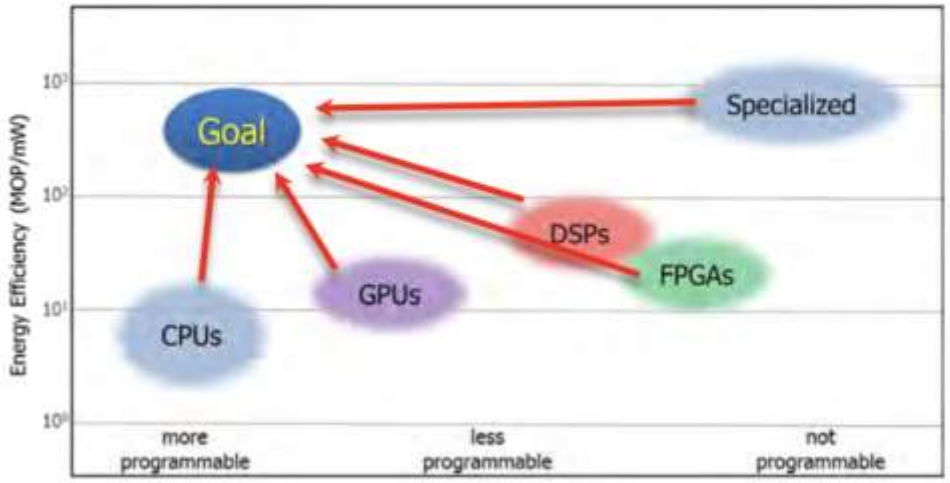
> An important feature of basing cryptography on the ring learning with errors problem is the fact that the solution to the RLWE problem can be used to solve the NP-hard shortest vector problem (SVP) in a lattice (a polynomial-time reduction from the SVP problem to the RLWE problem has been presented[1]).

DSSoC

- Domain Specific System on Chip enabling rapid development of multi-application, heterogeneous systems through a single programmable device
- Expert first users of performer tools covering device ontology for novel accelerators, intelligent scheduling and runtimes

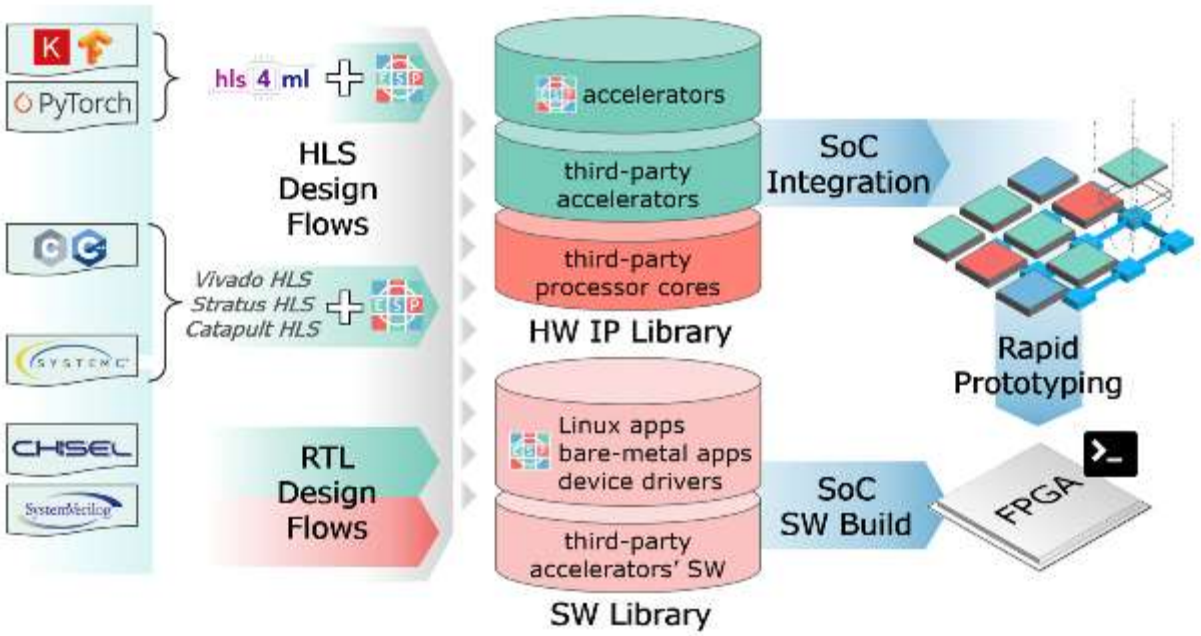






Designing a DSSoC

- Using ESP from Columbia University to design SoC accelerators via multiple design flows
  - hls4ml – TensorFlow, Pytorch to accelerator designs
  - Vivado, Stratus, Catapult HLS



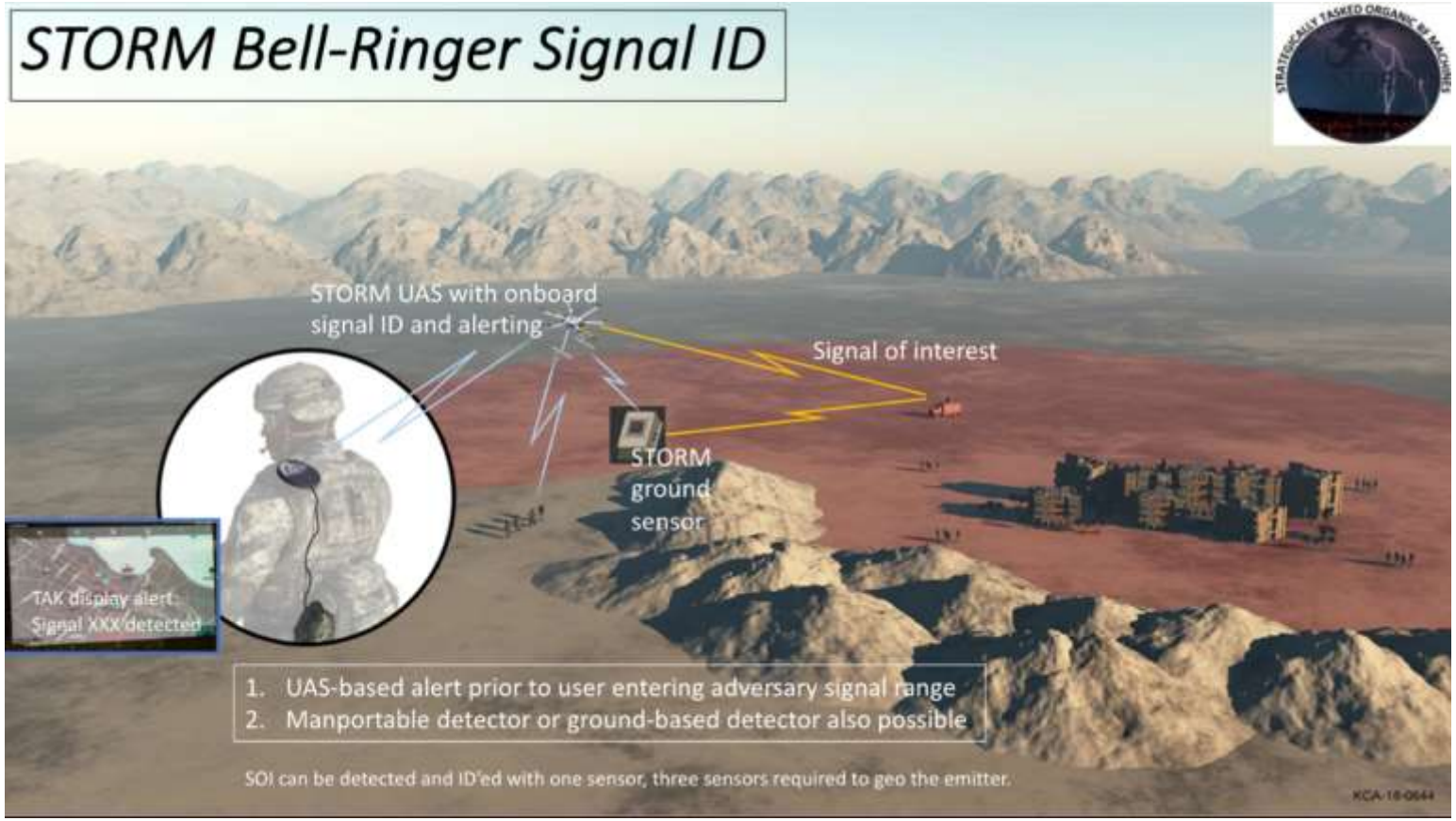
Co-design for Edge AI

"In general, ***compute*** can improve mission time and lower energy consumption by as much as 5X."

Boroujerdian, Behzad, et al. "Why compute matters for UAV energy efficiency?" (2018)

Objective 1: Application dependent SoC co-designs for stakeholders with accelerators that are 2x more energy efficient than GPU solutions and are 30% better in mission metrics.

Objective 2: A co-design flow methodology and **codriver** implementation that includes four tools per pipeline stage and enables co-design of RTL for an application that includes accelerator IP *within two months*.



Co-design for Edge AI

"In general, ***compute*** can improve mission time and lower energy consumption by as much as 5X."

Boroujerdian, Behzad, et al. "Why compute matters for UAV energy efficiency?" (2018)

Objective 1: Application dependent SoC co-designs for stakeholders with accelerators that are 2x more energy efficient than GPU solutions and are 30% better in mission metrics.

Objective 2: A co-design flow methodology and **codriver** implementation that includes four tools per pipeline stage and enables co-design of RTL for an application that includes accelerator IP *within two months*.

DARPA STORM UAV becoming a USASOC program of record

- STORM (Strategically Tasked Organic Radio Frequency Machine) Applications
  - Spectrum survey for support of the kill chain
  - RF mapping
  - Tactical alerting
  - Wideband scanning
  - Geolocation of signal sources
- ML-based signal identification characterization
- Cyclostationary signal identification
  - Heavy load of matrix multiplies
- DARPA – Strategic Technology Office

UAS with Prototype STORM Payload

What is a co-design flow?

- The co-design flow is made up of all the software components developed for and used to do the tasks, as well as documented procedures for executing the tasks.
- The flow documents how to concurrently design hardware and software components of complex electronic systems.
  - The co-design flow is captured in the script **codriver**

1. Characterization of application kernels to determine resource-intensive components.
2. Co-design a compressed ML model and a corresponding accelerator
3. Define a System on Chip (SoC) Co-design flow that produces hardware description language (HDL) specification
4. Simulate efficiency as a function of SoC efficiency and Size Weight and Power (SWaP) to showcase mission capability

PHITE: Portable, High-performance Inference at the Tactical Edge

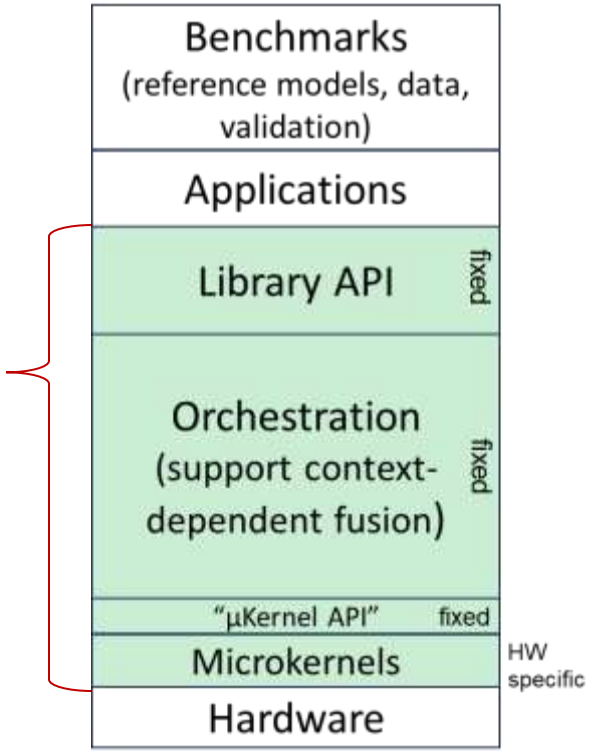
Research topic: Performance modeling and abstractions (µKernel APIs) for performance-portable neural network libraries with focus on low-power devices.

Current Approach:

- Library APIs: Tensorflow API and MLIR integration
- Orchestration: Efficient in-memory data layout and loop structure
- Microkernels: Small amount of platform-specific code

Results and Impact:

- Memory savings and layer fusion results in faster code and smaller footprint enabling larger models in resource-constrained environments



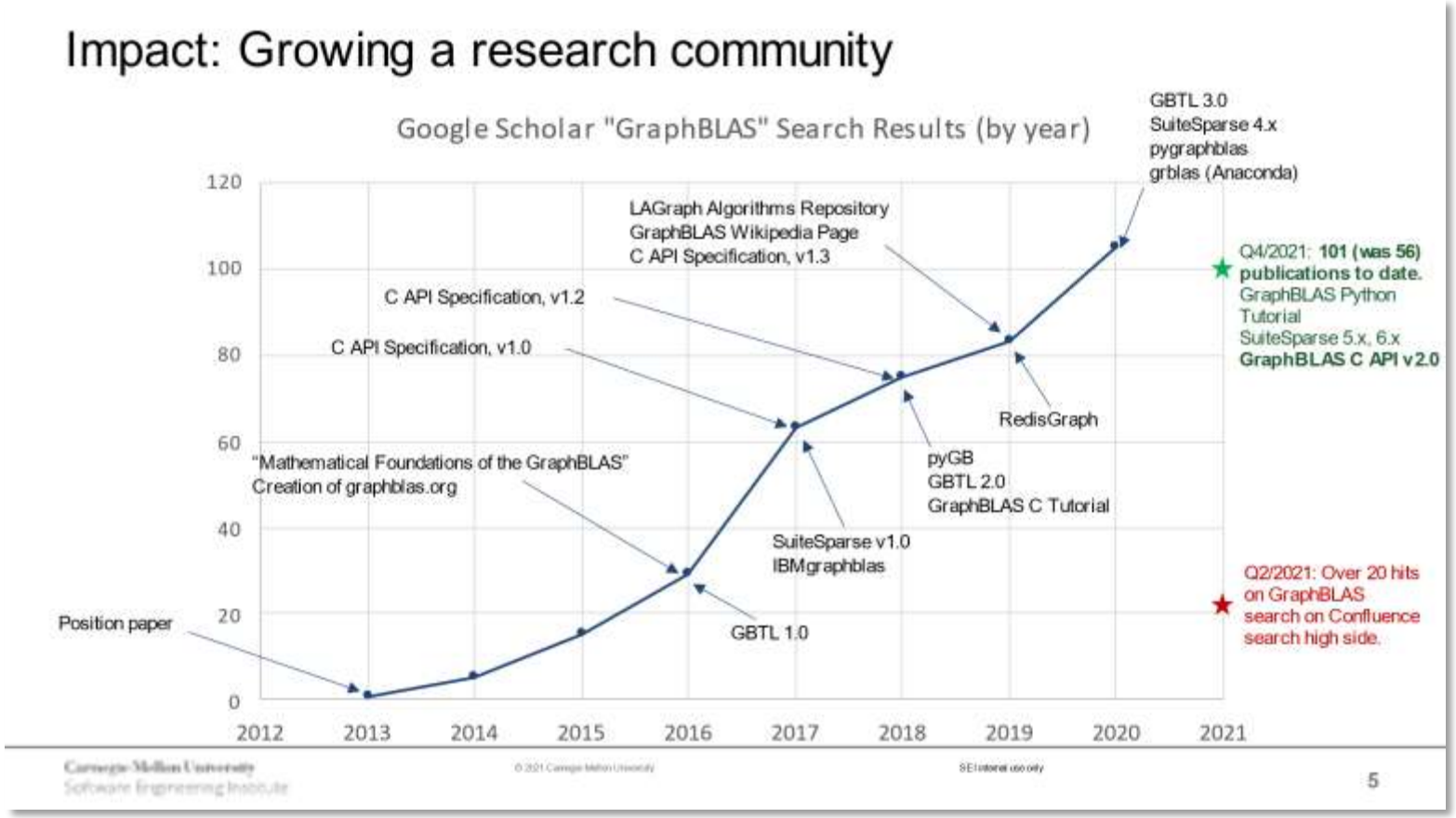
System:	Raspberry Pi Model 4-B SoC	Raspberry Pi PICO Microcontroller
Power	~3W (idle) – 7W (4 cores)	6mW - 330mW
Cost	$35-$75	$4
Peak	13.5 gigaFLOPS	266 megaFLOPS
Model/Size	AlexNet / 60M parameters	MobileNet V2 / 3M parameters

Prof. Tze Meng Low, CMU/ECE, co-PI

High-performance algorithms using formal methods and analytical models; performance portability through capture of interaction between software algorithms and hardware features; code generation and libraries for emerging domains.

GraphBLAS

- API specification of standard building blocks for graph algorithms in the language of linear algebra
- Describes how graph operations can be efficiently implemented via linear algebraic methods
- Graphs are everywhere: social networks, mobile networks, energy grid
- Types of computations: centrality analysis, community detection, connectivity analysis, path analysis, link prediction

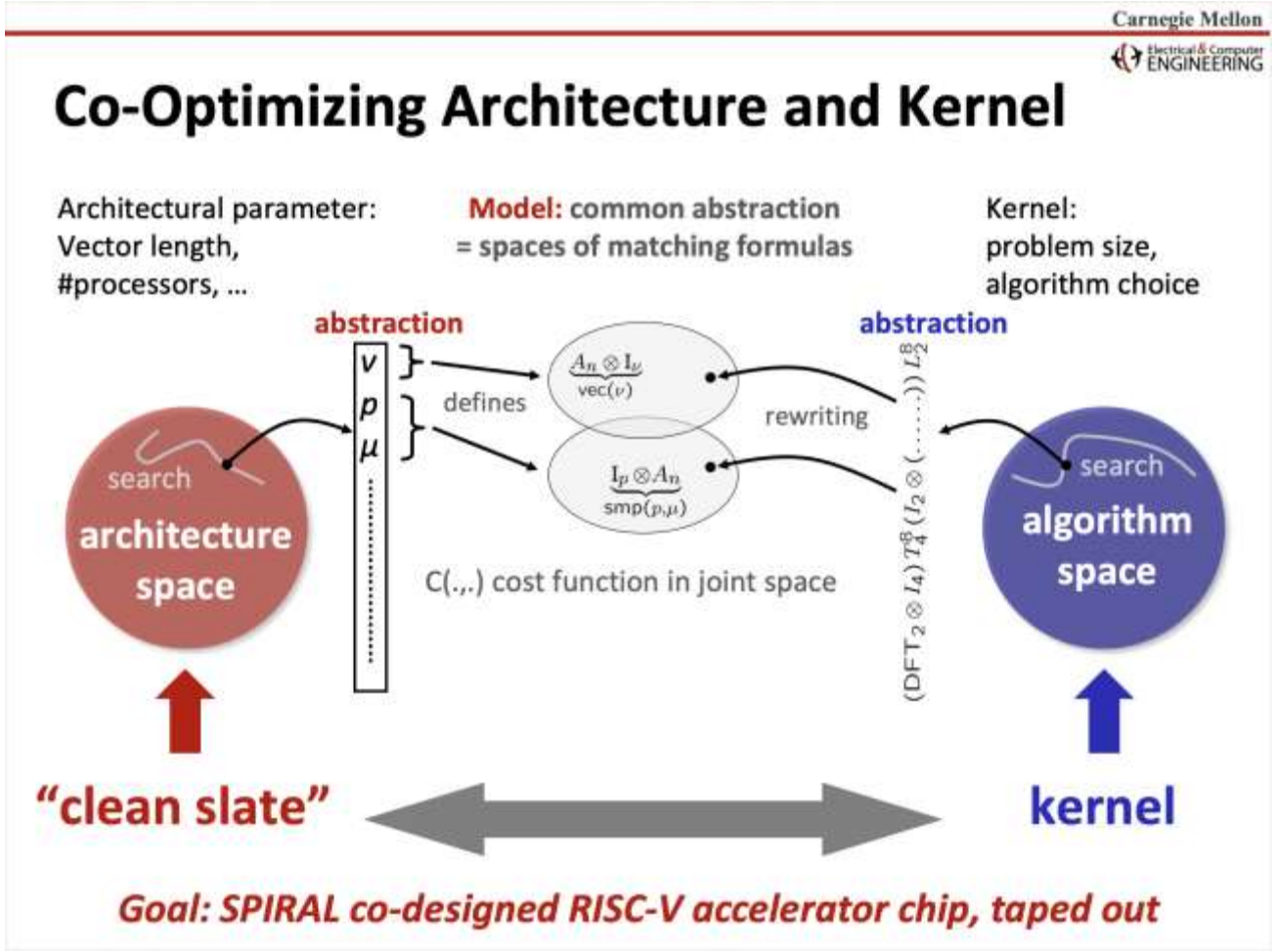


Spiral AI/ML: Optimizing data-intensive applications for changing hardware platforms

Co-Optimization for High-Performance, Data-Intensive Computing

- Hardware-software co-optimization for data-intensive computations and applications will allow DoD to take advantage of new and novel hardware architectures very rapidly
- Focus on performance models and their role in automated code generation

Prof. Franz Franchetti, CMU/ECE
Automatic performance tuning and program generation for emerging parallel platforms and algorithm/hardware co-synthesis. Targets multicore CPUs, clusters and high-performance systems (HPC), graphics processors (GPUs), field programmable gate arrays (FPGAs), FPGA-acceleration for CPUs, and logic-in-memory and 3DIC chip design.



CUI?

- CUI is the new umbrella term now used by the Federal Government for information that a “law, regulation, or government-wide policy” requires or permits an agency to handle using safeguarding or dissemination controls.
- CUI program was created by President Obama's Executive Order 13556 to create a streamlined method for information sharing and safeguarding
- The DoD version of the CUI Program is captured in DoD Instruction (DoDI) 5200.48, which was issued in March 2020. This DoDI became a formal contract requirement for the SEI in September 2021.
- The DoD Registry defines CTI as, *Technical information with military or space application subject to controls on the access, use, reproduction, modification, performance, display, release, disclosure, or dissemination.* ***Controlled technical information is to be marked with one of the distribution statements B through F, in accordance with Department of Defense Instruction 5230.24***