HE Hardware Acceleration Research

with a discussion of standardization of SW-HW Interfaces

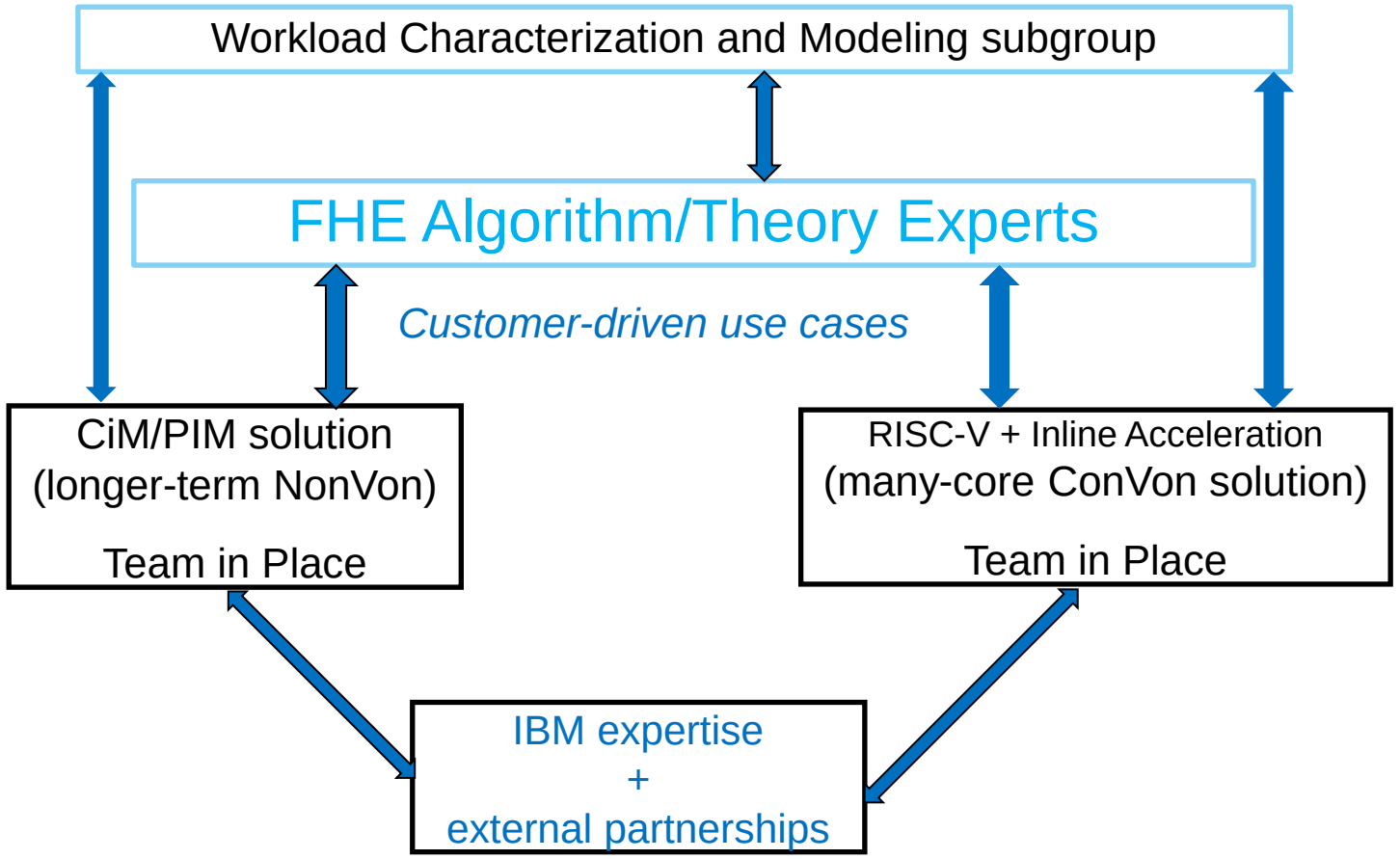
September 2, 2022

Pradip Bose (in association with)
Omri Soceanu, Nir Drucker, Karthik Swaminathan, Subhankar Pal, John Buselli et al.

IBM Research

Current Hardware Acceleration Research Strategy (IBM)

Focus on key workloads: e.g., credit card fraud detection



2022 Highlights

- SoC Performance Model in place
- Architecture and microarchitecture strawman in place
- RISC-V design and methodology (adapted over from EPOCHS)
- Leverage ESP hardware integration methodology from EPOCHS
- Resilient, low-voltage AI/ML acceleration research pursued under a linked IARPA-sponsored project
- Research collaborations with university and industrial research groups

The modeling and system-on-chip (SoC) definition methodology is linked, in part, to an ongoing DARPA project called EPOCHS

Research timeline and agenda partly inspired by DARPA DPRIVE program

Setting hardware-aided speedup targets

How are the speed-up targets chosen?
What is the baseline against which speedup is measured?

Use Case Selection and Workload Characterization

- Customer specifies real-time deadlines
- Measurement & characterization on existing systems tells us the gap

Baseline = today's state-of-the-art CPU chip/system

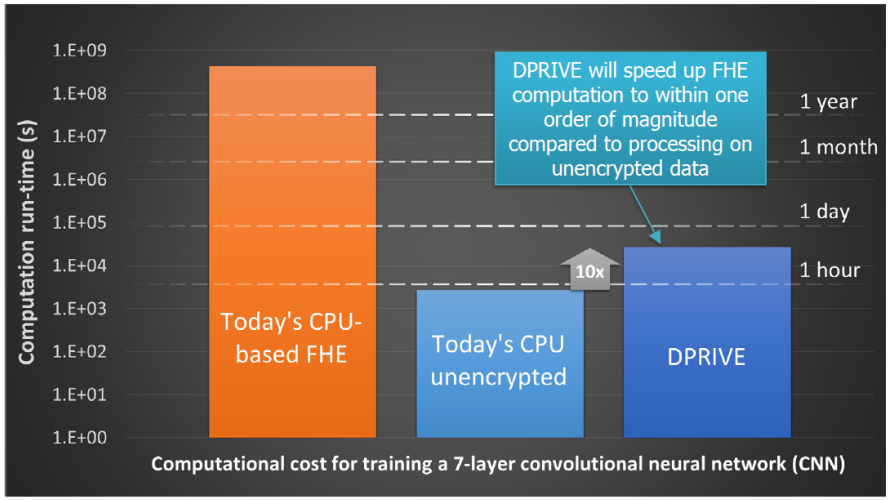
Competitive positioning

- DARPA DPRIVE program specifications
- Published claims from industry/academia
- Several sources:
  - ✓ HEAX paper from Microsoft (March '20)
  - ✓ Cheetah paper from Facebook (Feb. '21)
  - ✓ F1 paper from SRI International (Oct. '21)
  - ✓ BTS and Craterlake papers (June '22)

University partners involved in many cases

DARPA DPRIVE GRAND CHALLENGE

What are we trying to do?



Program Objective: Design and implement a hardware accelerator to reduce computational run time for FHE to make it comparable to similar unencrypted data operations (within 10x)

DPRIVE Program Metrics

Challenge		Phase 1: Building Blocks in emulation	Phase 2: Full Design in emulation	Phase 3: Prototype and Software Port
1	Ability to build and emulate all blocks (add, sub, mul, mod, shifts, transforms)	Binary yes/no		
1	Verification coverage of logic circuits	≥ 90%	100%	
1	Time to perform logic circuit verification	≤ 1 day	≤ 1 day	
2	Chip dimensions	≤ 150 mm^2 (estimate)	≤ 150 mm^2 (RTL)	≤ 150 mm^2 (real chip)
3	FHE parameter range: Plaintext Modulus	2 – 1024	2 – 1024	2 – 1024
3	FHE parameter range: Ciphertext Modulus	$2^{15} – 2^{500}$	$2^{15} – 2^{500}$	$2^{15} – 2^{500}$
3	FHE parameter range: RingSize	512 – 16384	512 – 16384	512 – 16384
Overall	Execution of a 1024-point logistic regression model	≤ 10 ms (100x)	≤ 1 ms (10x)	≤ 0.1 ms (1x)
	Execution of 7-layer CNN **inference** w/ CIFAR-10 data set per image		≤ 250 ms (100x)	≤ 25 ms (10x)
	Execution of 7-layer CNN **training** w/ CIFAR-10 data set over 10 epochs			≤ 10 hours (10x)

Metrics in parenthesis indicate expected penalty vs. plaintext operations

From Proposers' Day Briefing Charts (Tom Rondeau, 03/02/2020)

Initial Application/Use Case Focus Areas

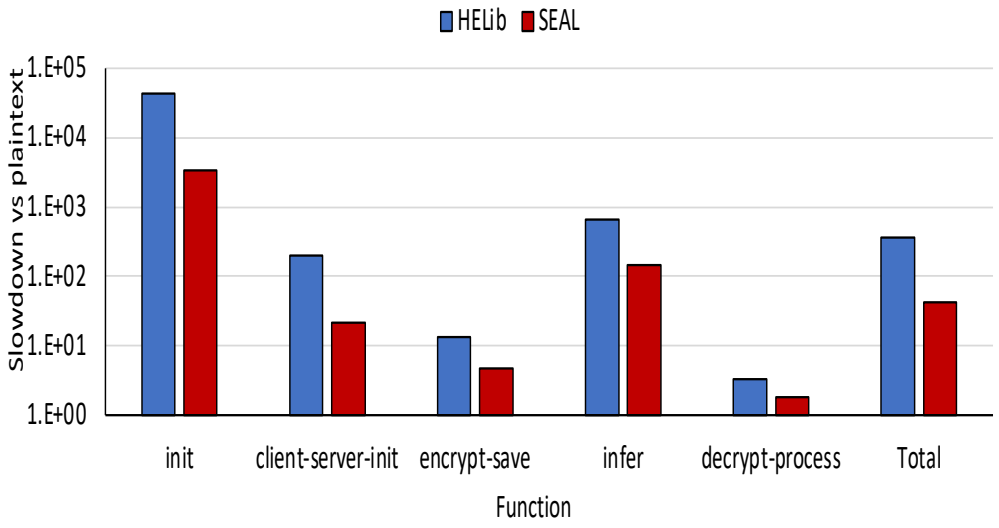
- Initially: AI-embedded financial domain transactions
- Bank/financial customers
- End-to-end use cases:
  - ✓ Credit card fraud detection – XGBOOST, 7-layer fully connected NN, inference & training
  - ✓ Loan application risk assessment (inference, training)
  - ✓ Some other applications
- Microbenchmarks derived from above.
- Microbenchmarks (DARPA DPRIVE)
  - ✓ 7-layer CNN inference /CiFAR-10 data set per image
  - ✓ 1024-point logistic regression model

Customer confidential performance specifications in many cases:
Throughput and real-time latency requirements

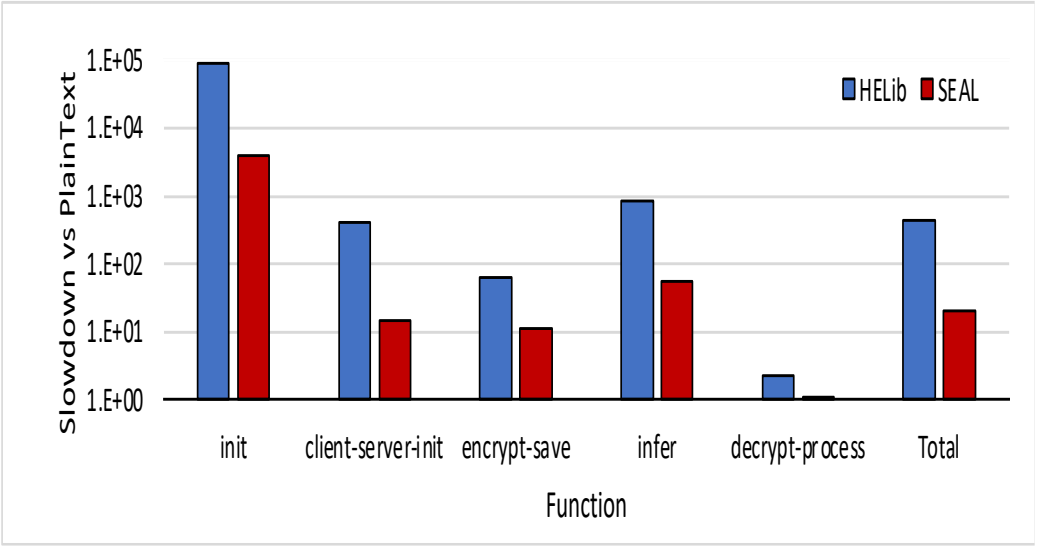
Workload Characterization – Exemplary Data

Credit Card Fraud Detection Example

Implementation	Exec Time (s) (ref m/c A)	Exec Time (s) (ref m/c B)
Plaintext	0.83	0.69
HELib	301.28	193.78
SEAL	34.85	10.09



- Small (toy) problem size used
- Slowdown from plaintext to ciphertext computation: up to 350X in this particular case.
- Scaling up problem size (NN size) exposes more challenging speed-up needs: e.g. ~1000X for 7-layer CNNs (as in DPRIVE program



BTS: An Accelerator for Bootstrappable Fully Homomorphic Encryption

Sangpyo Kim† Jongmin Kim† Michael Jaemin Kim† Wonkyung Jung†
Minsoo Rhu‡ John Kim‡ Jung Ho Ahn†
Seoul National University†, KAIST‡
{vnb987, jongmin.kim, michael604, jungwk, gajh}@snu.ac.kr, {mrhu, jjk12}@kaist.edu

ISCA 2022

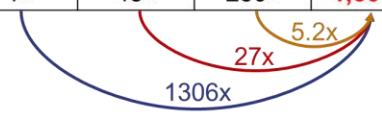
Experimental setup

- Simulation
  - BTS is sized **373.6mm²** and consumes **163.2W** of peak power using 7nm technology node
- Performance metric
  - amortized mult time per slot ($T_{mult,a/slot}$)
- Workloads
  - **[HELR]** Logistic regression training of a binary classification model
  - **[ResNet-20]** ResNet-20 inference for CIFAR-10
  - **[Sorting]** Sorting using 2-way sorting network to sort 2^{14} data
- State-of-the-art implementations of CKKS for comparison
  - **Lattigo** (CPU)
  - **100x** (GPU)
  - **F1** (ASIC)
  - **F1+** optimistically scaled version of F1 from 14nm to 7nm technology node

Results

- **[HELR]** Logistic regression training – training time per iteration (1024 samples)

	Lattigo (CPU)	100x (GPU)	F1+ (ASIC)	BTS
Execution time (ms)	37,050	775	148	**28.4**
Speedup	1×	48×	250×	**1,306×**



- **[ResNet-20]** BTS shows up to **5,556×** better performance than prior work in ResNet-20 inference
- **[Sorting]** BTS outperforms prior work by **1,482x** in sorting

System Architectural Directions Summary
(IBM Research)

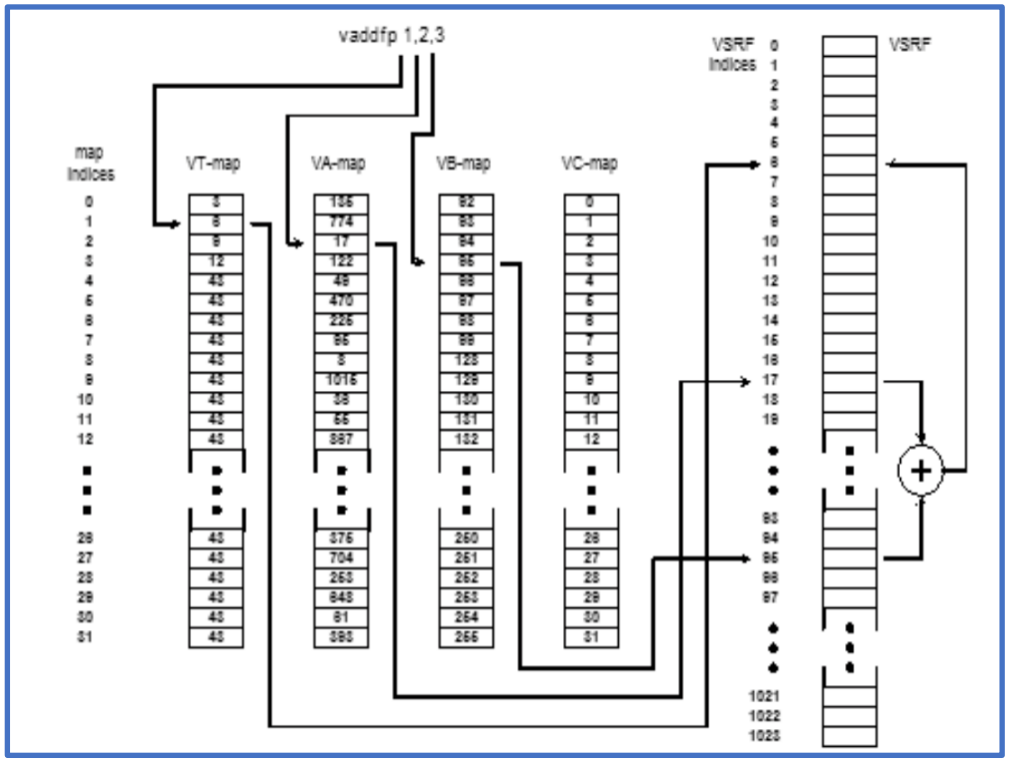
- **Paradigm#1:** focused on conventional von Neumann architecture (ConVon)
  - RISC-V/inline accel based massively parallel (dynamic SIMD) architecture

- **Paradigm#2:** focused on non-von Neumann architectures (NonVon)
  - CiM/PIM and other forms of NDP (near-data processing)
  - Collaborative partnership with university and industry research groups
  - Leverage relevant Heterogeneous Integration (HI) technology within IBM Research
  - And, ongoing collaborative projects with memory vendor groups

ConVon Architectural Paradigm:

Key Features and Performance

- Massive (64KB) register file for each compute core: indirectly addressed
- Organized such that it allows interleaved (“SPLIT”) access with register-embedded compute engines
- Matched innovation of split power supply: for lowering the power consumption of inline accelerator logic & data comm.



Very early power-performance estimates of our ConVon research

AI-compute, 2-byte: 1.3 PFLOPS @ 0.4 V, 0.66 GHz, ~400 W

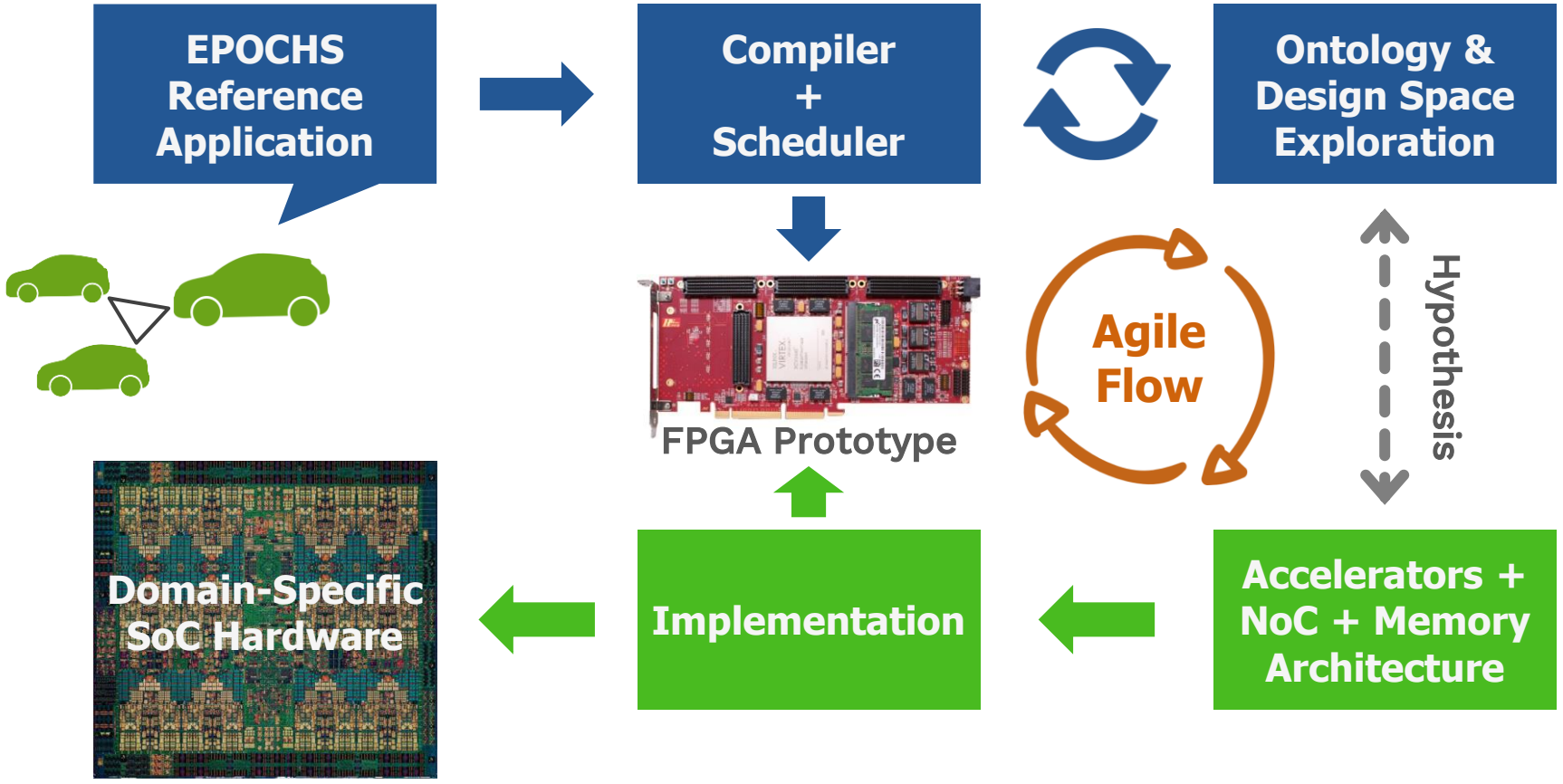
HPC/Linpack/DP: 192 TFLOPS @ 0.4V, 0.66 GHz, ~400 W

64K NTT (key to FHE perf): >3 orders of magnitude better than SOTA CPU

Efficient Programmability Of Cognitive Heterogeneous Systems (EPOCHS)

(We will use this agile design methodology)

EPOCHS Agile Flow Methodology

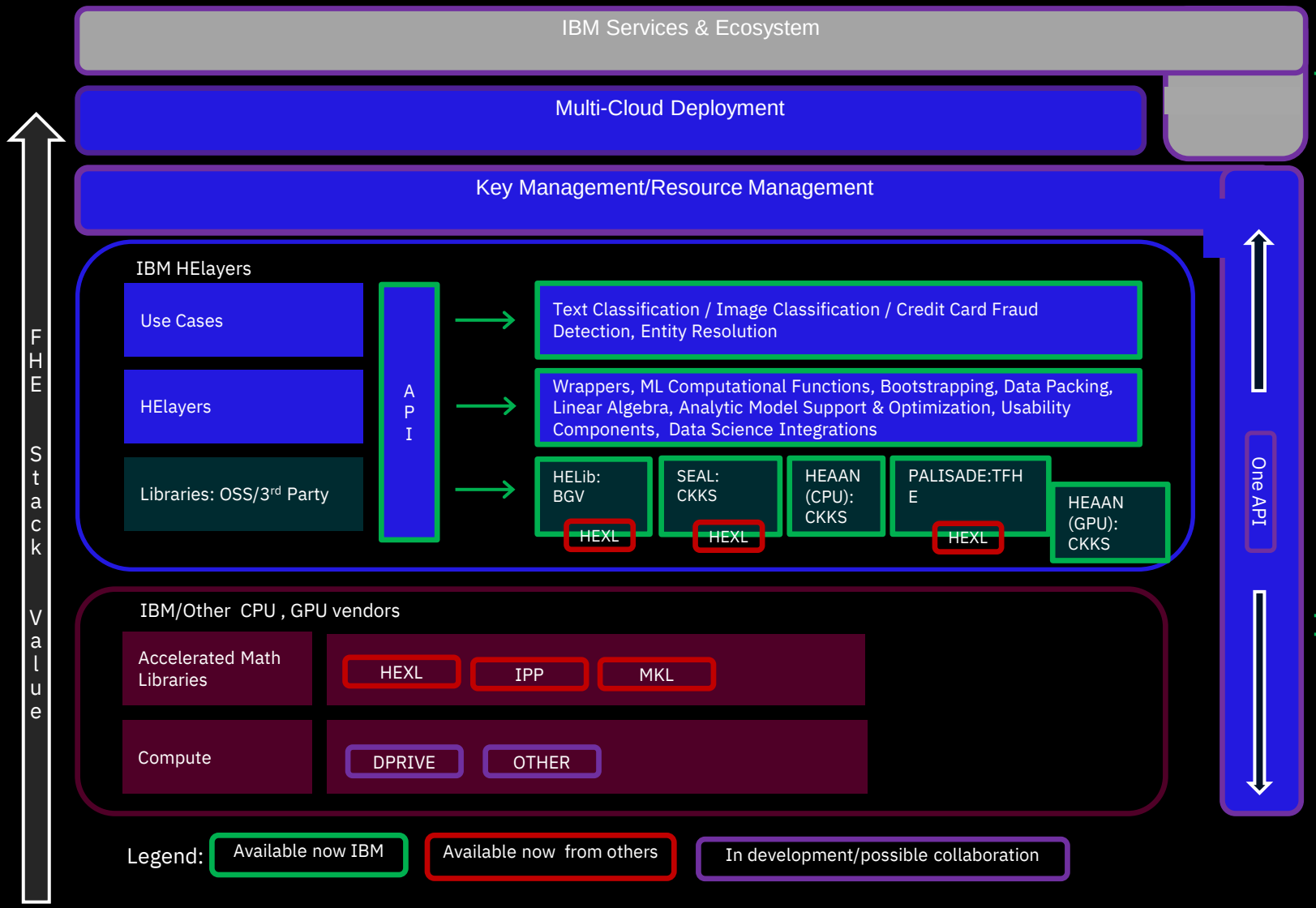


10×-100× reduction in person-years

Agile methodology to quickly design and implement an easily programmed domain-specific SoC for real-time cognitive decision engines in connected vehicles

All open-source hardware and software artifacts in this DARPA DSSoC sponsored, IBM-led research

FHE Open Collaboration Ideas: Full Stack Integration



Potential Collaboration Areas

HE Application Layer:

- Key Management/Resource Management
- HELayers– Packaged with additional optimizations for Intel HW
- FHE Standards and Cryptanalysis
- Open-source HE Benchmark Community initiative (HEBench)
- Client use scenarios and workloads
- Joint Thought Leadership
  - Research & Client publications
  - Academic partnerships
  - Clients/Conferences/Marketing events
- Open-source HE Acceleration library (HEXL) from another vendor

HW Acceleration:

- Explore joint research opportunities (IBM Research with others)
  - ✓ Hard system architectural issues: esp. memory sub-system, on-chip communication sub-system, verification, RAS, power mgmt
  - ✓ Joint research consortium – with external sponsored funding

Summary and Open Discussion

- IBM Research has an active interest in HE hardware-software co-designed acceleration
- We are pursuing a relative open collaborative research model
- Standardized hardware-software interfaces are a key driver of such a research objective
- We are open to discussion about collaborations and opne-source development methodologies and practices