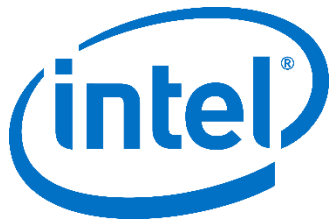


Software-driven Design for Domain-specific Compute

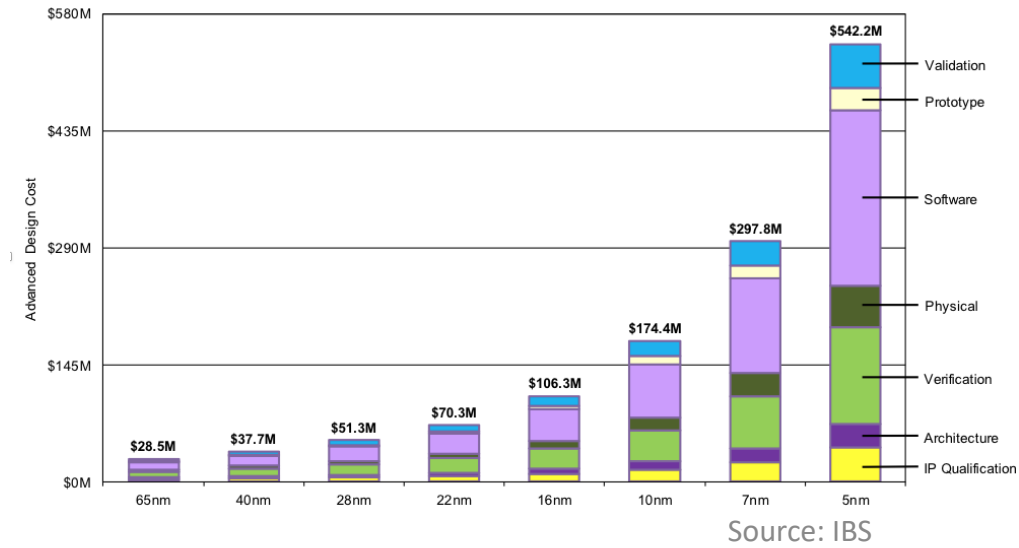
Desmond A. Kirkpatrick (Intel Corporation)



Domain Specific Design Dilemma: NRE Costs

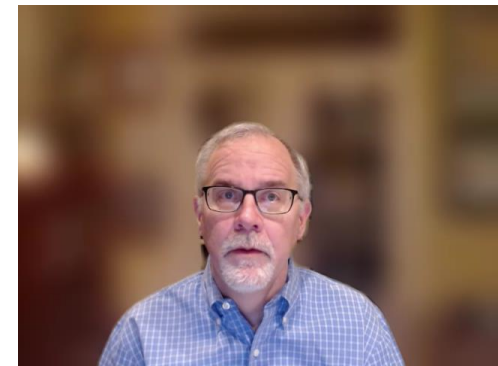
Specialization:

- Massive increase in compute/efficiency
- Lower volumes
- Increased NRE costs (esp. software)



Assumptions about Accelerators:

- Single-purpose products are exceedingly rare
- Constellations as part of a system are quite common
- Always part of a larger computation, part of a larger system
- Separate, highly-specialized, software stacks
- Workloads evolve rapidly, requiring rapid TTM



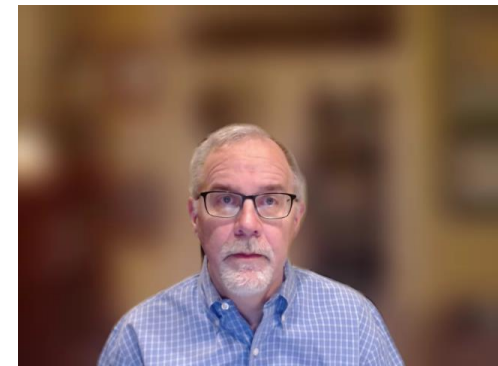
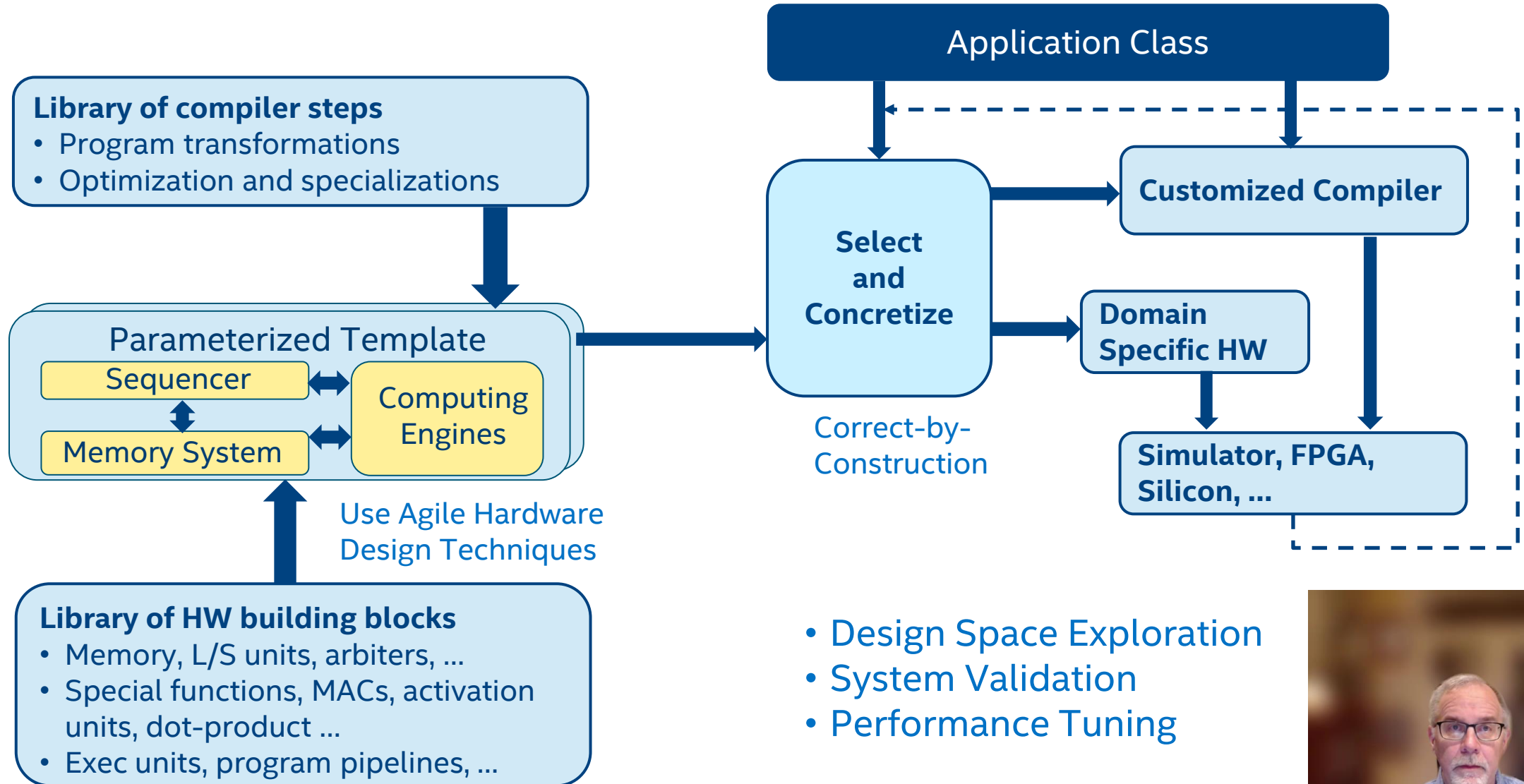
Domain-specific Design NRE Countermeasures

- Accelerator family targeting different markets
- Accelerator complexes / aggregate application support
- Frameworks/building blocks for domain accelerators
- Modular / flexible compiler frameworks
- Extendable HW APIs for future accelerator interfacing
- Trade efficiency for programmability for broader support
- Efficient codesign methodologies

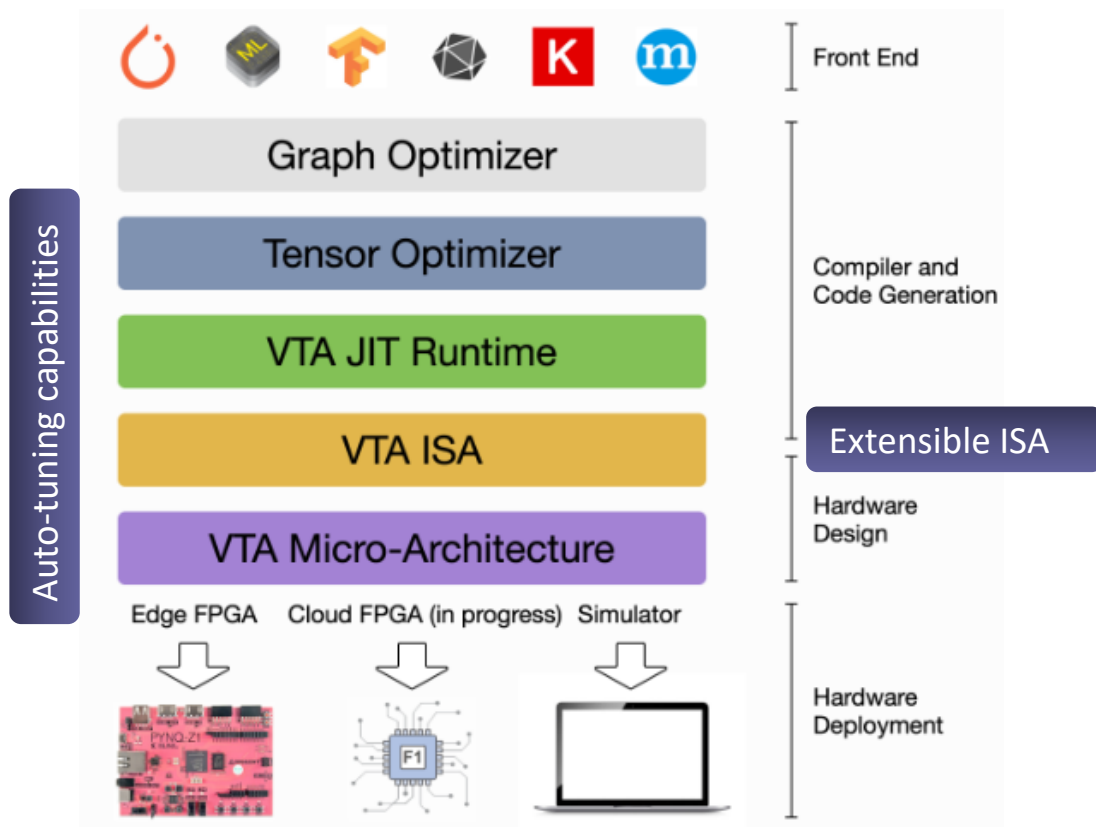
*Need an automation framework comprehending SW and HW changes
during expansive Design Space Exploration*



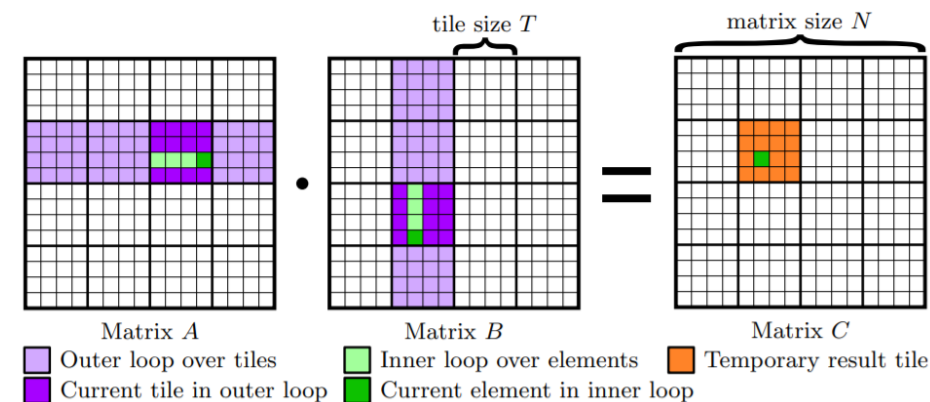
Template-Based Domain Specific Design Flow



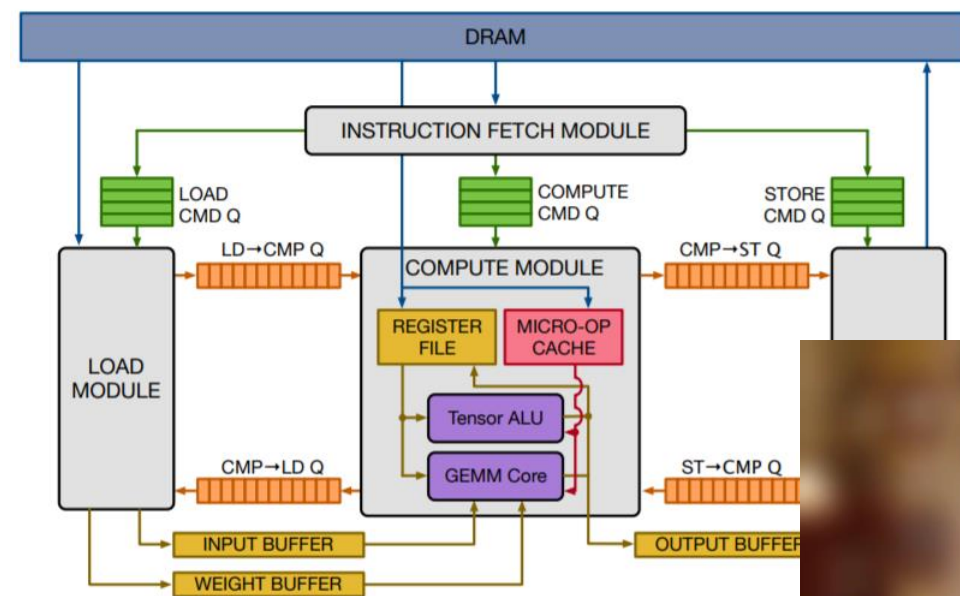
TVM $\rightarrow$ Versatile Tensor Architecture (VTA)



FPGA / Cloud FPGA and Silicon targets

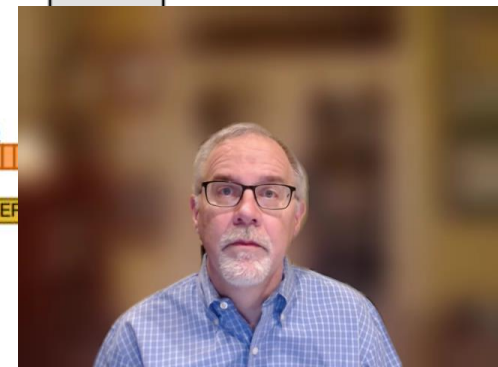


Tensor compiler with configurable passes



Parameterized Chisel HW

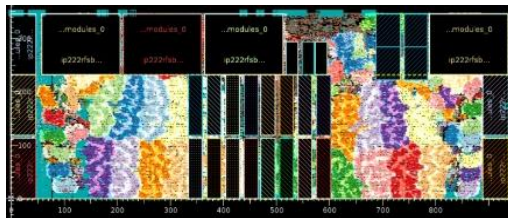
<http://tvm.apache.org/docs>



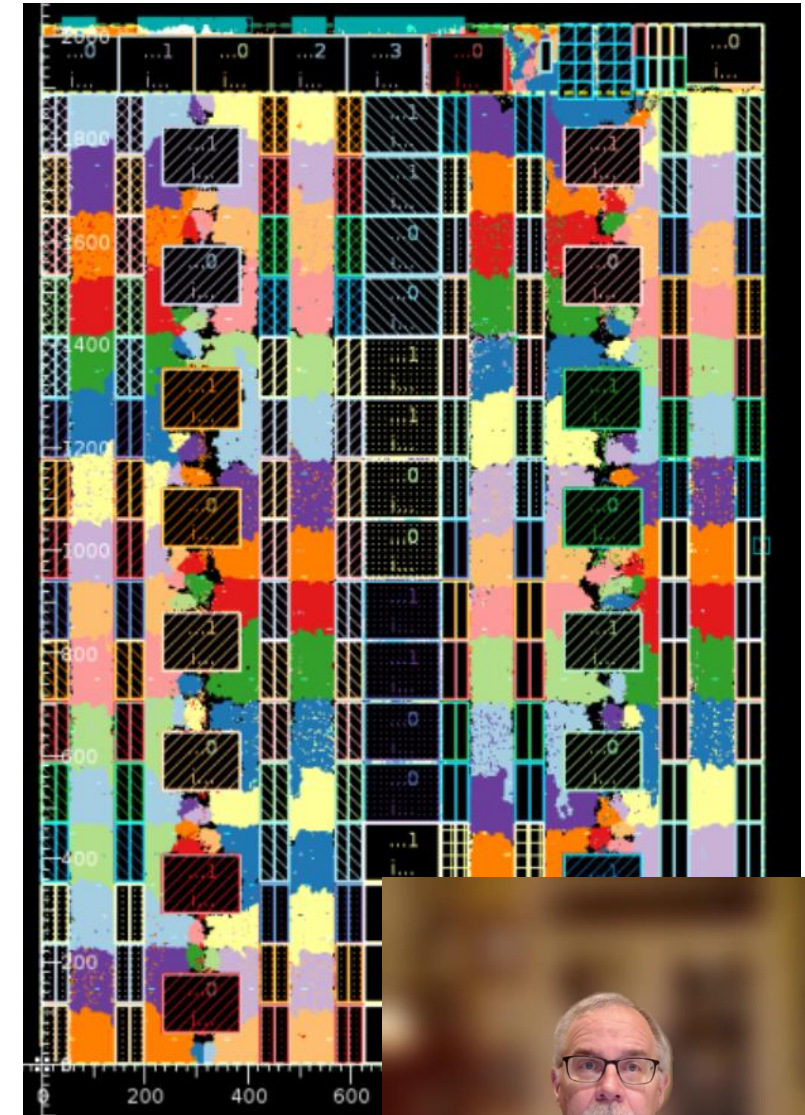
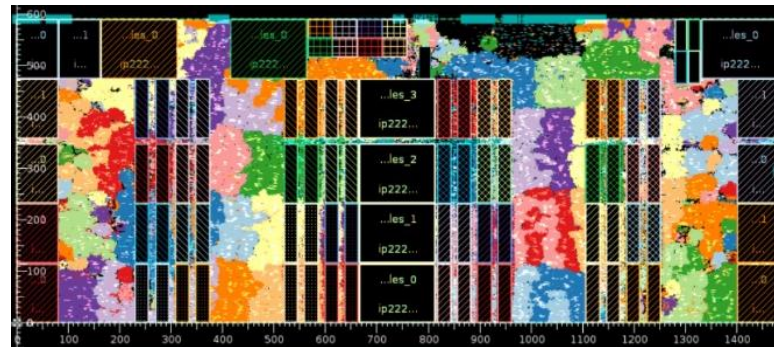
Three Pareto-optimal Implementations running ResNet and MobileNet

			Resnet 18 Computation	
MACS (#)	Area	Power	Latency	Energy
256	1X	1X	1X	1X
1024	3.04X	4.33X	0.25X	1.08X
4096	11.13X	11.17X	0.1X	1.11X

Agile Backend: Big-VTA converged in 8h using configurable wire-pipelining for largest configuration



Little-VTA



Big-VTA:
Configurable Wire-Pipelining



EDA for Domain-specific Compute

- SW/HW Codesign
 - Agile TFM driving combined SW and HW product development
- Flexible Compiler technologies
 - Halide language
 - TVM framework
 - MLIR Infrastructure
 - EXO language
- Design Space Exploration support
 - Parameterized design (both SW and HW)
 - Rapid backend feedback (ML predictors?)
 - Search space support (ML?)
- Domain-specific building block libraries
 - BasejumpSTL: but ...focus on domains: image, NN, DSP, media, networking, codecs, ...

