The Future of Heterogeneous Systems Design

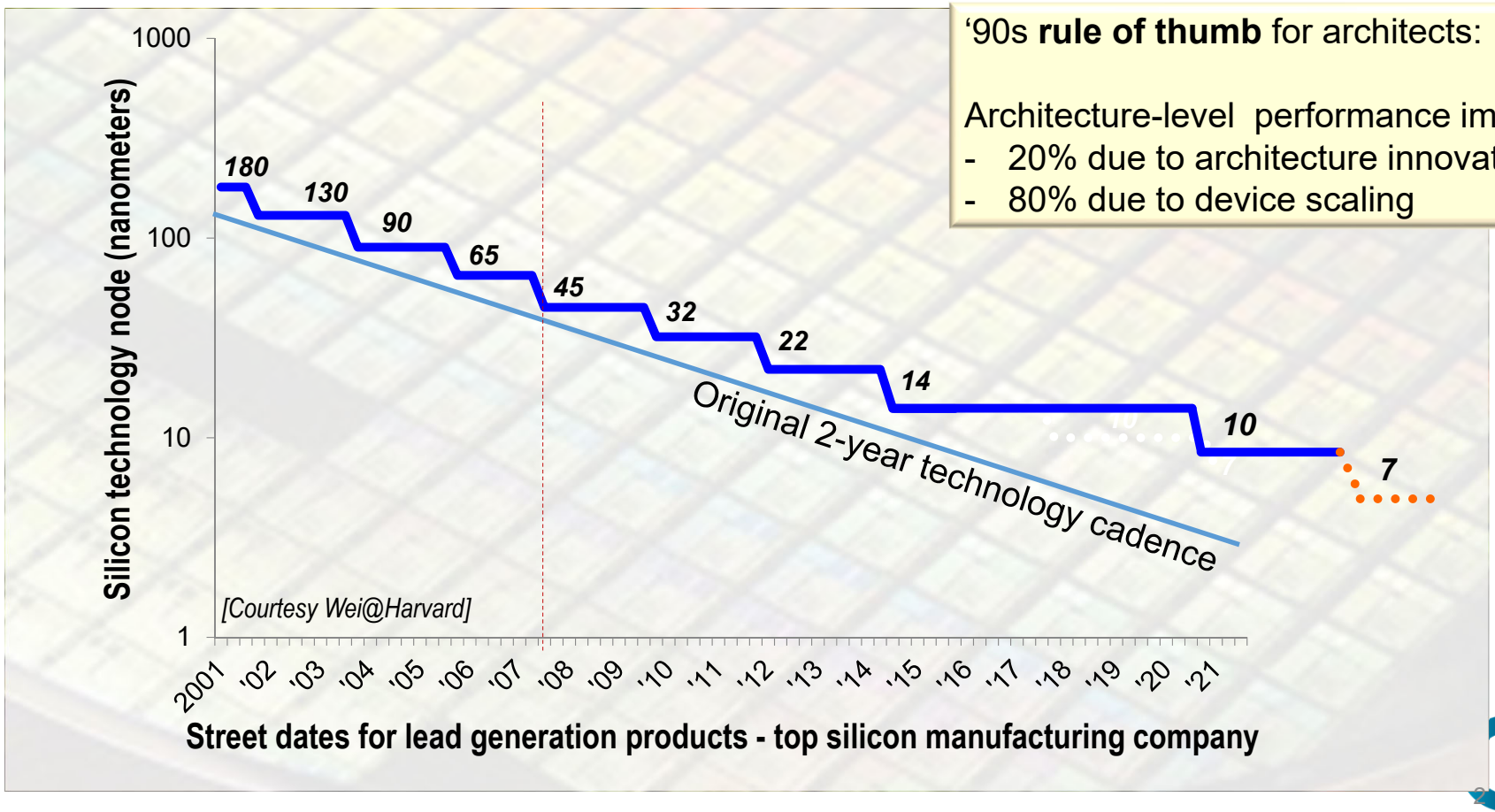
Valeria Bertacco
ADA Center Director

adacenter.org @ADA_Center

This work is supported by the Semiconductor Research Corporation (SRC) and DARPA

Applications Driving Architectures

Technology innovation – limited benefits from device scaling



'90s **rule of thumb** for architects:

Architecture-level performance improvement:
- 20% due to architecture innovation
- 80% due to device scaling

Applications Driving Architectures

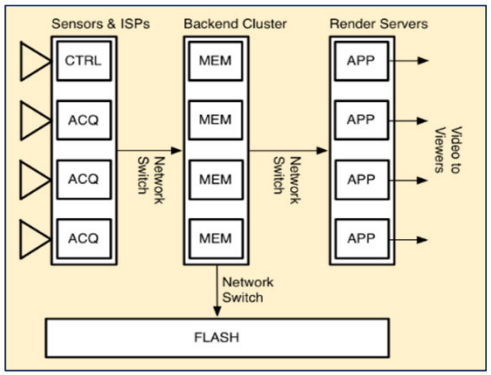
Emergence of specialized architectures

+Growing domain offerings

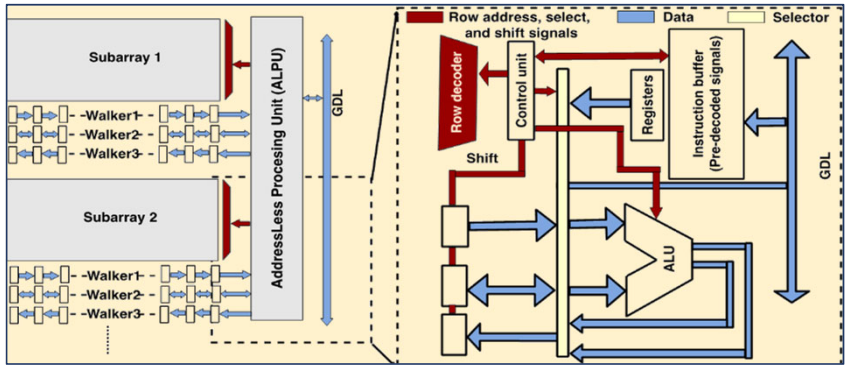
+Great performance/energy boosts

- 1 app → 1 accelerator

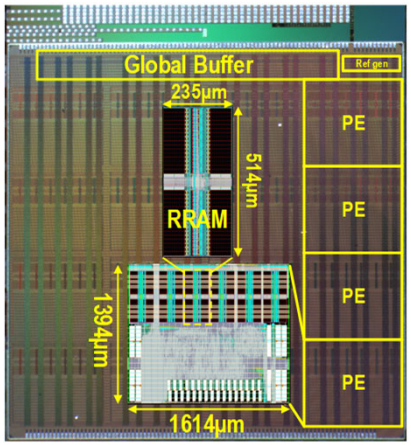
- Ad-hoc interfaces



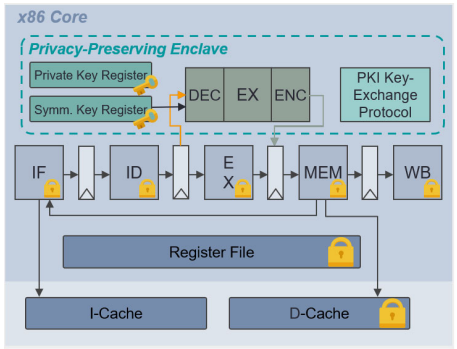
2016 - Lee et al. image processing accelerator



2020 – Skadron et al. – Fulcrum: bit-level parallel PIM accelerator



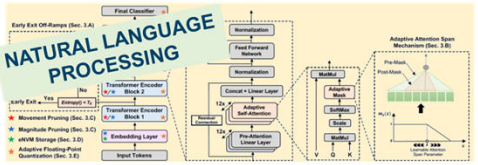
2019 - Sylvester et al. 22nm low-power DNN



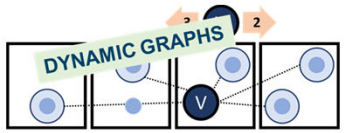
2021 - Austin et al. sequestered encryption accelerator

ada Applications Driving Architectures

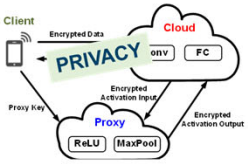
Accelerators from the ADA Center



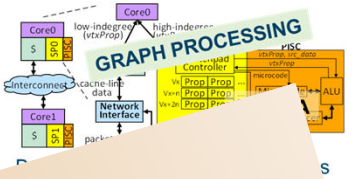
Brooks, Wei – EdgeBERT transformer-based NLP

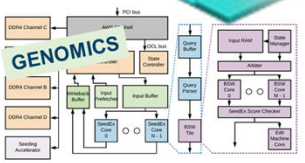


Bertacco – Dynamic graph acceleration

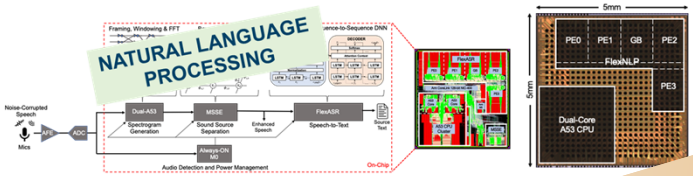


Brooks, Wei - IMPALA distributed ML privacy

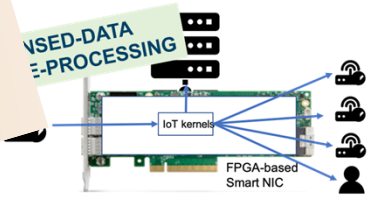




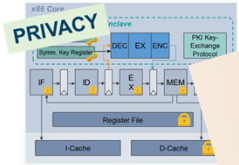
Das – SeedEx – optimal seed extension in genomic algorithms



Brooks, Wei – FlexNLP - Speech-enhancing ASR

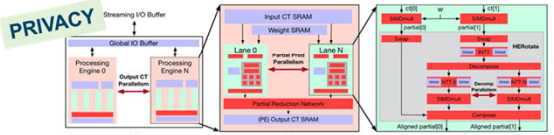


Gavriloska – Vasado: smart NIC acceleration of IoT analytics

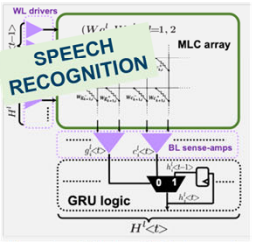


Austin – accelerated always-encrypted computation

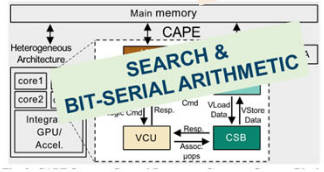
Wenisch – 3D Image reconstruction



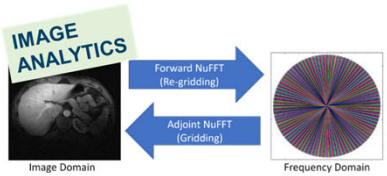
Reagen – CHEETAH – HE accelerator for NN



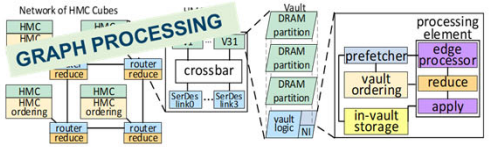
Brooks, Wei – PIM-based trigger-word detector



Batten – CAPE for bit-serial arith and search



Wenisch – 3D image analytics with Jigsaw



Bertacco – MessageFusion: specialized NoC for graph analytics

LARGE SEGMENTATION LEADING TO:
UNMANAGEABLE INTERFACE DIVERSITY,
INTEGRATION COMPLEXITY, PROGRAMMABILITY

Applications Driving Architectures

Applications Driving Architectures (ADA) Research Center

5-year endeavor: 2018-2022

21 faculty members, 130 graduate students

Co-sponsored by SRC JUMP and DARPA

GOAL: reignite computing system innovation for the 2030-2040 decade through:

- sustained scalability and

- sustained value creation

ada Applications Driving Architectures

Managing design under vast heterogeneity

1. ***[ENABLE MORE IDEAS TO TRANSFORM INTO NEW DESIGNS]***
   Lower expertise required to design hardware systems
   → ***reignite innovation***

2. ***[BOOST OPTIMIZATION OPPORTUNITIES]***
   Blur hardware abstraction layers and cross-optimize

3. ***[IMPROVE SILICON USE EFFICIENCY]***
   Need flexible fabric for specialized accelerator synthesis
   → ***lower carbon emissions associated with computing***

Solution 1: Lower expertise needed to design hardware systems

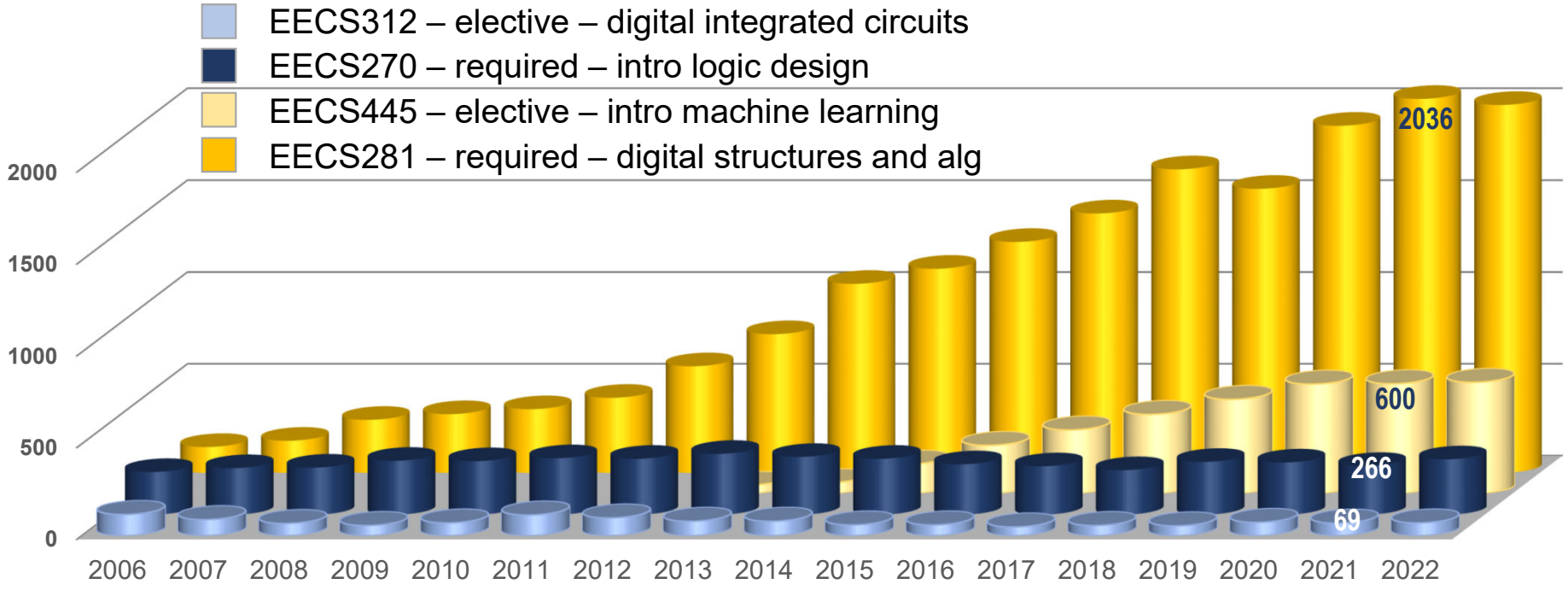
- Domain-specific languages (DSL) boost programmer's productivity
- DSLs are approachable by a broad population of software engineers
- High-level compilers today are unable to leverage specialized accelerator hardware (APIs are the practice)

How does ADA approach this goal?

- Enabling compilation flows from DSLs to accelerator-rich heterogeneous architectures (3LA)

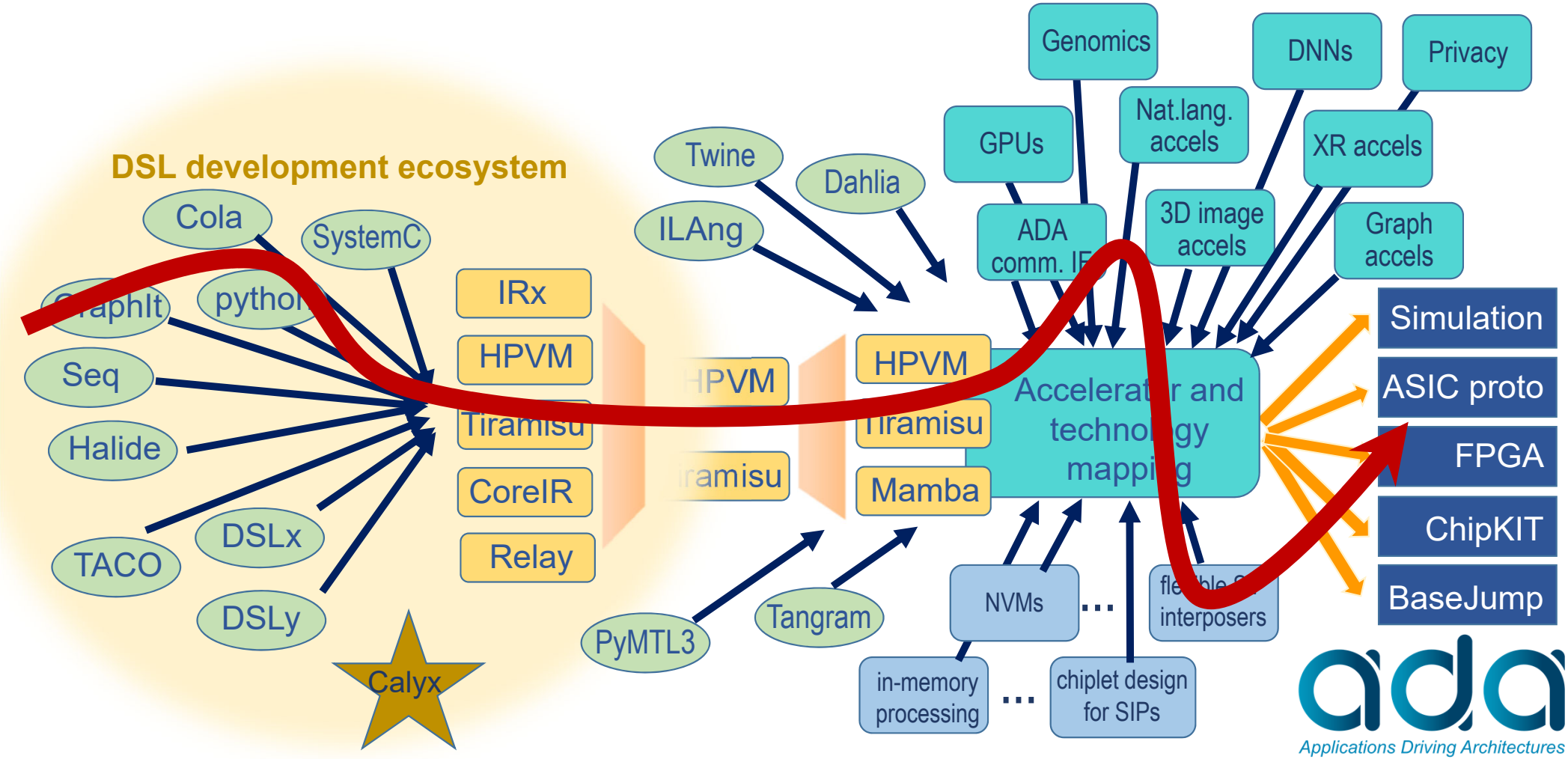
Innovation is powered by people

University of Michigan – selected EECS course enrollments



Applications Driving Architectures

ADA today – design flows for the 2030s



Mapping DSLs to accelerators: 3LA

SOFTWARE PRIMITIVES

Applications provided in different DSLs

Tensorflow

Pytorch

Keras

ONNX

Relay IRModule (computation graph)

tvm

Relay-to-HW program fragments mapping


```
FlexNLP func A:
instr. 1 (STR r2, 0xffff0000)
instr. 2 (LDR r3, 0xffff0010)

FlexNLP func B:
instr. 3 (STR r2, 0xffffaa00)
instr. 4 (LDR r3, 0xffffaabb)

NVDLA func C:
instr. 5 (STR r4, 0xffff0100)
instr. 6 (LDR r5, 0xffff0110)
```

HARDWARE FUNCTIONS

Executable/runtime leveraging HWs

```
; CPU instructions
CMP   r0, r1
SUBGT r0, r0, r1
BNE   loop
; Access accel 1 (MMIO)
STR   r2, 0xffff0000
LDR   r3, 0xffff0010
; Access accel 2 (MMIO)
STR   r4, 0xffff0100
LDR   r5, 0xffff0110
; CPU instructions
MOV   r3, r2
SUBGT r0, r0, r1
B     lr
```

Heterogeneous hardware backends

arm

NVIDIA CUDA

General purpose

FlexNLP

HLSCNN

NVDLA

VTA

Application specific

End-to-end compilation steps:

1. Take Relay as the representation
2. Define Relay & HW ILA formal models
3. Provide Relay-to-HW program fragment mappings for each accelerator/functionality
4. Verify the correctness of the program fragment pairs via ILA-based methodology
5. Pattern matching and code-gen.

ILAng

Translate ILA program fragments into:

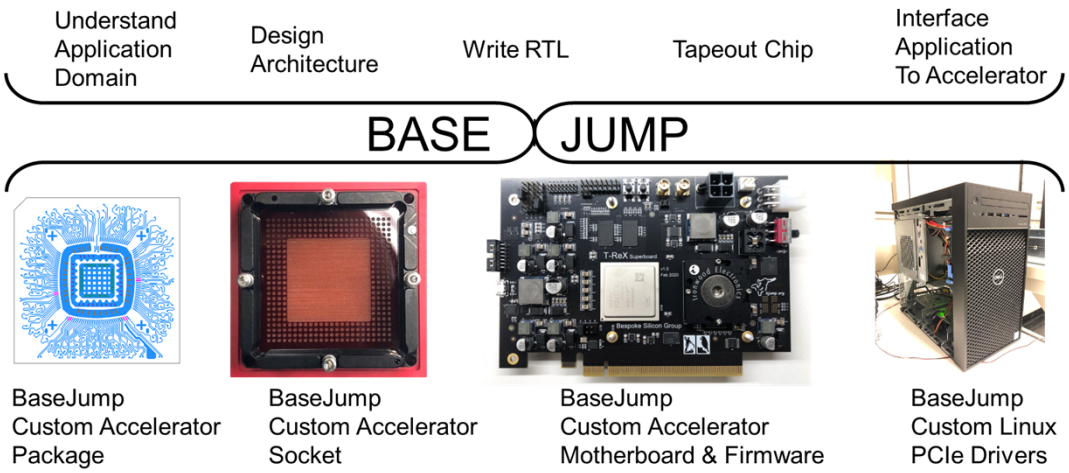
- SystemC modules for simulation validation
- SMT formulas for formal verification

[Malik, Tatlock, Wei]

ada

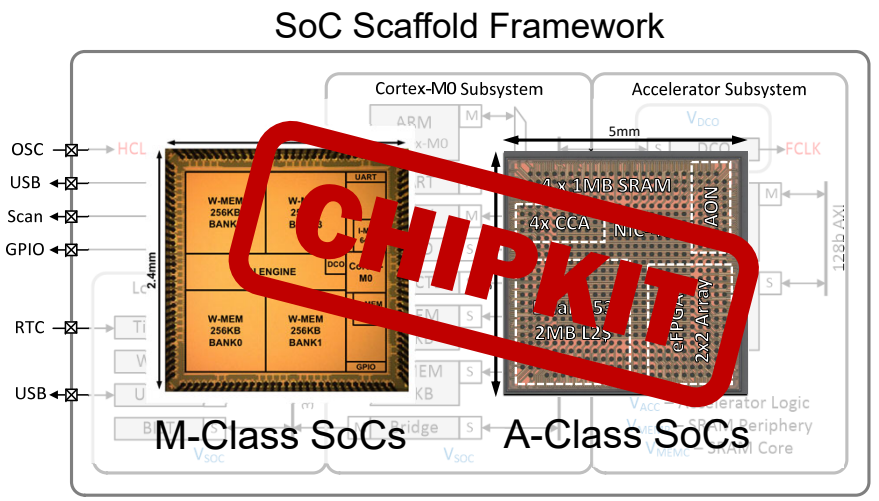
Applications Driving Architectures

Enabling agile research: Test chip frameworks



Allow new accelerator chips to be rapidly deployed into a platform without having to design custom systems that support them -- includes

- Logical and physical socket
- I/O links and networks
- Accelerator motherboard



SoC scaffold library and example to simplify SoC design integration

- AXI + protocol checker
- Fully-synthesize-able, all-digital DDR1 PHY and memory controller
- Programmable DMA controller
- SoC examples to highlight HLS flow, e.g., FlexASR

[Brooks, Taylor, Wei]

Applications Driving Architectures

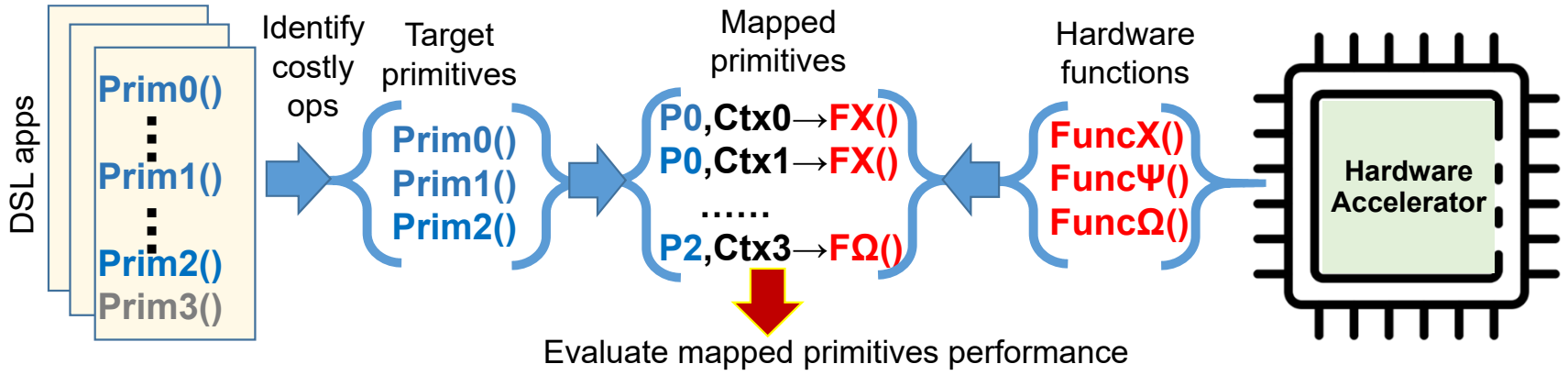
Solution 2: blur hardware abstractions layers, cross-optimize

Why: it provides many additional optimization opportunities, which have traditionally been overlooked

How does ADA approach this?

1. Explore cross-optimization while compiling in end-to-end design flows (PriMax)
2. Design exploration tools that allow computer architects to explore device parameters (NVMexplorer)

PRIMAX: selective primitive mapping



- Mapping DSL primitives → accelerator functions leads to mixed performance results
- PRIMAX identifies when the mapping is beneficial and applies it selectively

Breadth-First Search

```
func updateEdge(src : Vertex, dst : Vertex) -> output : bool
  output = CAS(&parent[dst], -1, src);
end

func toFilter(v : Vertex) -> output : bool
  output = parent[v] == -1;
end

func main()
  % Declare an active set and make Vert 0 the starting point
  var active : vertexset{Vertex} = new vertexset{Vertex}(0);
  active.addVertex(0);
  parent[0] = 0;

  % Loop until the active set is empty
  while (active.getVertexSetSize() != 0)
    #s1# active = edges.from(active)
                  .applyModified(updateEdge, parent);
  end
end

schedule:
  program->configApplyParallelization("s1","dynamic-vertex-parallel");
  program->configApplyDirection("s1", "SparsePush");
  program->configApplyAcceleration("s1", "OMEGA");
```

```
func updateEdge(src : Vertex, dst : Vertex) -> output : bool
  output = CAS(&parent[dst], -1, src);
end
```

parent[dst] : irregular access ➔ map to SPM

CAS : atomic on SPM data ➔ map to PISC

[Bertacco]

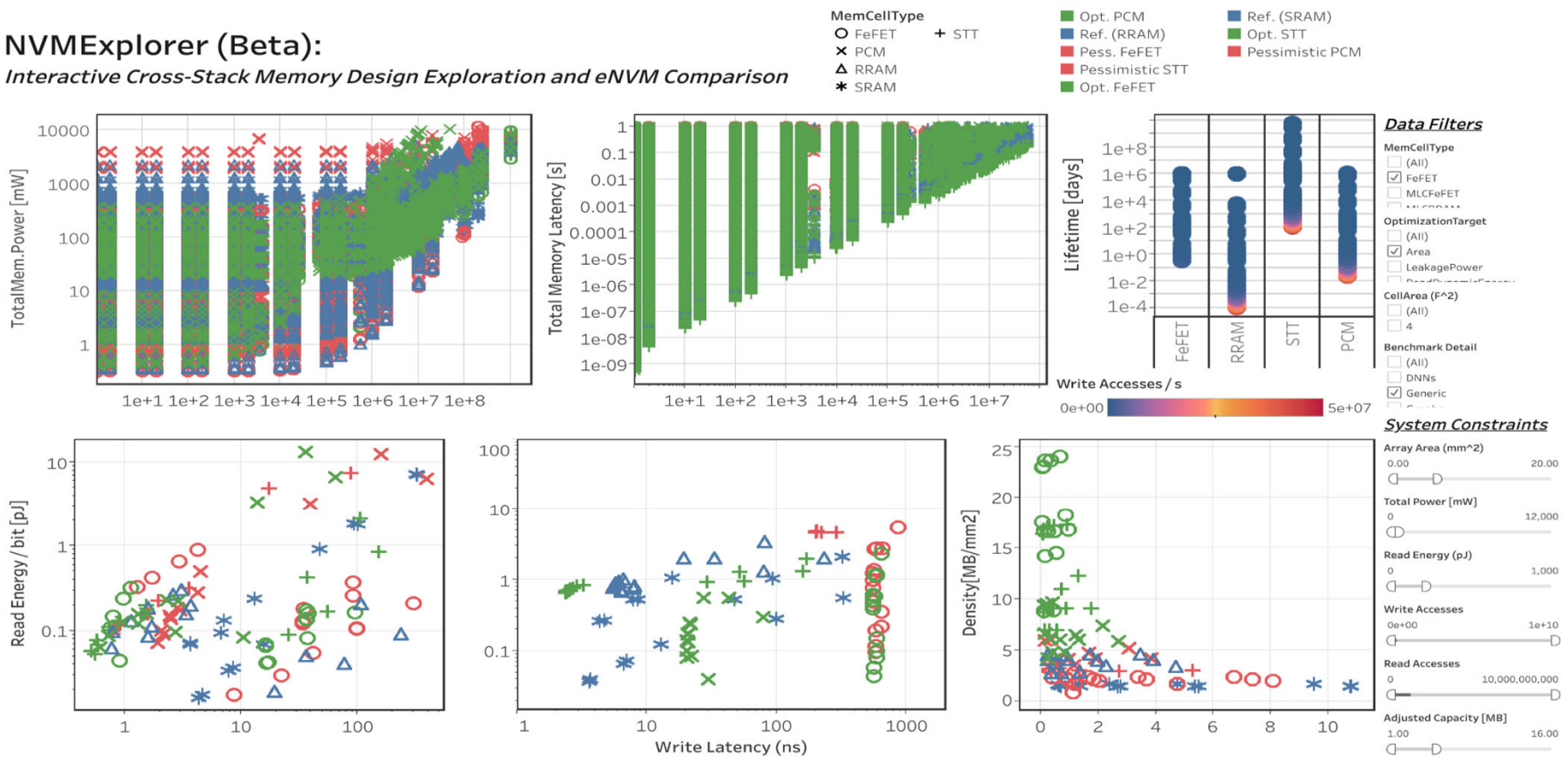
Case study:

DSL	Accel
GraphIt	→ OMEGA
	↘ GraphPull

Geomean Speedups	
PriMax	**1.57x**
Optimal: Both Targets	1.58x
Optimal: OMEGA Only	1.45x
Optimal: GraphPull Only	1.17x

Applications Driving Architectures

Design exploration tools: NVMExplorer



[Brooks, Wei]

Applications Driving Architectures

Solution 3: Flexible fabrics for accelerator synthesis

- It is impractical to produce chips with hundreds of accelerator types
- Many computing systems must be capable of running a wide range of applications

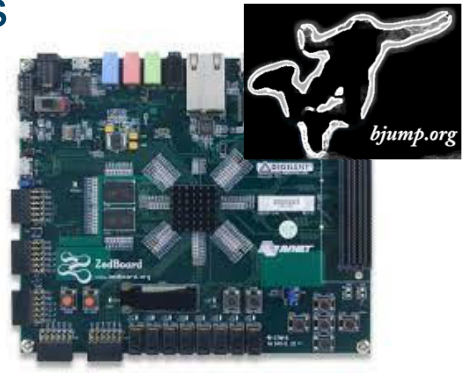
Must fit many different accelerators in a small silicon footprint

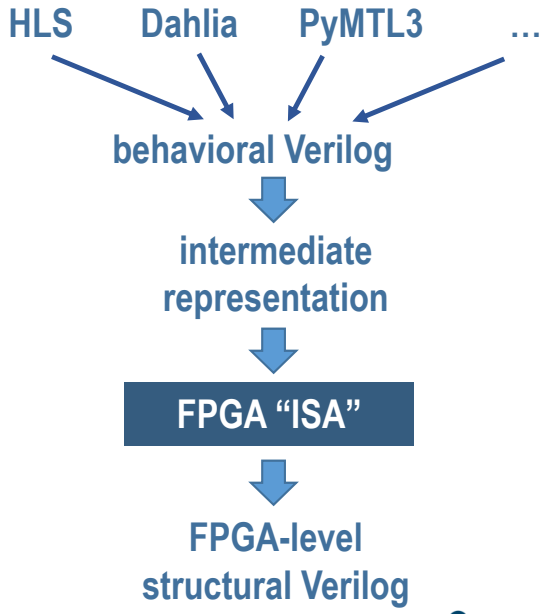
ADA: designing for reconfigurable hardware

- [Kasikci] SignalCat & LossCheck – debugging support for FPGA designs
  Monitor signals over time, identify data losses in datapaths

- [Tatlock] Lakeroad - ISA synthesis for FPGAs
  make FPGA-synthesis similar to software compilation,
  to improve compiler predictability

- [Taylor] BaseJUMP flow for FPGAs





Applications Driving Architectures

In summary:

1. *[ENABLE MORE IDEAS TO TRANSFORM INTO NEW DESIGNS]*
   Lower expertise required to design hardware systems
   → reignite innovation

2. *[BOOST OPTIMIZATION OPPORTUNITIES]*
   Blur hardware abstraction layers and cross-optimize

3. *[BETTER SILICON EFFICIENCY]*
   Need flexible fabric for specialized accelerator synthesis
   → lower carbon emissions associated with computing

Thank you

ada
Applications Driving Architectures

CRISP
Center for Research on Intelligent Storage and Processing in Memory

Intelligent Memory and Storage

Kevin Skadron

Director, CRISP Center

Dept. of Computer Science

University of Virginia

New Trajectories for Memory and Storage

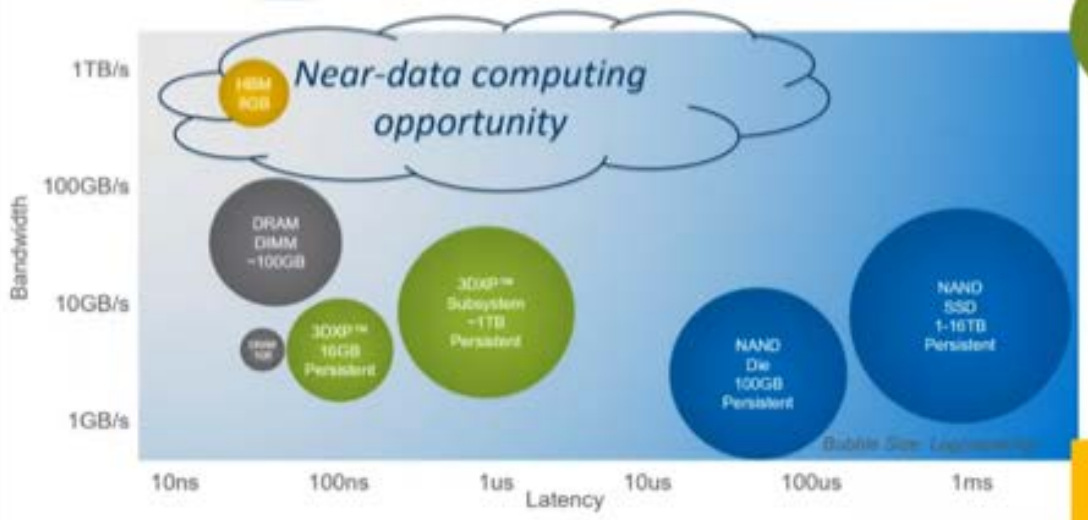
[From **Decadal Plan for Semiconductor**s presentation by Sean Eilert, Micron]

Heterogeneous Computing Brings New Memory Requirements

Bandwidth for GPU & AI accelerators in Mobile applications

Bandwidth and capacity for domain specific datacenter

Persistence & low power for edge applications



3D integration & Low temperature memory processes enable improved logic

Reduced data movement enables Bandwidth & Energy improvement

Energy-efficient near-data computing

. . . But adoption is limited by ease of programming

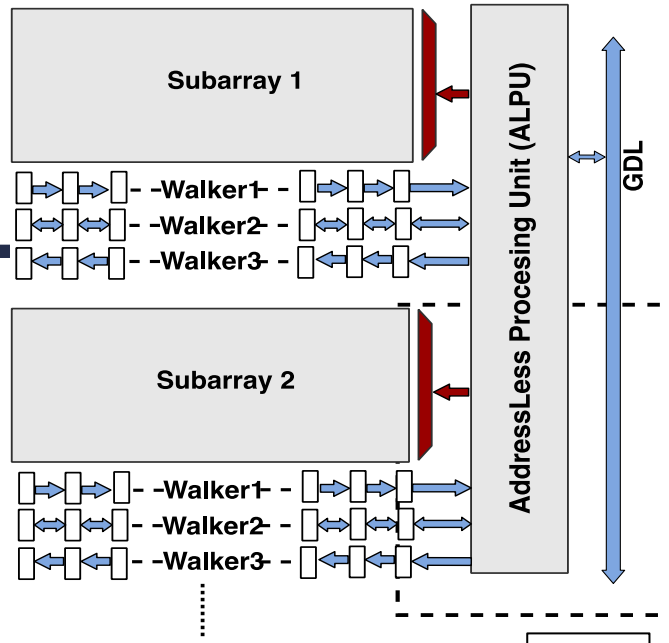
Near-data computing framework is a key enabler for compute bandwidth scaling and energy savings

Why Intelligent Memory and Storage?

- "Memory wall" has been discussed for nearly 30 years
  - But caches, interfaces etc. can no longer hide this wall
  - Big data, irregular access patterns, poor reuse
  - High energy costs to move large volumes of data
  - More algorithms that are data-intensive (ie, low ops/byte)
  - More and more tasks are stalled on memory/storage access
    - Tail latencies also getting worse
- Memory and storage have much higher internal bandwidth than they can transmit
- The closer computation is to the data, the lower the power

CRISP

Design Questions



Fulcrum, HPCA 2020

- Where to put the intelligence? Huge design space!
  - In the bitcells? At the chip interface? In the controller? Etc.
  - As we move further away from the bitcells, we lose bandwidth but also reduce design and area overhead
  - CRISP identified several candidate designs at different performance/complexity design points
- How to orchestrate placement of data and compute?
  - "Near data computing" is hard in heterogeneous/distributed systems if inputs are in different places
  - Important to look at workflows, not just kernels
- For memory, do we want
  - Memory that can accelerate some computations?
  - Accelerators that happen to use memory technology?
- For storage, do we still need overheads of a block-based interface?
- What does it take for emerging device technologies to find a market?
- Making the programmer's life easy is essential, or nobody will use it
  - High-level, portable abstractions

CRISP

CONIX RESEARCH CENTER

Computing on Network Infrastructure for Pervasive Perception, Cognition, and Action

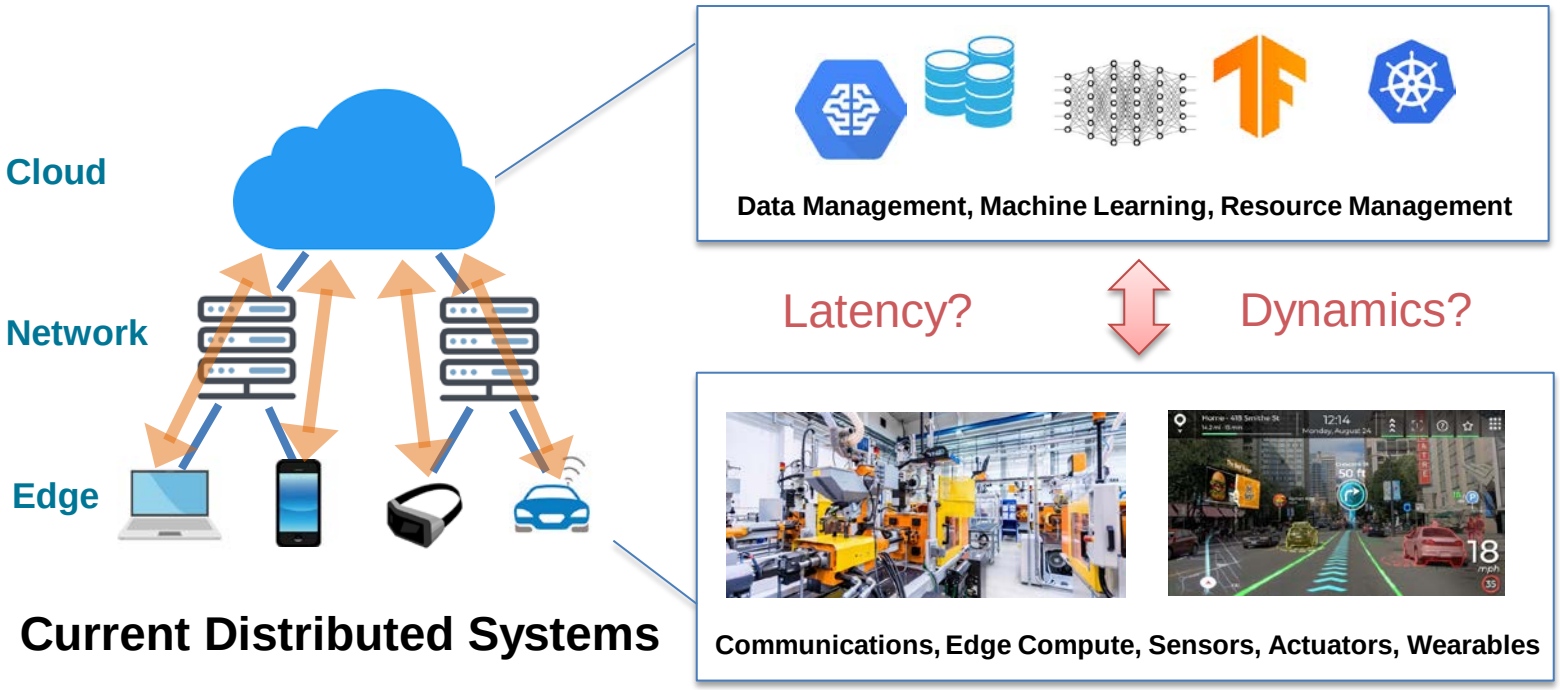
CONIX Perspective on Advances and Challenges in Semiconductor Design

Anthony Rowe
Carnegie Mellon University

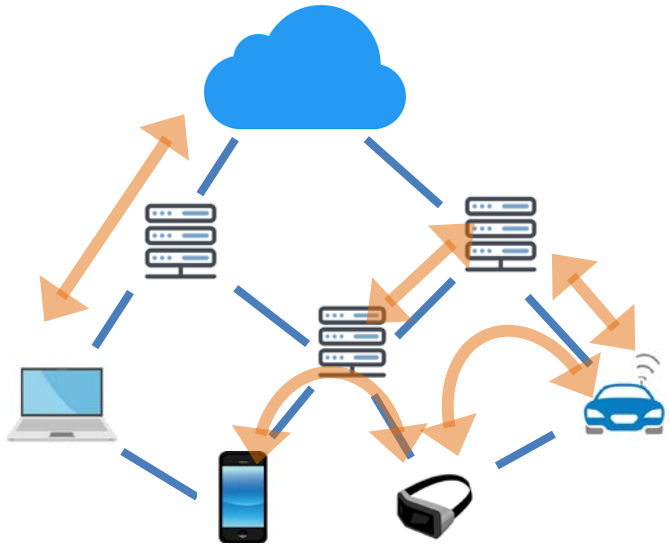
SRC JUMP

Carnegie Mellon University
George Washington University
University of California, Berkeley
University of California, Los Angeles
University of California, San Diego
University of Southern California
University of Washington

CONIX: A Distributed Compute Paradigm Shift

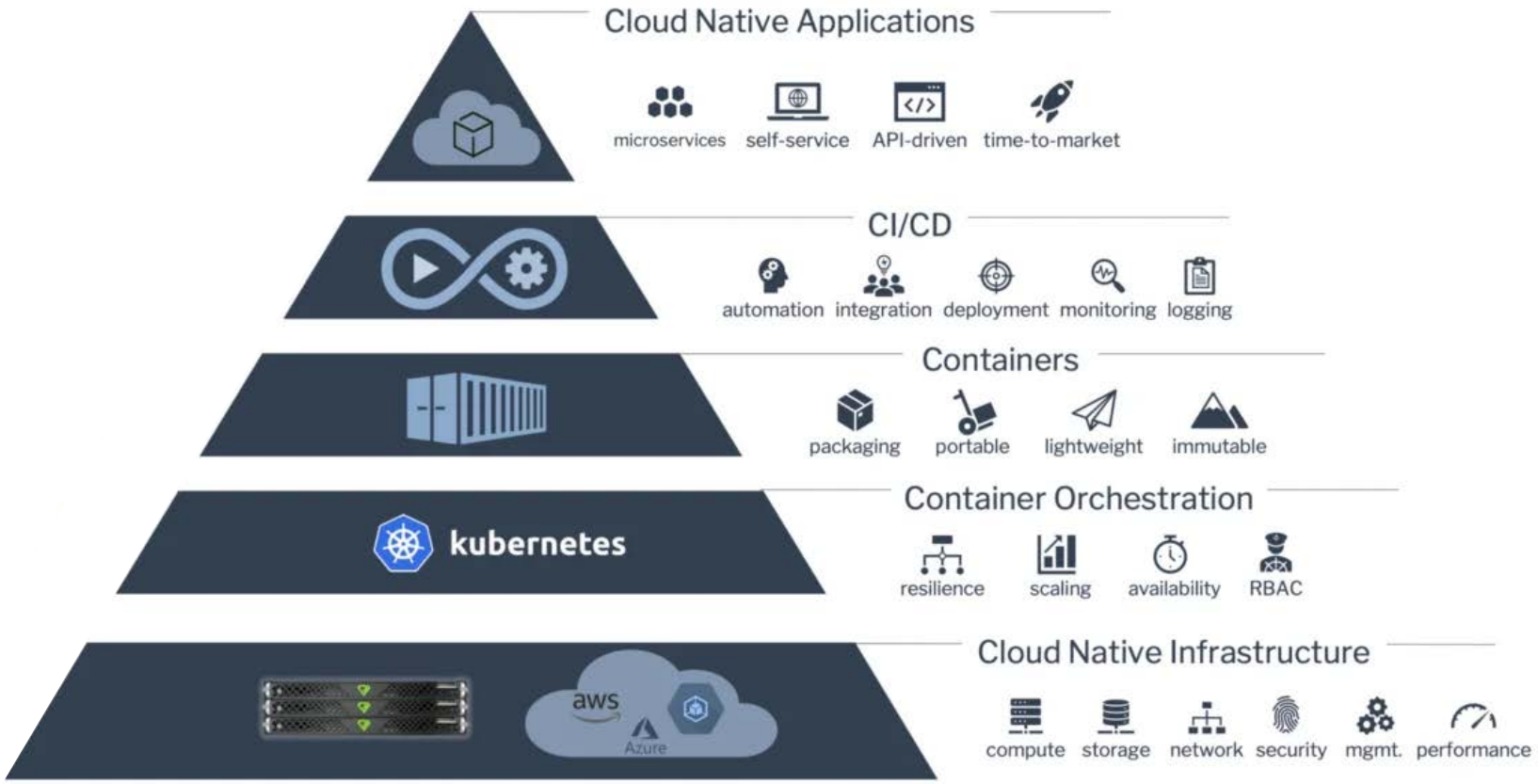


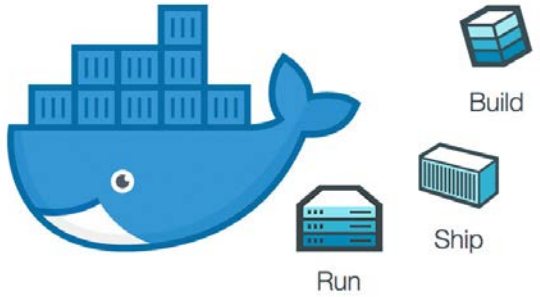
Current Distributed Systems



Next Generation Systems

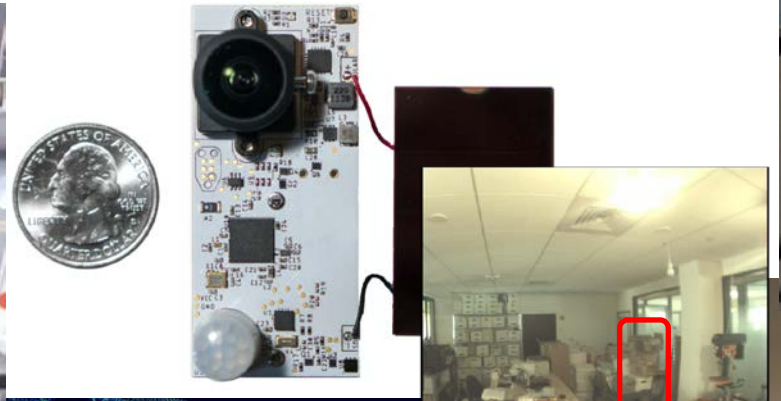
Simply bringing cloud(-native) to the edge won't cut it…

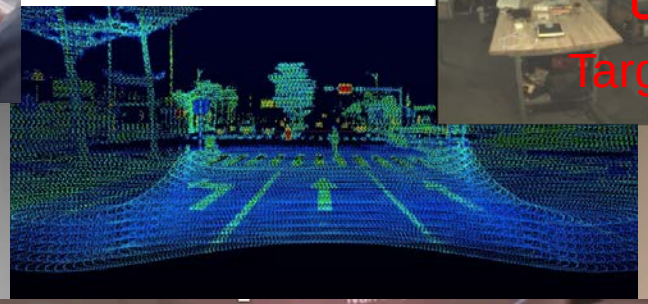


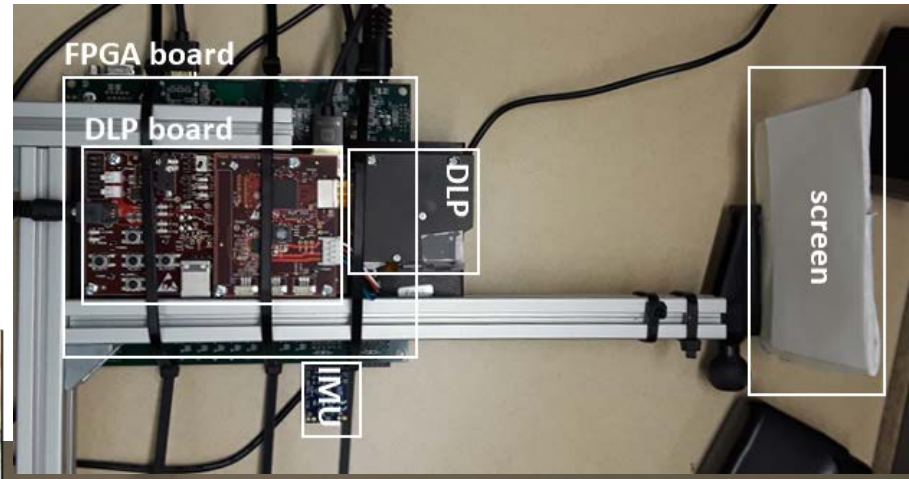


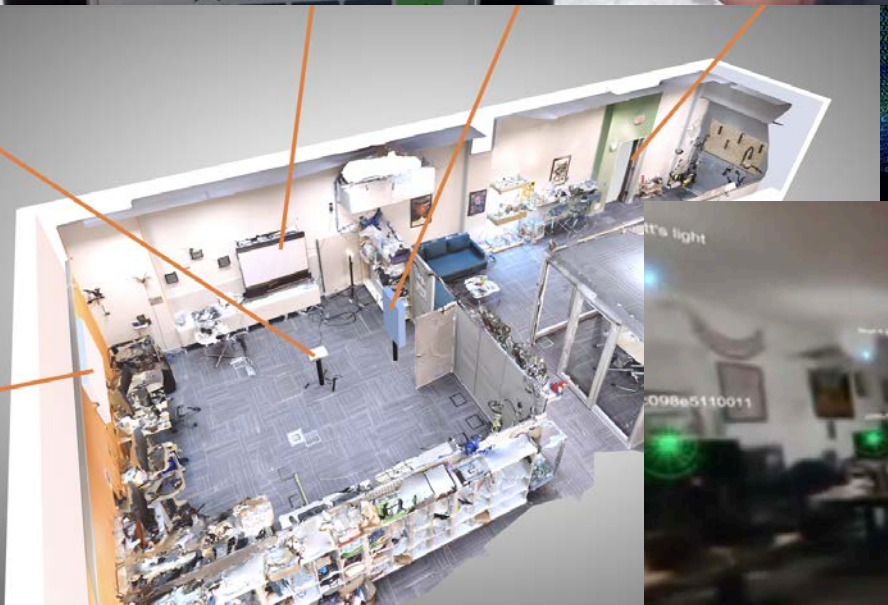
WEBASSEMBLY

Our benchmarks need to evolve...







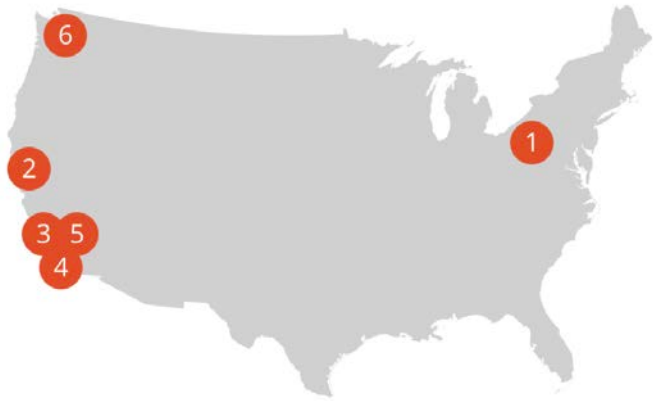


Hardware design cycles are still too slow…

Thanks!

PARTNER INSTITUTIONS

1. Carnegie Mellon University (headquarters)
2. University of California, Berkeley
3. University of California, Los Angeles
4. University of California, San Diego
5. University of Southern California
6. University of Washington



CONIX RESEARCH CENTER

Computing on Network Infrastructure for Pervasive Perception, Cognition, and Action

intel. Micron Raytheon LOCKHEED MARTIN ARM EMD PERFORMANCE MATERIALS

ARM ANALOG DEVICES IBM NORTHROP GRUMMAN tsmc SK hynix SAMSUNG

Collaboration towards Decadal Plan Goals: Advances and Challenges in Semiconductor Design Panel

Ada Gavrilovska

School of Computer Science, Georgia Tech

Applications Driving Architectures Center (ADA)

SRC JUMP

adacenter.org

@ADA_Center

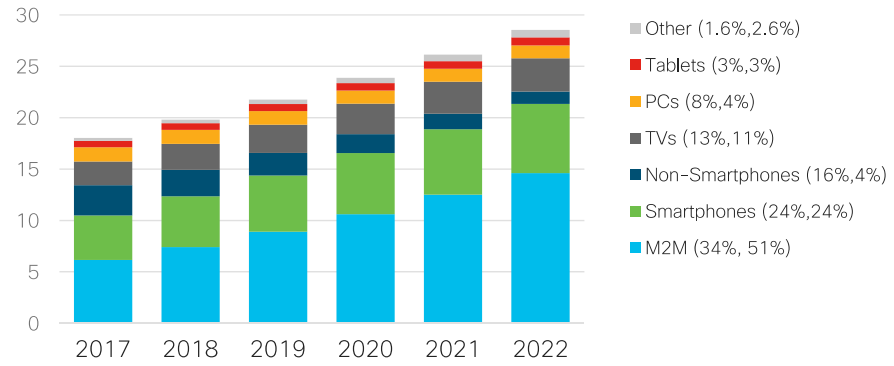
This work is supported by the Semiconductor Research Corporation (SRC) and DARPA

ada Applications Driving Architectures

Growth in data movement demand

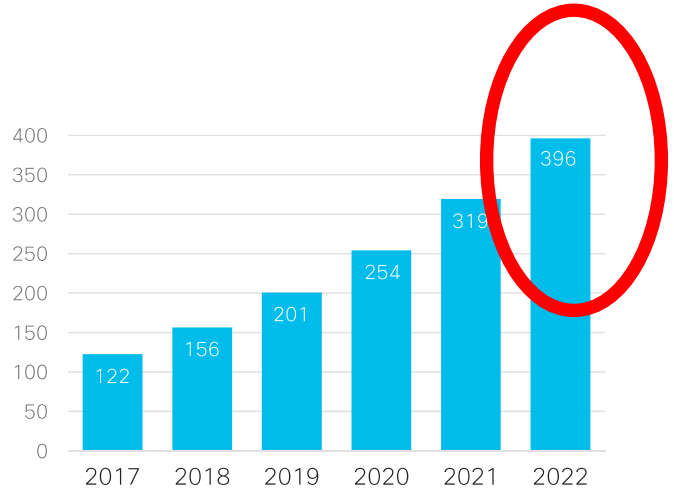
- Increase in traffic volume, number of devices, wireless

10% CAGR
2017–2022



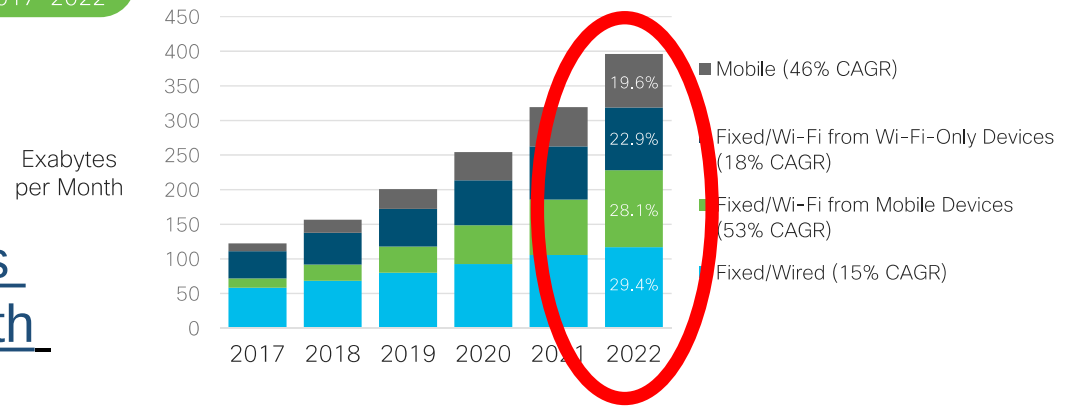
devices

26% CAGR
2017–2022



data

26% CAGR
2017–2022



wireless bandwidth

Source: Cisco Virtual Network Index

Applications Driving Architectures

Growth in data movement demand

- New bandwidth-intensive and latency-sensitive workloads

high definition video

AR/VR

SmartCity, automation

Applications Driving Architectures

Growth in data movement demand

- New bandwidth-intensive and latency-sensitive workloads

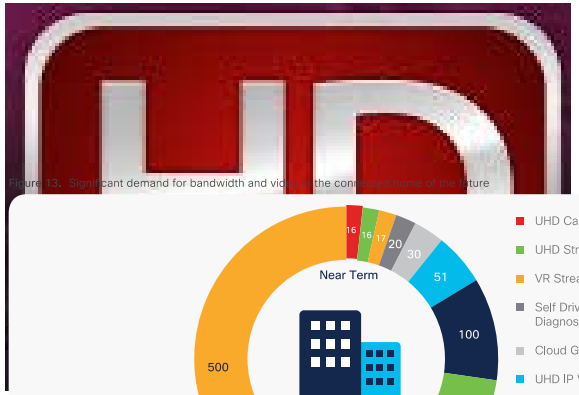


Figure 13. Significant demand for bandwidth and video in the connected home of the future

high definition video

Source: Cisco Annual

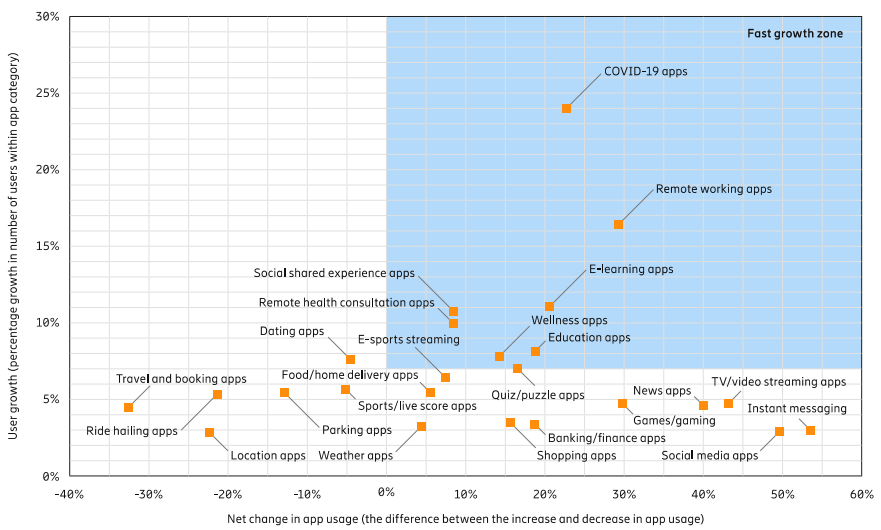
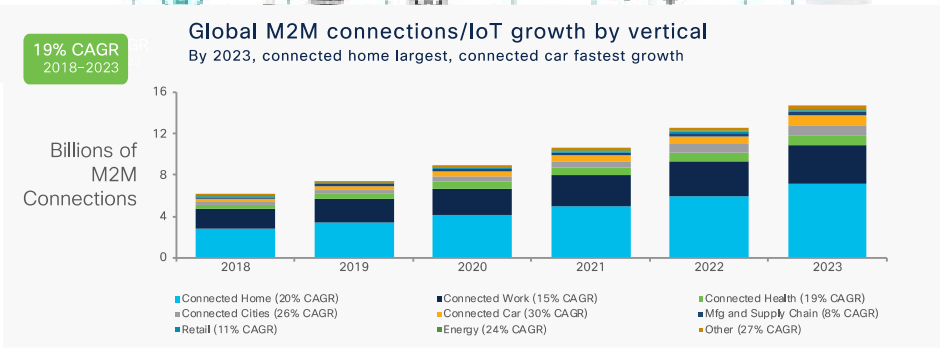


Figure 5. Global M2M connection growth by industries



Source: Cisco Annual Internet Report, 2018-

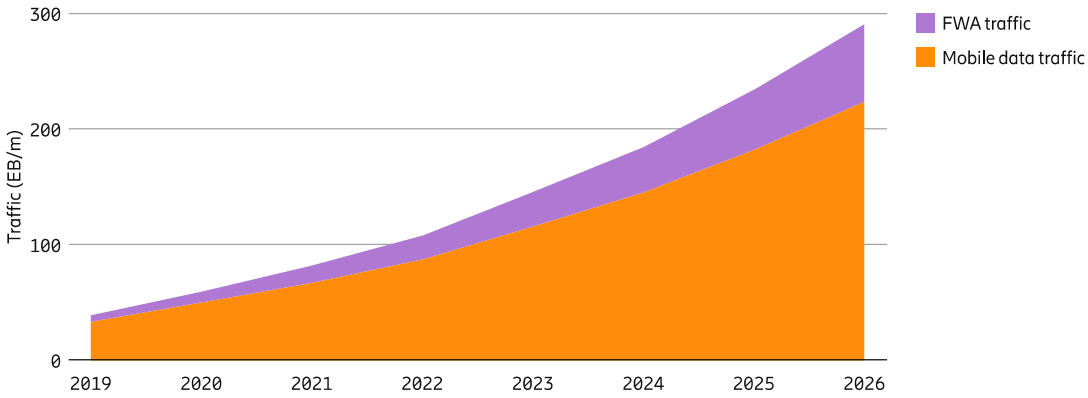
Applications Driving Architectures

What does this mean?

- Past and recent datapoints:
  - 70 TWh to run the Internet, LBNL, 06/2016
  - 50 TWh to run China's mobile network, Huawei, 07/2020
- Updated traffic predictions – no slowdown!
- EB/month cost?
  - wide range based on factors: technology, distance, system scope, … *
  - **1.8 TWh /EB**
  - => **1.2 million tons of CO2** (EPA calculator)
  - **per EB**

* https://www.wholegraindigital.com/blog/website-energy-consumption/

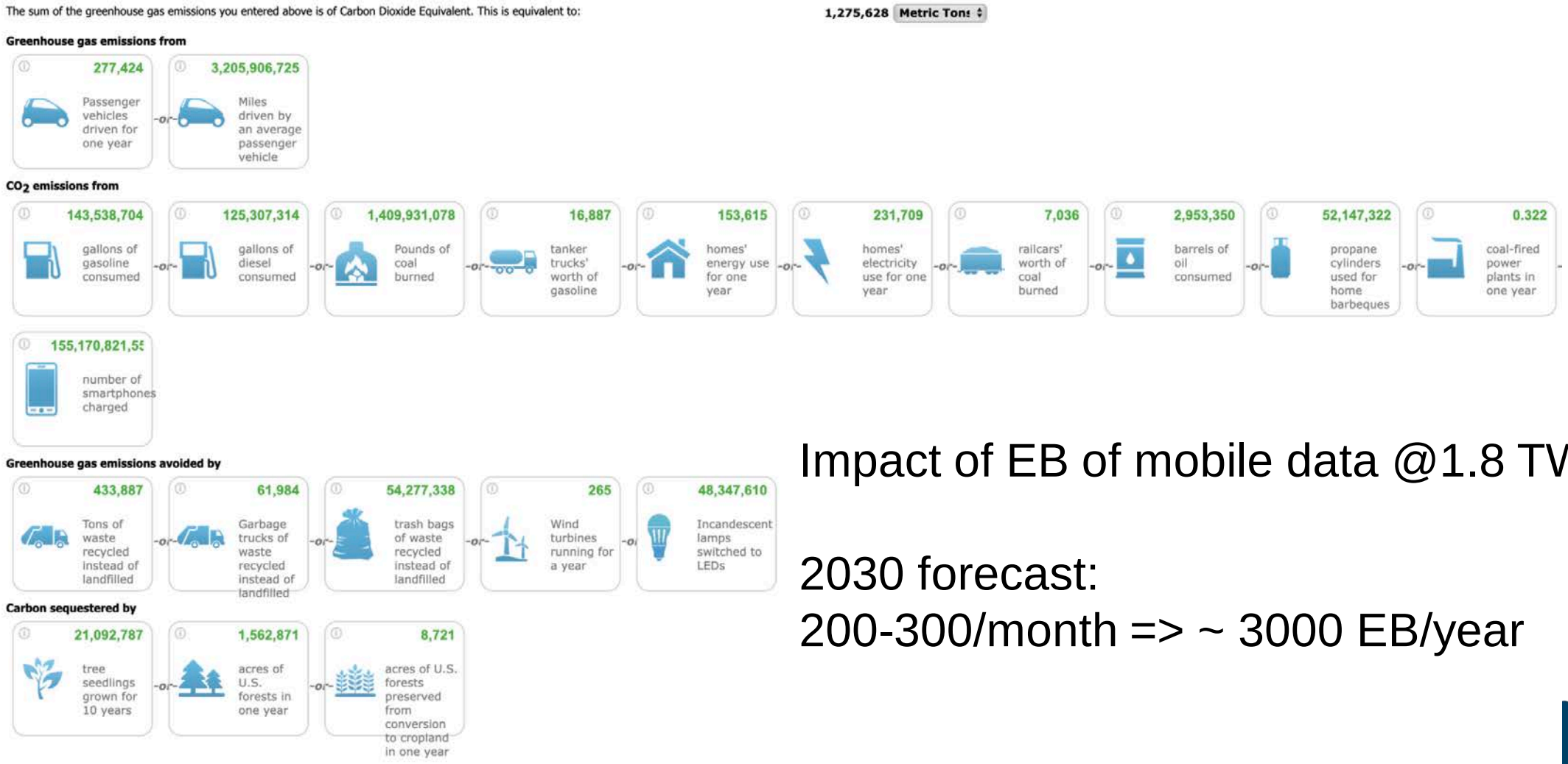
Figure 8: Mobile data and FWA traffic



Ericsson Mobility Report (11/2020)

Applications Driving Architectures

What does this mean?



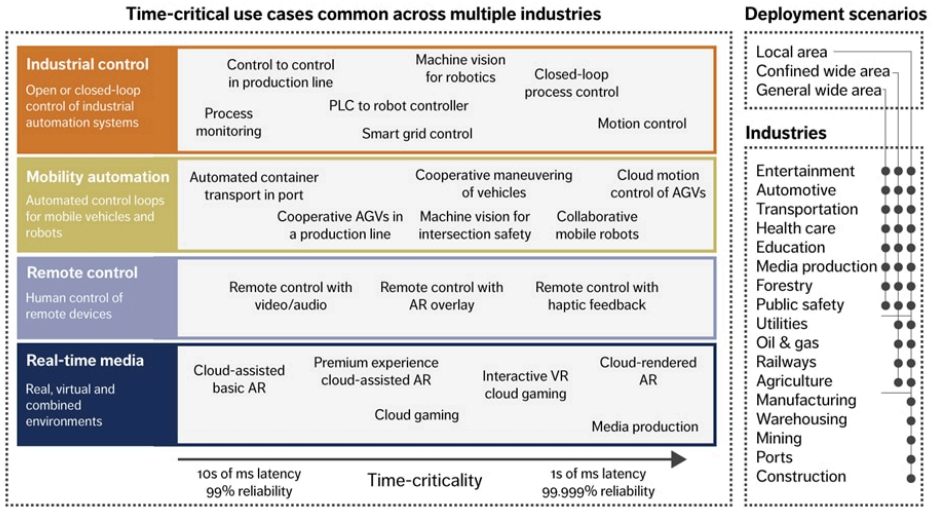
Impact of EB of mobile data @1.8 TWh/EB

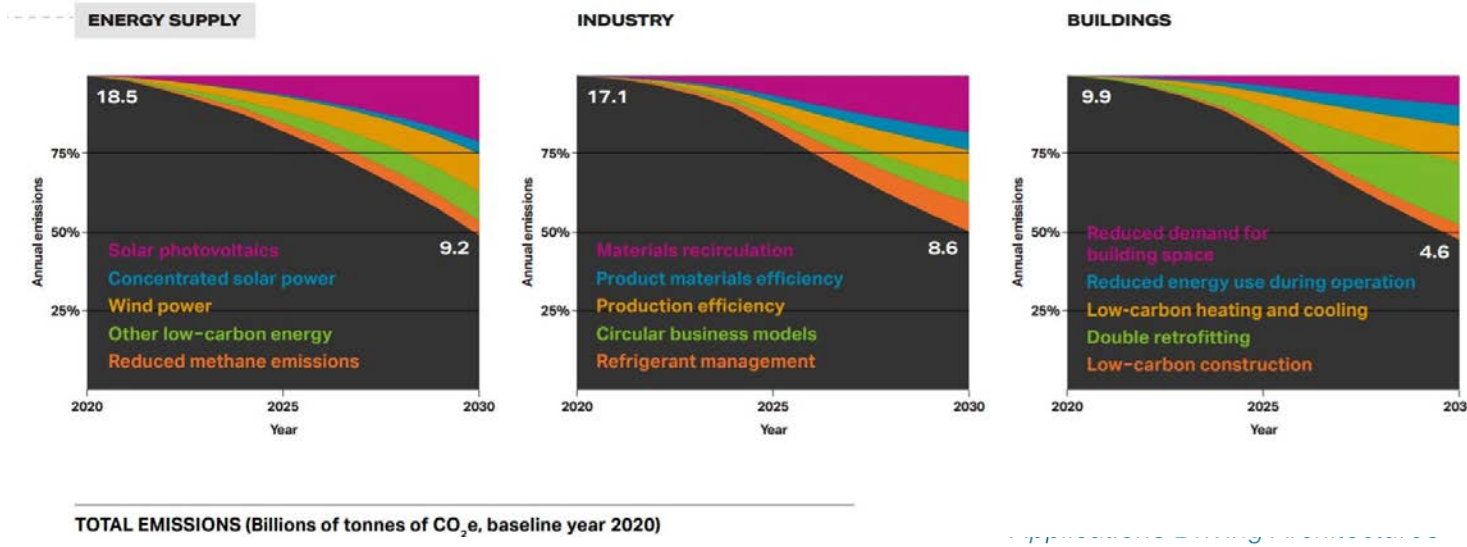
2030 forecast:
200-300/month => ~ 3000 EB/year

Applications Driving Architectures

Edge Computing and NextG Networking Opportunities

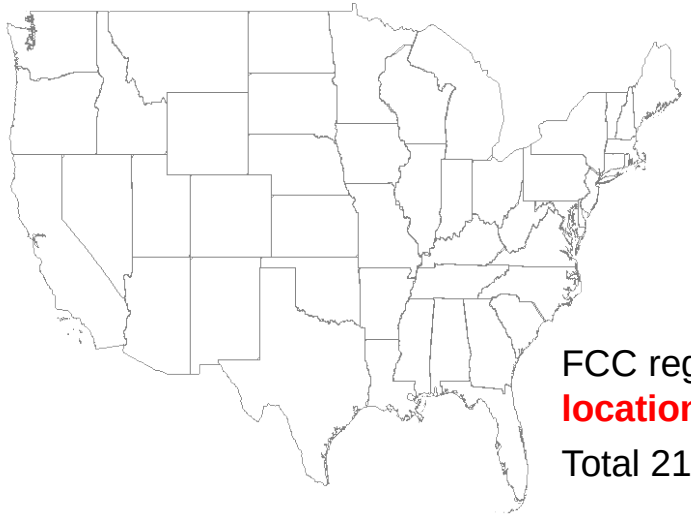
- New technologies => Energy efficiency in the data path
  - 5+G/6+WiFi/…, software functions/network server, …
- Edge computing => Reduce/remove data movement
  - Enabler for new applications
  - Aligned with UN SDG, Exponential Energy Roadmap





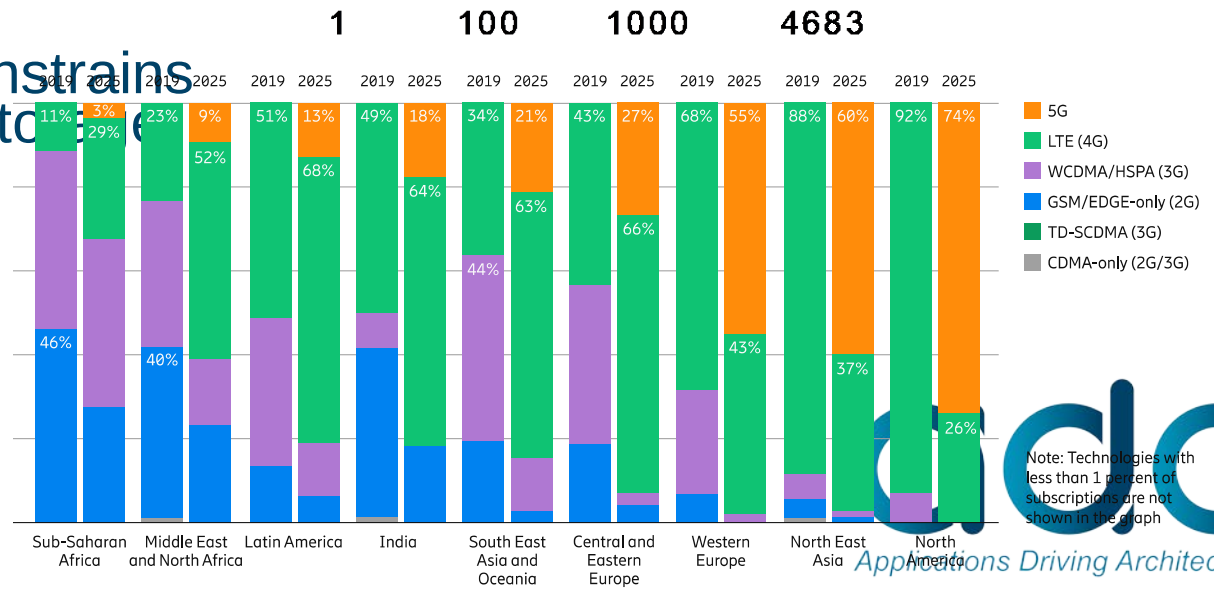
Edge Computing and NextG Networking Challenges

- Growth in demand
  - Huawei estimate 5G transition from 50TWh to 100TWh mobile network
- Deployment cost, scale, and challenges
  - O(US$1000) per location
  - Densification of infrastructure, urban deployment, ensuring coverage
  - New power, thermal, packaging constrains for compute/accelerator, memory/sto… (…) … technologies
- Datacenter-native technologies
  - Natural cooling? PUE efficiency?
- Sustainability of access



FCC registered cellular tower **locations** (Crown Castle, …)

Total 217,346, as of Mar. 2017



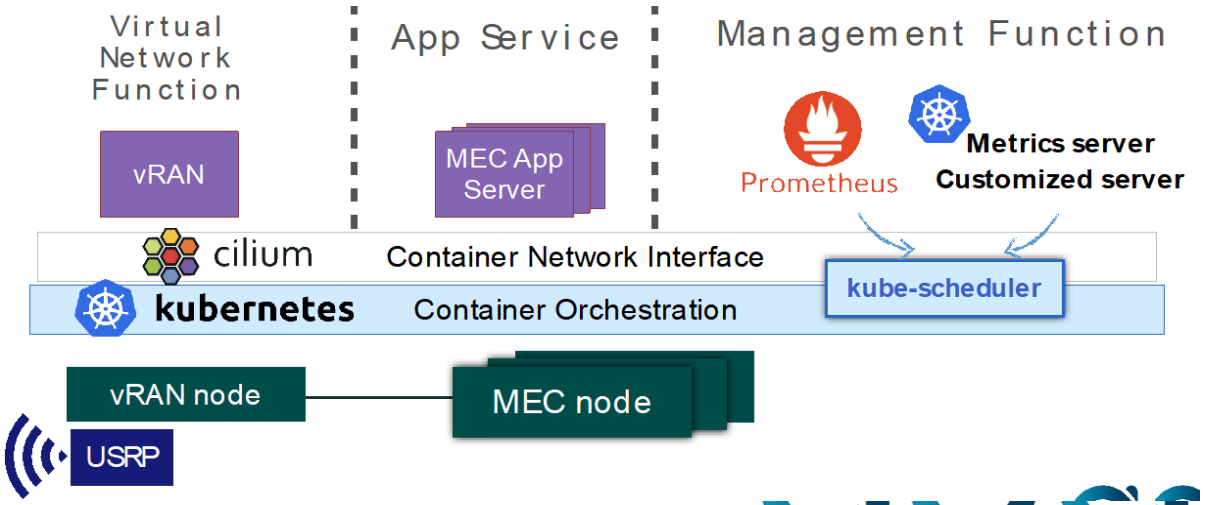
Applications Driving Architectures

End-to-end Benchmarks for Edge Computing

collaborations with UIUC/ILLIXR & other JUMP centers

benefits: Flexible deployment models, end-to-end characterization, support for different application libraries, accelerators

- **End-to-end system prototyping:**
  - client/workload; edge stack + application(s); cloud backend
- Edge infrastructure stack and services
  - resource discovery, orchestration and allocation
  - telco/mobile network stacks
  - *privacy*, based on MPC



Applications Driving Architectures

JUMP and the Decadal Plan: the challenges and opportunities for HW security

JUMP has contributed many advances to the field of HW security across the stack:

- ADA: improving the performance, communication and storage of privacy-enhancing techniques
- CONIX: major contributions to the security of accelerators, and securing Wasm for distributed compte
- CRISP: advancing security issues related to in- and near- memory processing

But there are multiple challenges ahead:

- **Increasing complexity**: accelerators, chiplets, individual components
- **Increasing connectivity**: more ‘smart’ things - 29.2 bn Arm chips shipped in 2021
- **Increasing specialization**: there is no standard next-gen chip any more

What does all of this mean for hardware security?

How and where can the academic community make meaningful contributions?

Andrea Kells, Director Research Ecosystem

arm

JUMP and the Decadal Plan: the challenges and opportunities for HW security

Security has become everyone's responsibility:

- Growing number and diversity of attack surfaces, increasing (potential) impact of breaches, complex global supply chains

The opportunities for academic contributions are therefore huge, e.g.:

- Security v. energy efficiency
- Improving memory protection architecture
- Confidential compute – still in its infancy
- Self-healing components and systems
- Aging, reliability and security

Solutions will require a holistic approach, and therefore collaboration

JUMP Centres are ideally placed: scale, convening power, visibility, reputation

Andrea Kells, Director Research Ecosystem

arm