Decentralized Edge AI Systems with Data-Sharing Trust Groups and Hardware-Accelerated zkSNARKS

Christopher Torng, Zain Asgar, Riad Wahby

Stanford AHA Weekly -- Thursday, 10/22/2020

AHA
Agile Hardware Center

Edge AI has taken off

Why... compared to centralized cloud AI?

- Latency and bandwidth resources
- Robustness of internet connection

Typically, only **inference** is done at the edge. The cloud is responsible for training and then deploying a model to the edge.


builtin.com › blockchain-ai-examples

31 Companies Making AI and Blockchain A Powerful Pair

builtin TECH JOBS TECH TOPICS TECH HUBS FOR EMPLOYERS SIGN UP | LOG IN

TASTIER COFFEE, HURRICANE PREDICTION AND FIGHTING THE OPIOID CRISIS: 31 WAYS BLOCKCHAIN & AI MAKE A POWERFUL PAIR

Sam Daley

April 2, 2019 . Updated: April 6, 2020

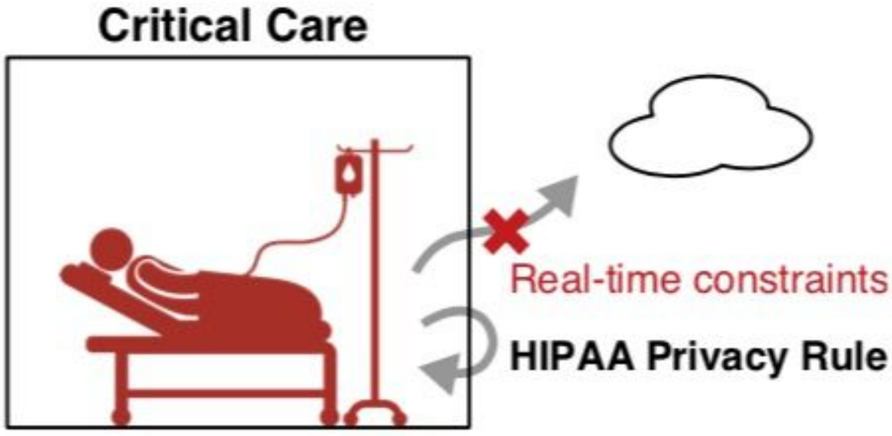
They used to be little more than buzzwords, but that's not the case anymore. Blockchain

AHA
Agile Hardware Center

Decentralized learning separates startups and research

Both inference and training at the edge

- Motivated by **data silos**
    - Personal information in your home
    - Medical info (e.g., HIPAA)
- Customize and tune models at edge
- Feed high quality data/models back



Managing blood transfusion, crystalloids, vasopressors

AHA
Agile Hardware Center

Decentralized learning at the edge is challenging...

- **Resources**
    - Accuracy tradeoff vs propagating back to cloud
    - Any single edge device is very constrained and unlikely to train to high quality
- About **federated learning** and preserving the value of new data
    - High quality data is valuable
    - Propagate value back up to the centralized model
- About **quality of service**
    - Can we guarantee the edge model has been trained for high quality?

Decentralized learning at the edge is challenging...

But the most important challenge is security and privacy...

- Today's solutions require a **single trusted central datacenter**
- Preventing **model inversion**
- Preventing **data leakage during training**
- Preventing **model stealing**
- Inference with consumer privacy/confidentiality (privacy of **raw sensor data**)
- Proofs of **data provenance** (lineage of data used to train the edge model)

What does it take to implement a simple edge AI system?

Consider a simple example of **counting people** entering a building as a shop owner of a small business:

- Track and predict foot traffic

- Plan for growth

- What kind of people enter my shop?

- Identify which spaces are available

- Identify which spaces are wasteful

(Stanford Shopping Mall)

What does it take to implement a simple edge AI system?

Forget AI... can I hack together a simple people counter?

1. IR beam at the door

2. Double IR beams at the door

3. Turnstiles

4. Security guard

But what if I am a tech-savvy owner?

Do I go for ML now??

(Stanford Shopping Mall)

AHA
Agile Hardware Center

What does it take to implement a simple edge AI system?

Go on Amazon

Unfortunate bad reviews

- "Very small and cheap for $92. Used it for our corporate gym to count entrance and exit. Around **30 people use the gym** and **it registered 12,000 in 4 hours**. Useless"

Roll over image to zoom in

People Counter, Wireless, Non Directional | Visitor Traffic Counter for Retail | Footfall Counter | Door Counter | Customer Counter | Patron Counter

Brand: SmartCoounter

108 ratings | 12 answered questions

Price: $92.45 ✓prime & FREE Returns

This item is returnable

- Attention! The device is extremely sensitive to changes in lighting conditions. Read the description carefully and do not hesitate to contact us if you have any questions.
- Easy to use: insert the batteries, install it at the exit/entrance and watch it count the visitors!
- Resettable digital footfall counter, up to 99'999 passages displayed on the LCD monitor
- Wide range of 6,5 ft / 2 m guaranteed (Up to 16 ft / 5 m in good lighting conditions)
- The door traffic counter for shop, wireless, counter sensor device

(A people counter on Amazon)

AHA
Agile Hardware Center

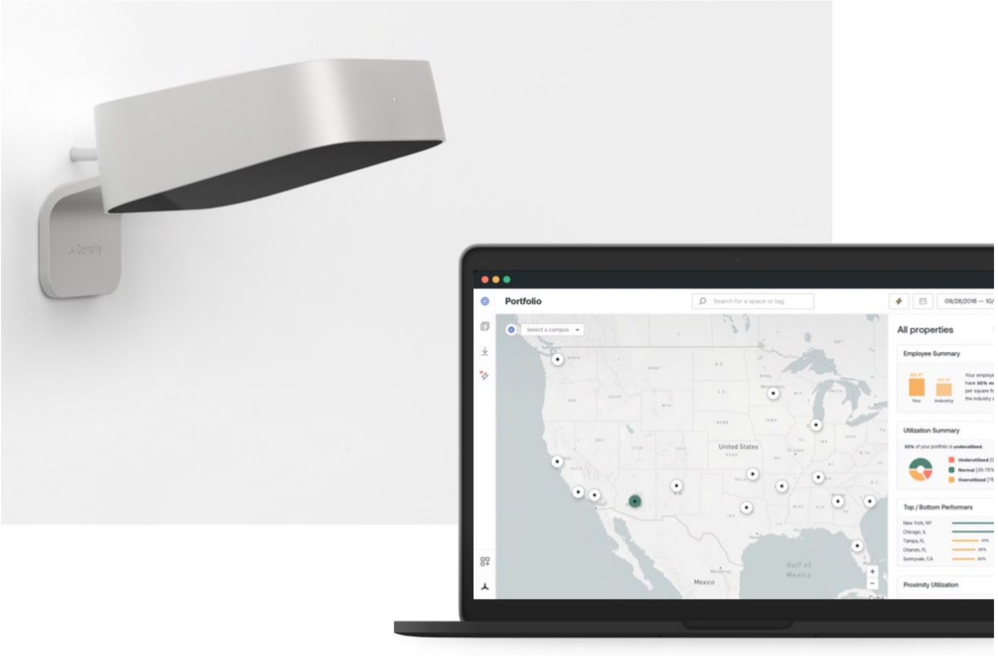
Simple edge AI system? Case study with Density.io..

Continue setting the stage for modern edge AI... can we learn lessons from *density.io?*

- Use **density sensors** and edge AI to detect people at each entrance and exit

- Businesses **count capacity** during the pandemic

- Received $51M funding



Density

Simple edge AI system? Case study with Density.io..

Technical details? Very sparse...

- Centralized learning model
- Periodic samples sent to the central controller for high-accuracy check followed by a potential update

Privacy? How do they market?

- Focus on consumer ("don't collect anything", no cameras with face rec)
- Encrypt data at rest (centrally)

WIRED

"Since it's tracking movement, it doesn't know who's coming through those doors, alleviating privacy concerns."

Simple edge AI system? Case study with Density.io..

Expanding on privacy for more feature-rich edge AI systems:

- Cannot necessarily continue the policy for "don't collect anything sensitive"

- Who is a unique visitor? If someone goes off camera / off sensor, do they suddenly count as a new visitor? This requires some storage... how long to keep that storage?

Performance and other concerns:

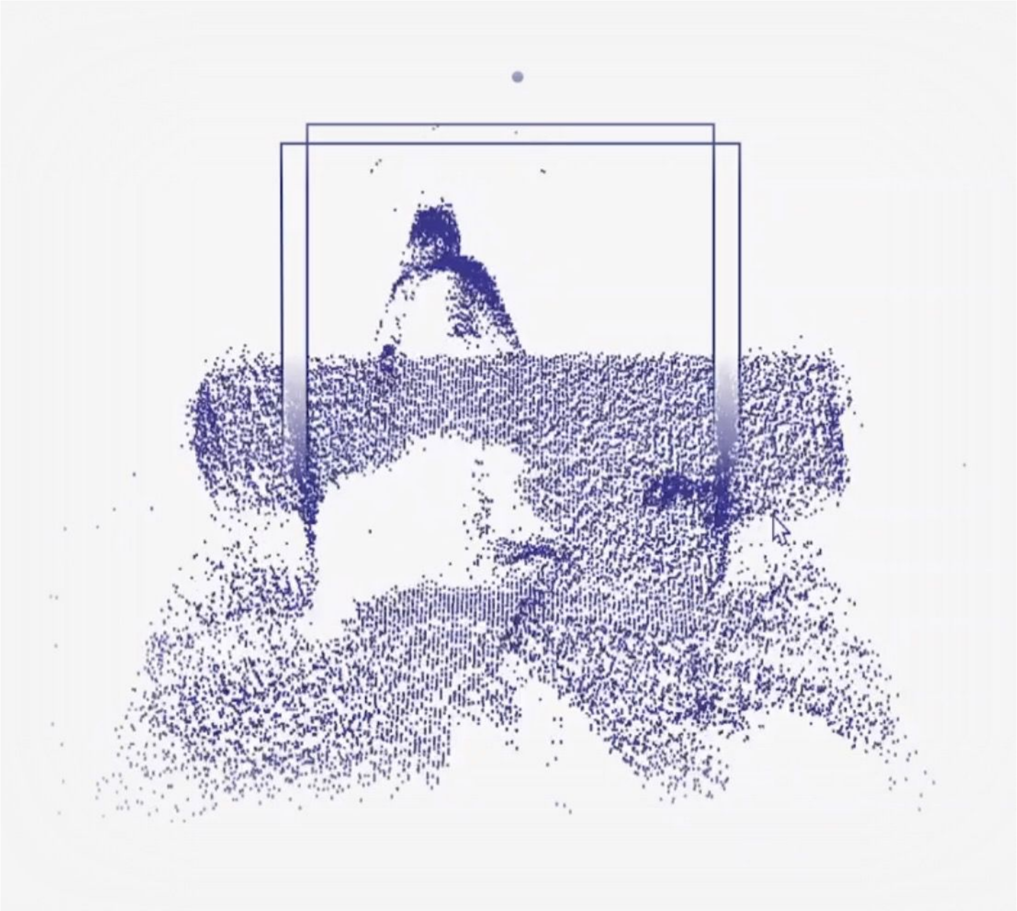
- This is not decentralized learning, so there is no way to accommodate data silos

- Security and performance are often at odds

Flashing back to the previous challenges...

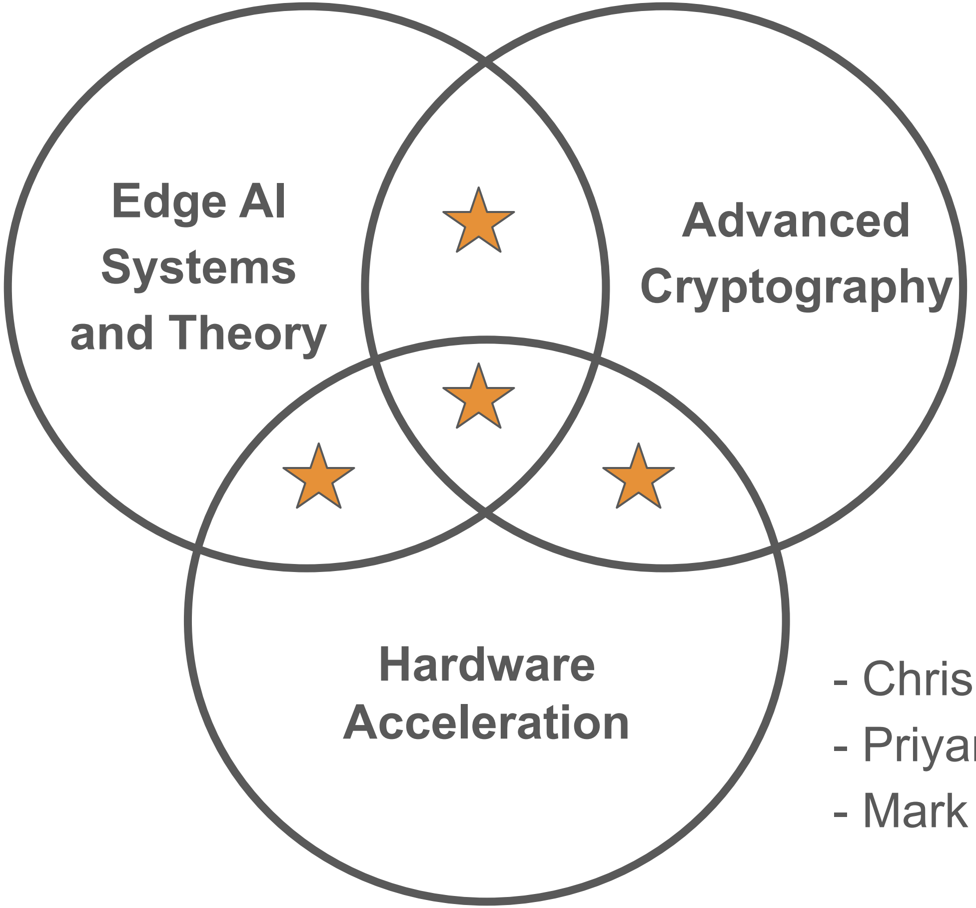
The most important challenge is security and privacy...

- Today's solutions require a **single trusted central datacenter**
- Preventing **model inversion**
- Preventing **data leakage during training**
- Preventing **model stealing**
- Inference with consumer privacy/confidentiality (privacy of **raw sensor data**)
- Proofs of **data provenance** (lineage of data used to train the edge model)



Folks interested

- Zain Asgar
- Pete Warden
- Mert Pilanci
- Sachin Katti



- Dan Boneh

- Chris Torng
- Priyanka Raina
- Mark Horowitz

Trust Groups (Zain)

Kartik Prabhu, Brian Jun*, Pete Warden**, Sachin Katti, Zain Asgar

*   Apple, Former Stanford Student
** Google, Tensorflow Lite/Micro TLM

AHA
Agile Hardware Center

Privacy issues are introduced by machine learning systems

But… Is the problem really machine learning?

or, just that machine learning allows us to look at lots of data?

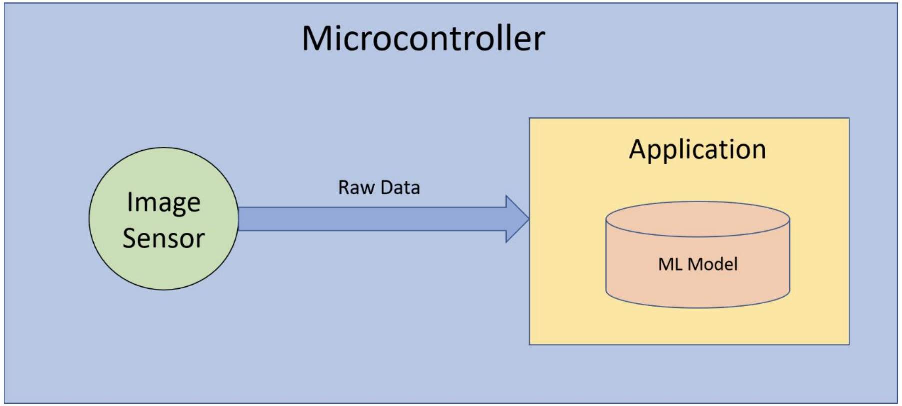
Privacy concerns and regulatory interventions

Privacy concerns and regulatory interventions

However, this is just the new Prop. 65 label.

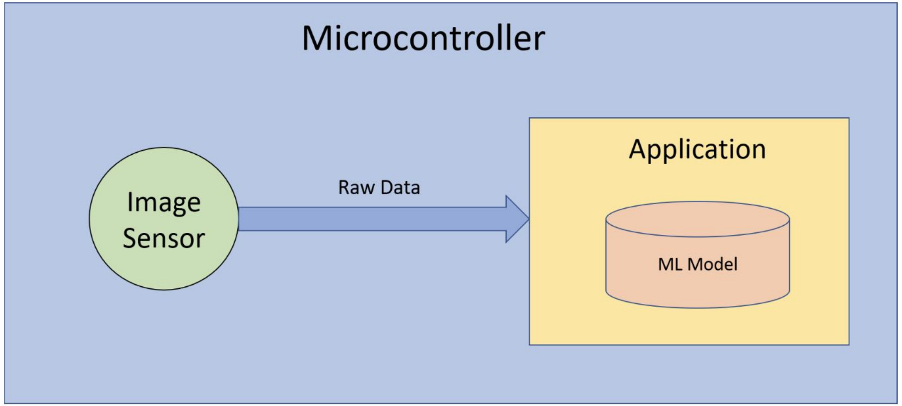
Many future devices will always be watching or listening while users become numb to privacy concerns.

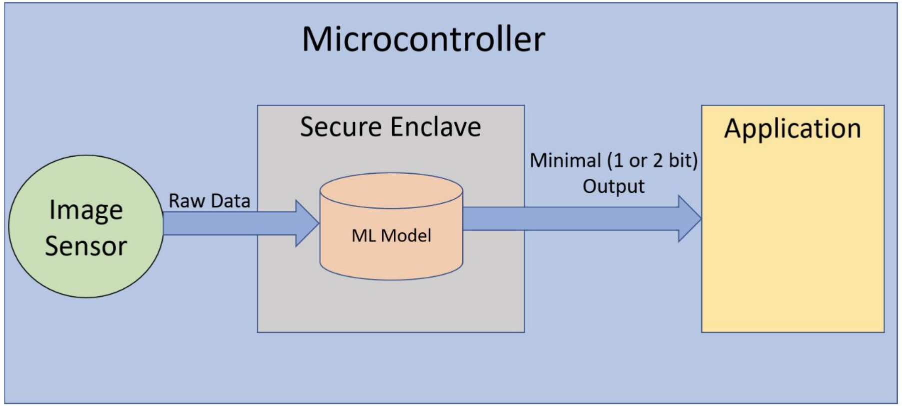
What if we can make machine learning the solution



Real threat is the raw data. Use machine learning models to help hide the raw data.

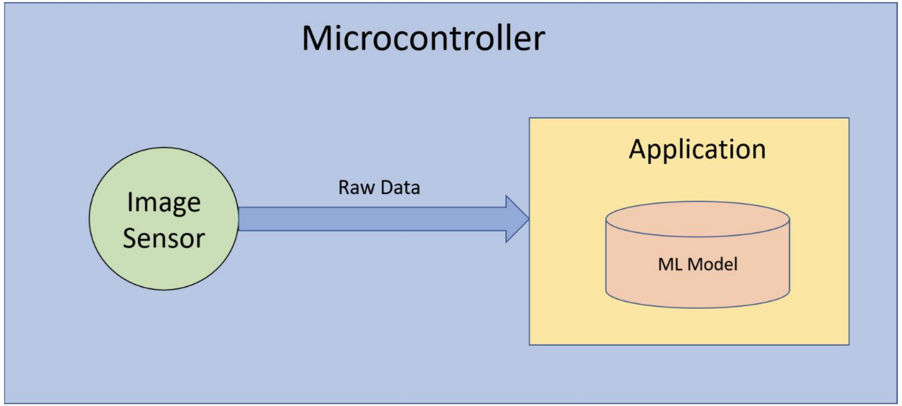
What if we can make machine learning the solution



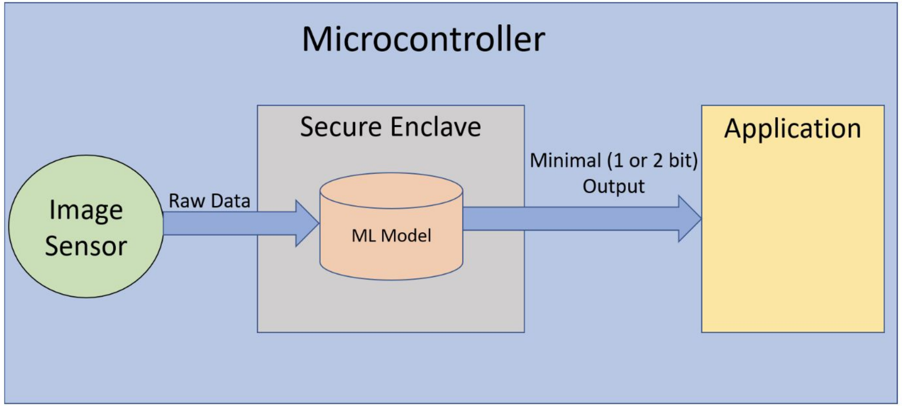


AHA
Agile Hardware Center

What if we can make machine learning the solution



Policies can be built based on model primitives and bandwidth restrictions to reduce data leakage.



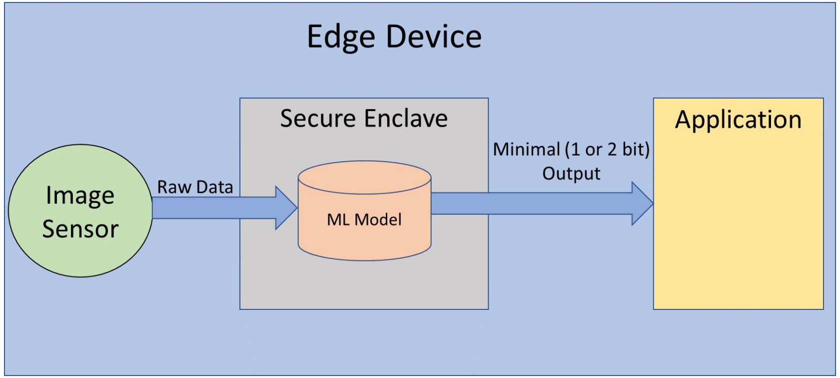
Permissions can be more fine-grained:

“applications is allowed to know that I am on the camera”

AH Agile

Using TEE for ML

- Edge devices typically have always-on sensors that feed into a machine learning model
  - Always-on sensors are a major privacy concern and access to them should be considered a threat model
- Alternate approach: use TEE to run ML models entirely within the secure world
  - Raw sensor data is contained entirely within the secure world
  - Application can only see a minimal bit output from the ML model



AHA
Agile Hardware Center

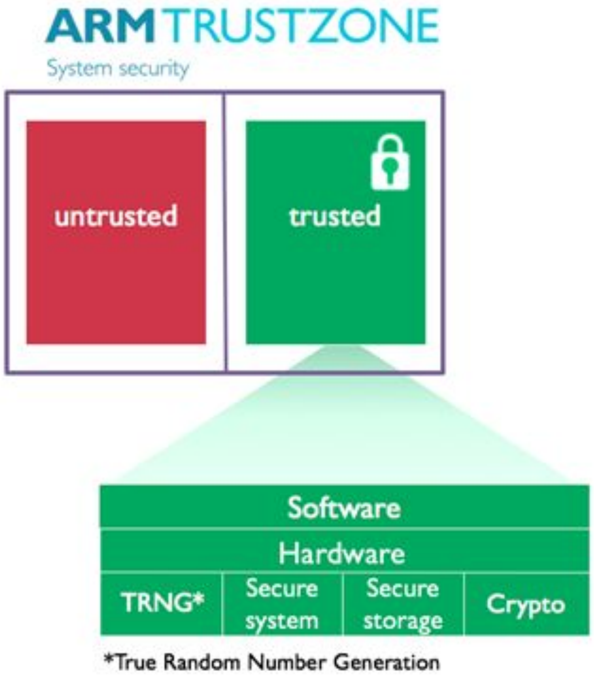
ARM TrustZone

Security extension that provides a trusted execution environment (TEE)

Memory Protection Unit (MPU) enforces accesses to memory

- Provides a sandbox for executing untrusted code
- Isolates sensitive data from nonsecure code

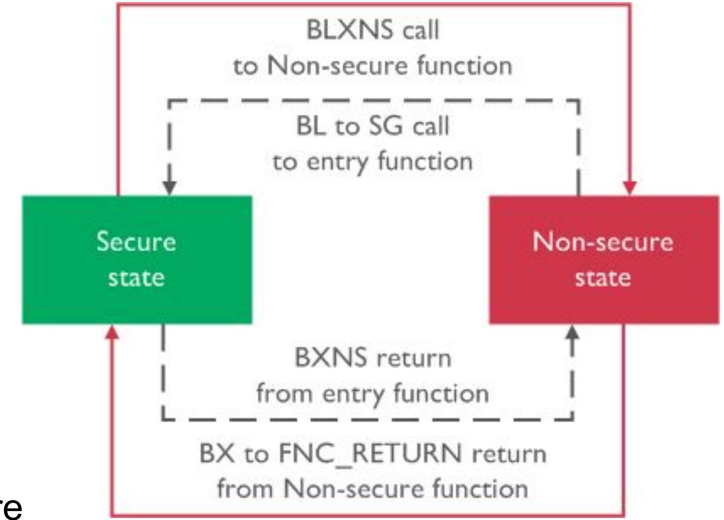
Trusted code partitions the memory space into secure and nonsecure regions at runtime



AHA
Agile Hardware Center

Transitions between Worlds

Compiler inserts instructions for transitioning between worlds



Secure -> Nonsecure:

- Variant of branch instruction that switches state

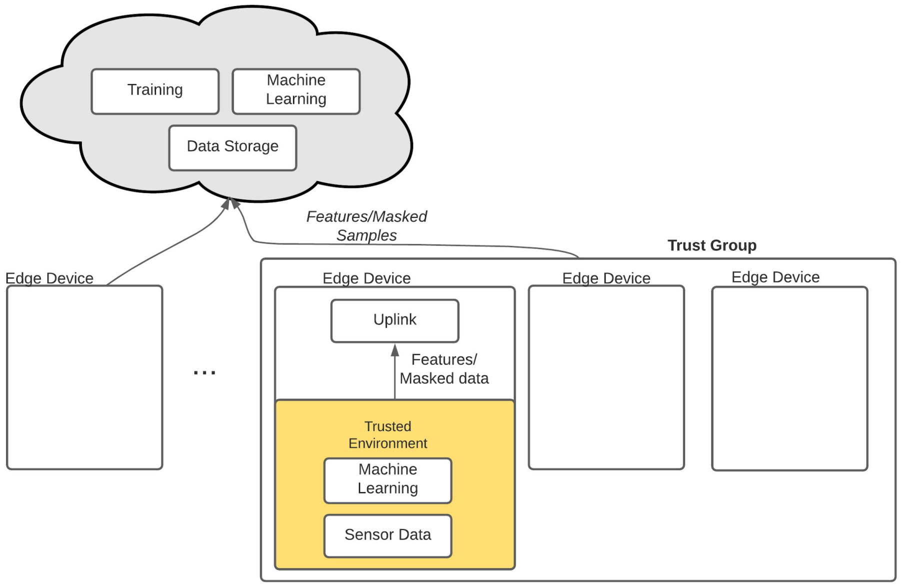
Non-secure -> Secure:

- Secure functions that can be called from nonsecure world need to be placed in a region marked as “non-secure callable”
- Prevents nonsecure code from branching into arbitrary areas in secure region

Decentralization and Trust Groups

- An environment will usually have several edge devices, e.g. smart assistant devices in a home
  - Devices may have different sensors and capabilities
- Create trust groups, which are subsets of devices that can communicate with each other
- Enable devices within trust groups to share data
  - Allows for collaboratively training models across devices
- Use a TEE to secure data from on-device sensors

Key: "Trust groups" can balance data sharing and system cost



Trust groups live "in between" the extremes of silos vs naive cloud sharing

- Trade cost (privacy, latency) against data sharing (knowledge, compute)

Our **technical approach**

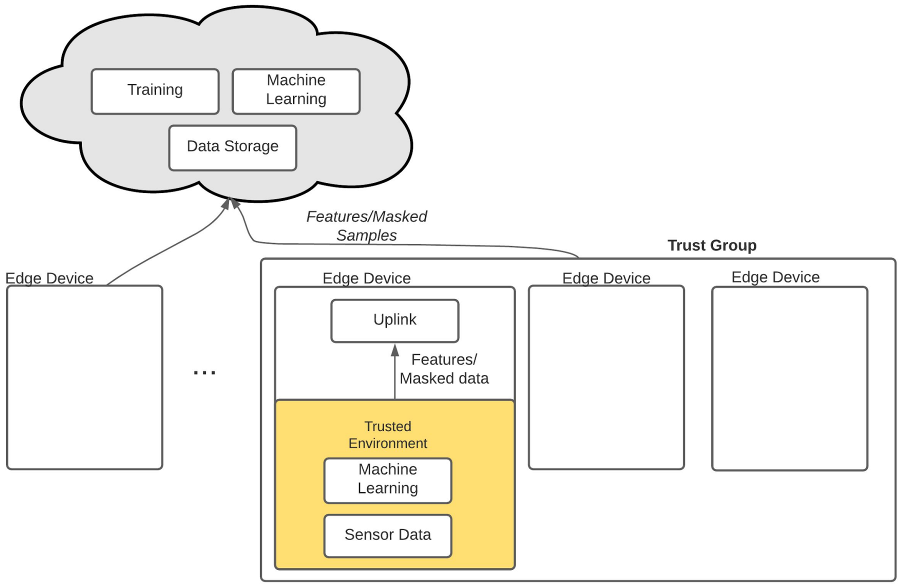
- Use TEEs
- Proofs on data provenance
- Utility functions to balance tradeoffs

AHA
Agile Hardware Center

Proof Systems and zkSNARKs (Riad)

Hardware and Summary

Key: "Trust groups" can balance data sharing and system cost



Trust groups live "in between" the extremes of silos vs naive cloud sharing

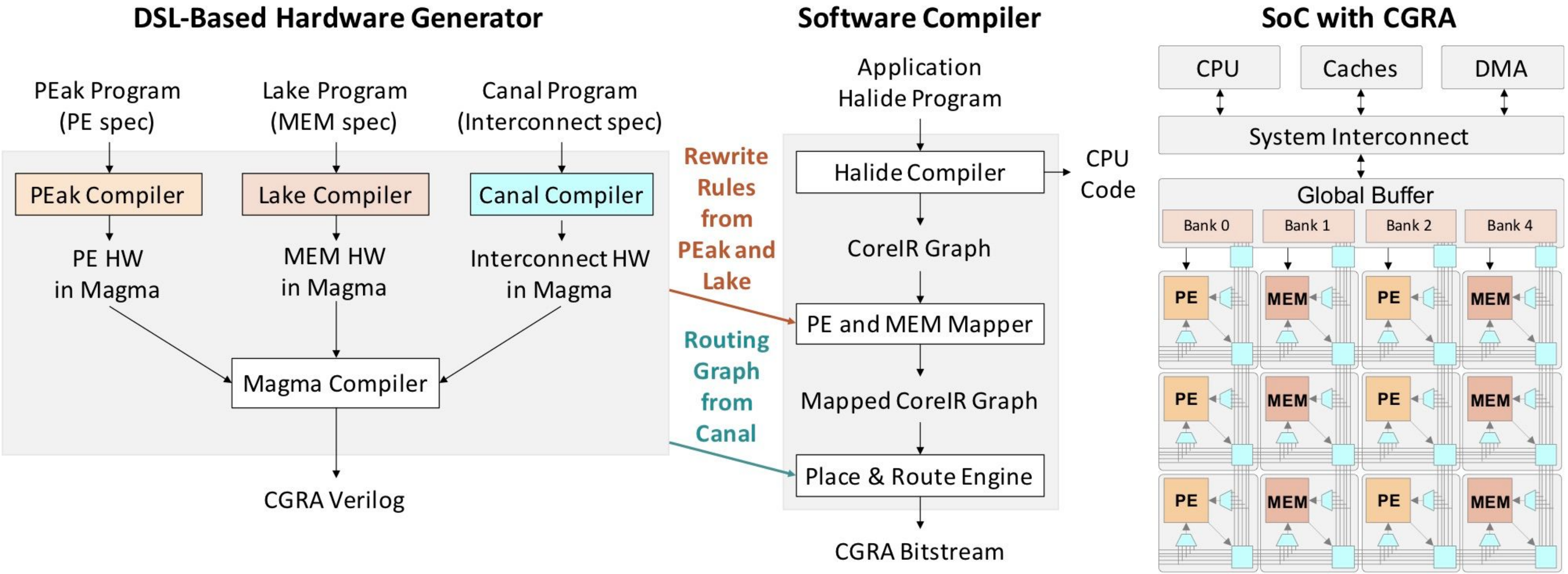
- Trade cost (privacy, latency) against data sharing (knowledge, compute)

Our **technical approach**

- Use TEEs
- Proofs on data provenance
- Utility functions to balance tradeoffs

AHA Agile Hardware Center

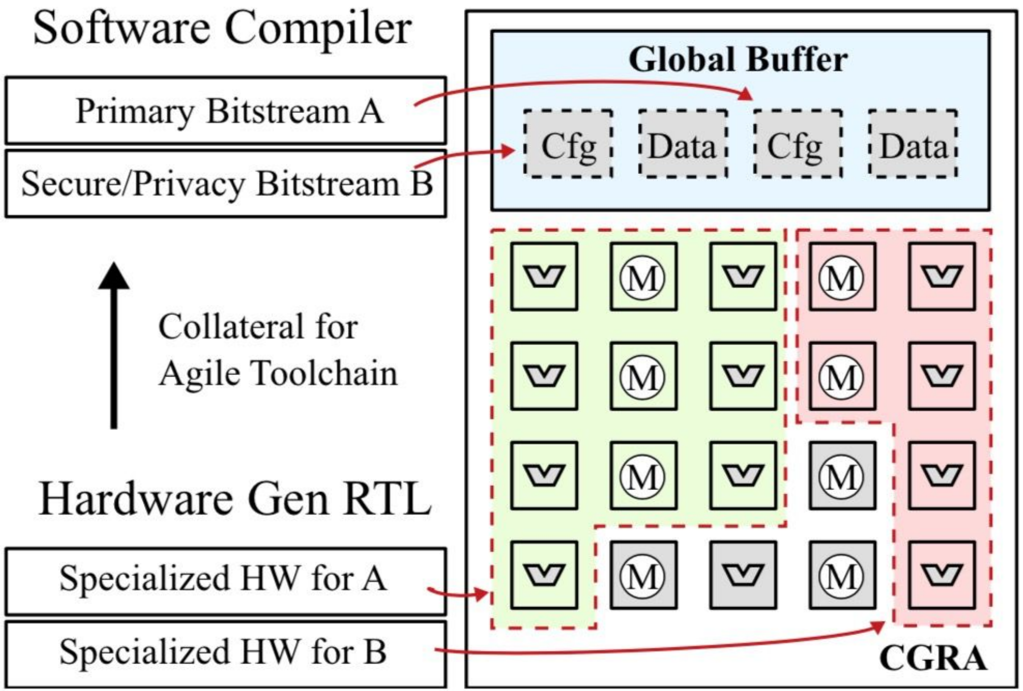
Key: Hardware acceleration can enable advanced cryptography



Existing zkSNARKs can take tens of seconds or minutes to run and can benefit from **hardware acceleration**. An agile approach can enable co-evolution of software algorithm and hardware.

AHA
Agile Hardware Center

Key: Hardware acceleration can enable advanced cryptography



Aside from AHA-style software/hardware co-evolution...

- Accelerators in secure enclaves

- CGRAs with virtualization to accelerate both primary and security/privacy kernels

AHA
Agile Hardware Center

Thinking points

TEE-based inference with policy enforcement – Create a system that can securely perform inference and provide features conforming to a set of privacy policies. Extend this to task-oriented content filtering.

Extend training to trust groups – Extend work on TEE-based inference to allow for trust groups to train models whose optimality can be verified.

Exposing neural network models in a privacy-preserving manner – Leverage zkSNARKs to allow mutually distrusting entities (e.g., trust groups, cloud service providers) to verify properties.

Thinking points

Agile hardware CGRA-based accelerator for ML workloads and advanced cryptography running inside a TEE – Create an agile hardware CGRA implementation that can be run as part of a TEE. This will pave the way for enabling flexible co-evolution of software (for both ML workloads and advanced cryptography) and hardware as part of a TEE in support of decentralized verifiable learning.

End-to-end system – A capstone project of this proposal is to create a CGRA chip and related asserts that leverage the work above to for efficient and verifiable edge AI systems.

More things to think about

Aggregate statistics in edge AI -- Prio (2018) from Dan Boneh's group has an approach to secret-share individual information (e.g., footsteps, age) across multiple servers, who conduct some process that only exposes aggregate trends.

How to interact with IP in the AI space -- DeepCloud AI (2018) wants to create a marketplace for small models. Match resource providers to applications (users) so people don't have to spin up their own clouds. E.g., exact license plate matching.

Blockchain usage -- Bias is unavoidable (e.g., favor white patients over blacks for treatment for the same disease). Avoid bias using data diversity, then be transparent about diversity by baking it into a blockchain so everyone can verify.

Backup