Single-Tenant Cloud FPGA Security

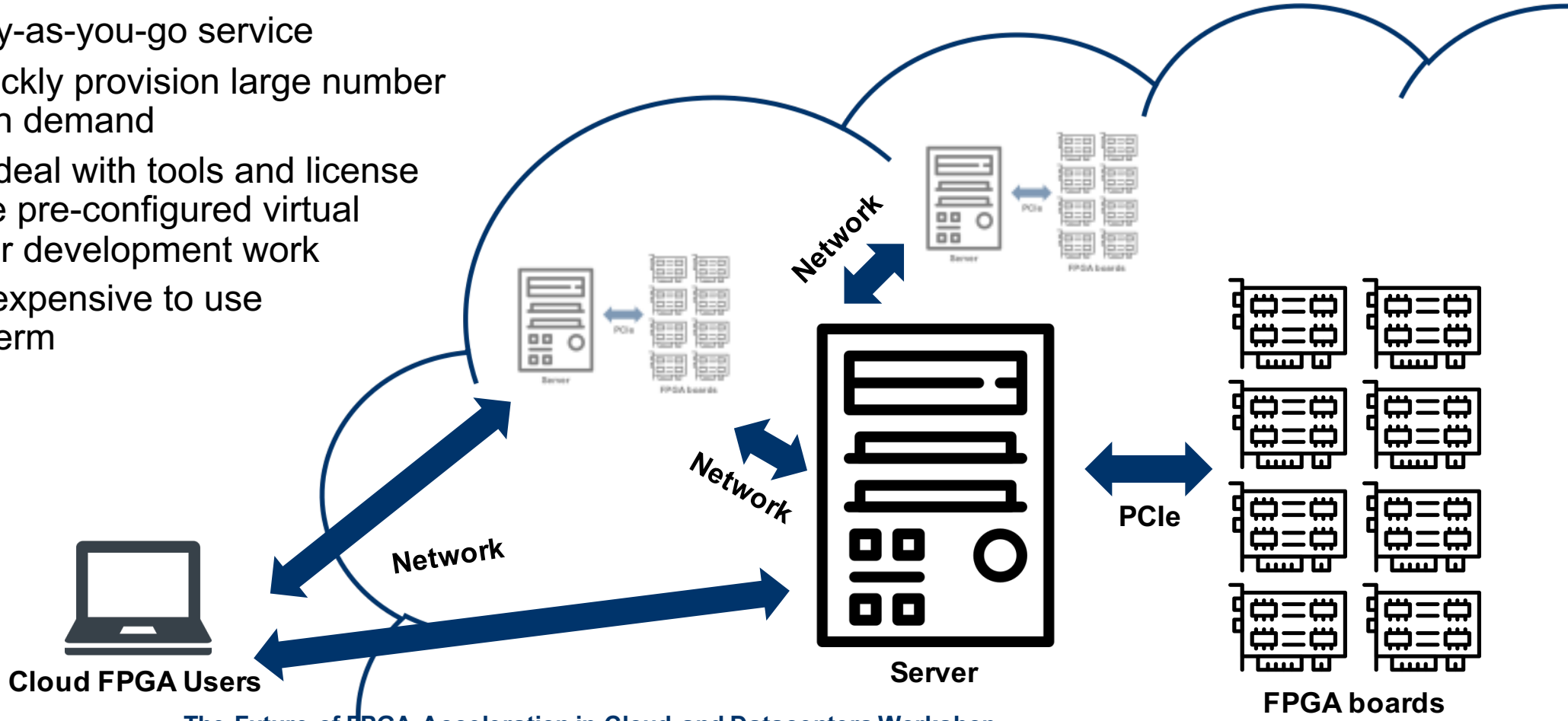
Prof. Jakub Szefer
Dept. of Electrical Engineering
Yale University

https://caslab.csl.yale.edu/

Presented at the FCCM 2020 Workshop:
The Future of FPGA-Acceleration in Cloud and Datacenters

Cloud FPGAs: FPGAs in the Cloud

- Cloud FPGA is the paradigm where FPGAs are made available to the users remotely
  - No need to purchase FPGA hardware
  - Typically pay-as-you-go service
  - Ability to quickly provision large number of FPGAs on demand
  - No need to deal with tools and license issues – use pre-configured virtual machines for development work
  - But can be expensive to use in the long term

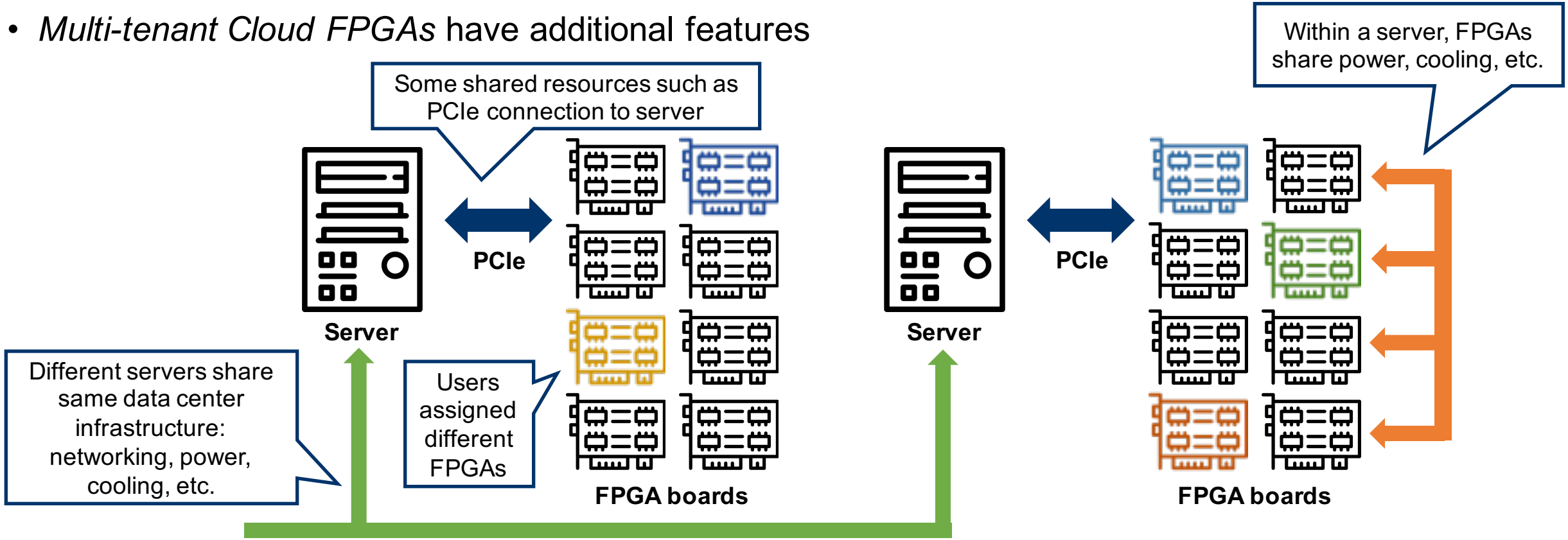


Single-Tenant Cloud FPGAs Today

- **In recent 2~3 years, there has been an emergence of public cloud providers offering FPGAs for customer use in their data centers** (as of 2019):
  - Xilinx Virtex UltraScale+: Amazon AWS, Huawei Cloud, and Alibaba Cloud
  - Xilinx Kintex UltraScale: Baidu Cloud and Tencent Cloud
  - Xilinx Alveo Accelerator: Nimbix
  - Intel Arria 10: Alibaba Cloud and OVH
  - Intel Stratix V: Texas Advanced Computing Center (TACC)
  - Intel Stratix 10: Microsoft Azure (for AI applications)
- Most infrastructures let users load any hardware design (with limitations imposed by the underlying FPGA and design rule checks)
- Some infrastructures only give indirect access to FPGA, e.g., via HLS

Sharing Resources in Single-Tenant Cloud FPGAs

- *Single-tenant Cloud FPGAs* allow only one user to use each FPGA at a time
  - Temporal multiplexing of same FPGA among different users
  - Spatial sharing of the server, sever rack, and data center by different users
- *Multi-tenant Cloud FPGAs* have additional features



Single-Tenant Cloud FPGAs Threat Model

Threat model for single-tenant Cloud FPGAs:

Additional threats exist in multi-tenant Cloud FPGAs

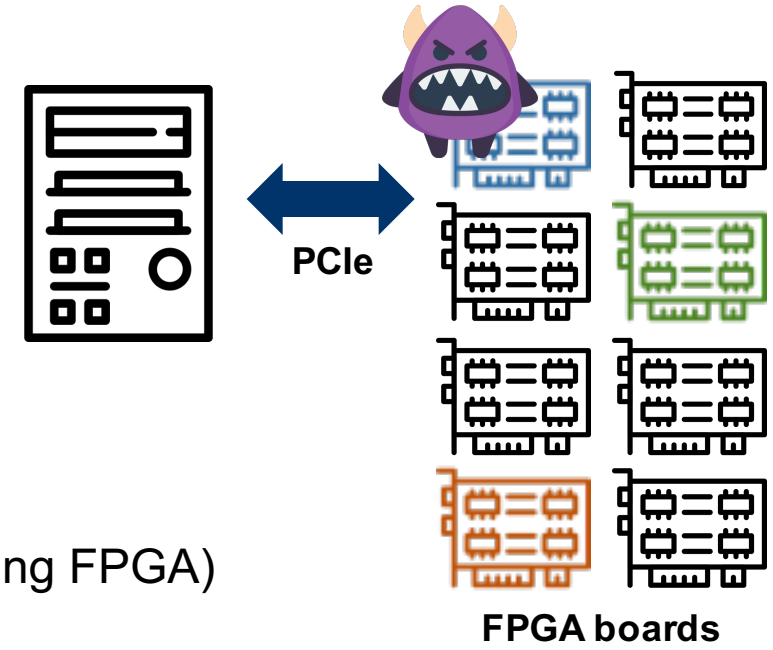
- Cloud FPGA provider and data center are secured
  - Physical attacks on FPGAs are not considered (side-channels using physical probes, etc.)

The providers are in business of rending resources, so have reputation to keep, and thus implement protections

- Users could load potentially malicious FPGA bitstreams (AFI)

Detecting malicious circuits in bitstreams or DCPs is a difficult problem

  - Leak information from one FPGA to another (covert channel)
  - Steal information from another FPGA (side channel)
  - Steal information from the shell (side channel)
  - Reverse engineer Cloud FPGA infrastructure
  - Induce faults, waste power, waste resources (e.g. generate PCI traffic that blocks others form accessing FPGA)
  - For all, use: thermal, cross-talk, or power attacks

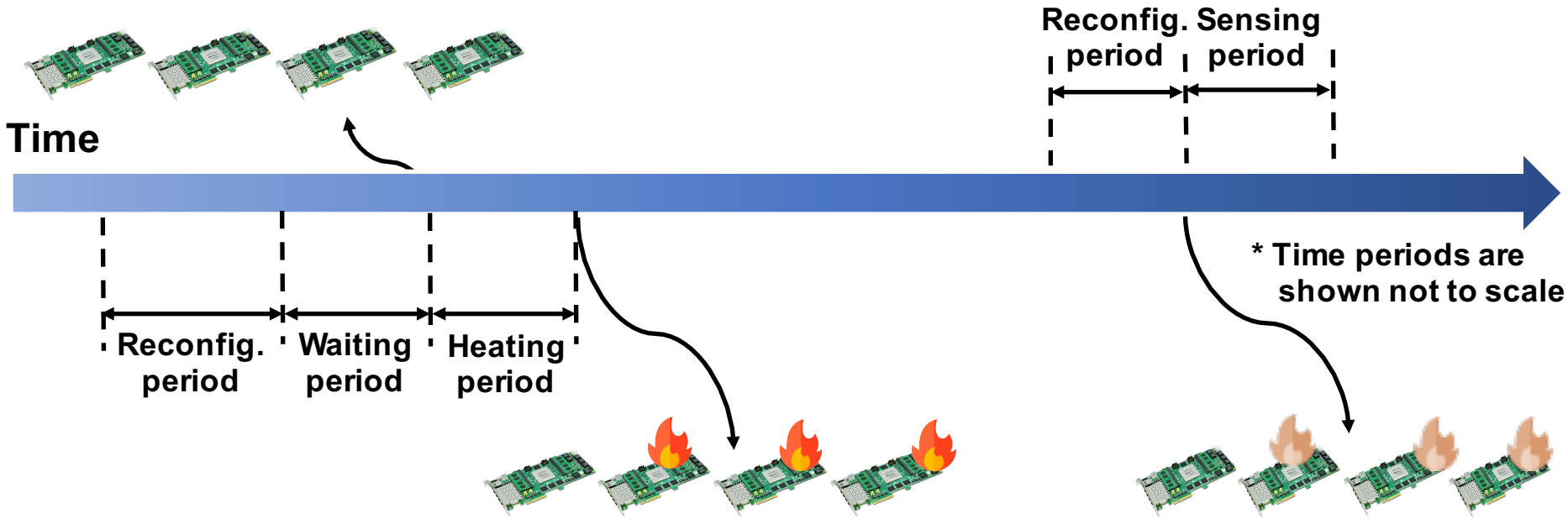


Thermal Channels in Cloud FPGAs

Covert Channels Using Temperature (Heat)

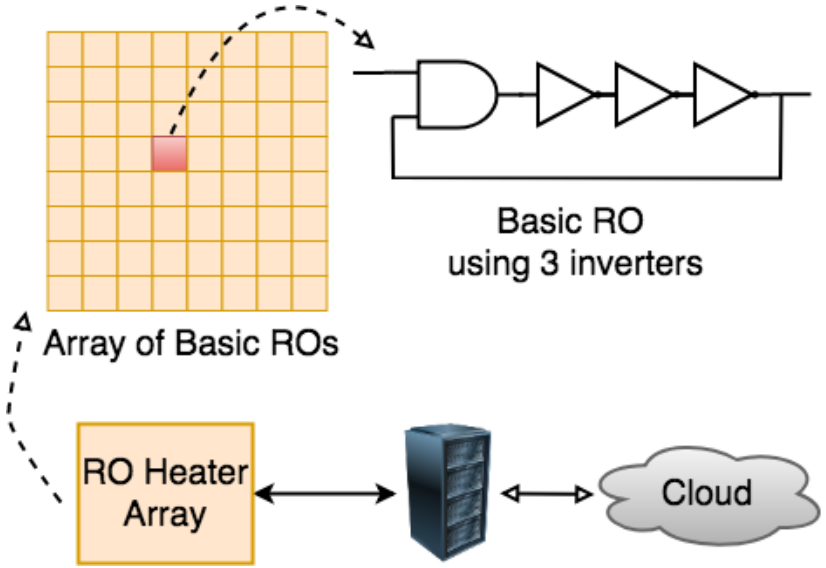
Shanquan Tian and Jakub Szefer, "Temporal Thermal Covert Channels in Cloud FPGAs", in Proceedings of the International Symposium on Field-Programmable Gate Arrays (FPGA), February 2019.

- Cloud FPGAs leverage temporal sharing of the FPGA resources among users
- The sender and receiver share or can access the same set of FPGAs
- **Idea: use thermal state of FPGA to send information between users**
  - Custom circuits can heat up (or not) FPGA to send 1 (or 0)
  - Heat dissipates on order of few minutes
  - Another user can be loaded onto FPGA before FPGA fully cools off, to receive the information



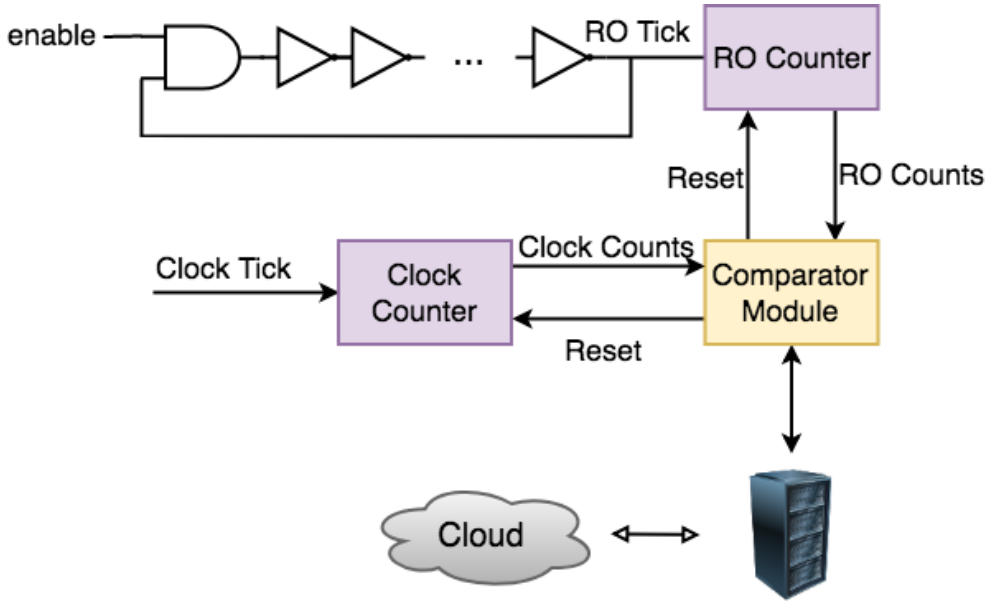
Heating FPGAs

- An array of free-running Ring Oscillators (ROs) can be used to generate a lot of heat
- Size of the array determines the amount of heat that can be generated
  - Need sufficient size to heat the FPGA faster than it is being cooled off
  - Too large array can overheat FPGA
  - Too large array can drain too much power and crash FPGA
  - Too much heat or power may be detected – need to balance all the considerations



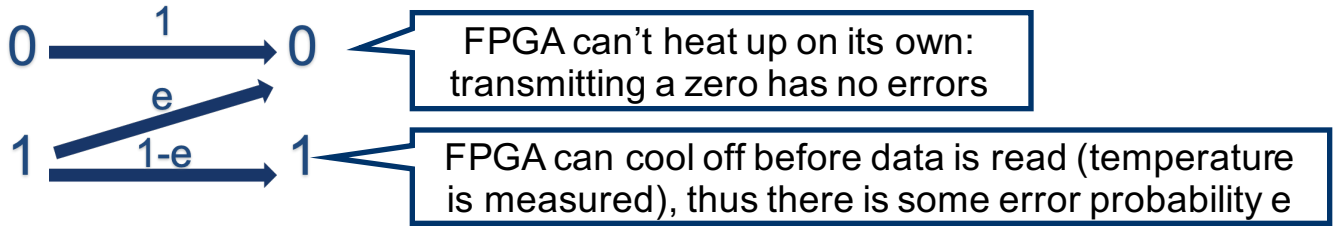
Measuring FPGA Temperature

- Ring Oscillators (RO) can be used as a temperature-to-frequency transducer suitable for thermal monitoring on FPGAs
- RO frequency depends on temperature (and voltage)
- Counting RO oscillations in a fixed time period, can be used to detect temperature (and voltage) changes

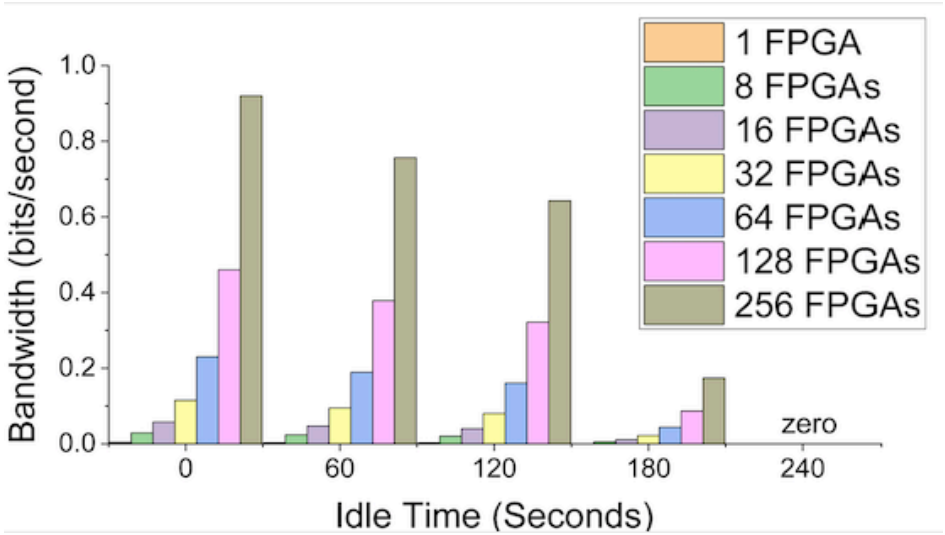


Thermal Covert Channel Results

- During transmission, there is possibility that FPGA cools off, loosing some information
  - Possible transitions



- Many FPGAs can be used in parallel to easily increase the bandwidth and error correction can be applied
- Can demonstrate practical channel on Cloud FPGAs
- Research challenges:
  - Not easy send more than 1 bit per FPGA
  - Need to locate correct FPGA and measure temperature within the time it remains heated
  - Effects of cooling and changing data center temperatures



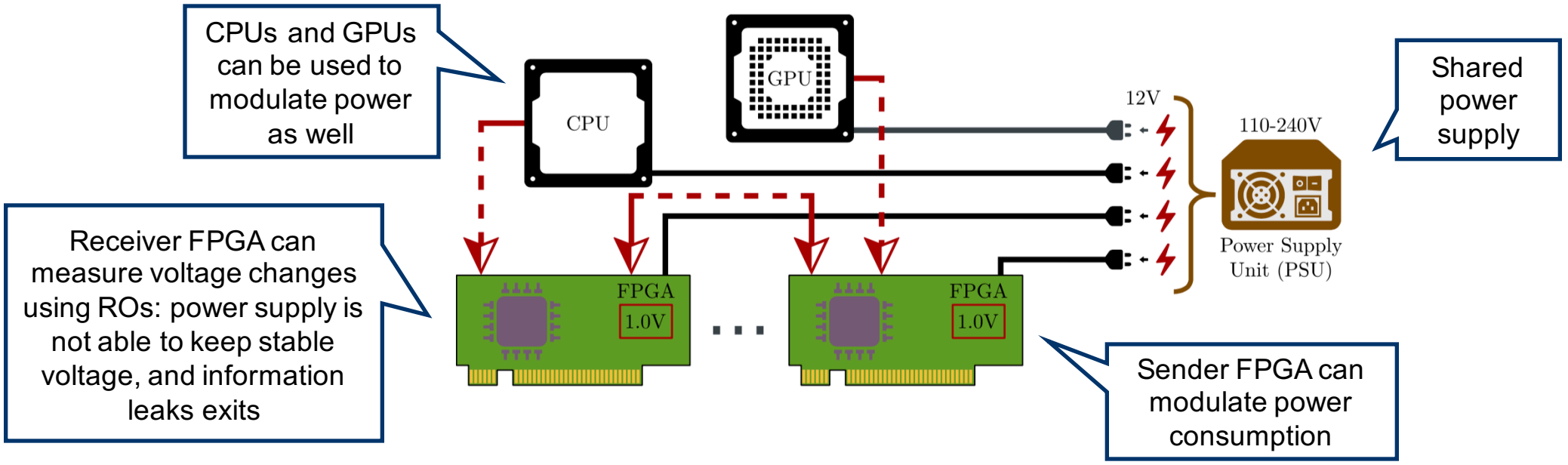
Voltage Channels in Cloud FPGAs

Shared Datacenter Infrastructure and Information Leaks

Ilias Giechaskiel, Kasper Rasmussen, and Jakub Szefer, "CAPSULe: Cross-FPGA Covert-Channel Attacks through Power Supply Unit Leakage", in Proceedings of the IEEE Symposium on Security and Privacy (S&P), May 2020.

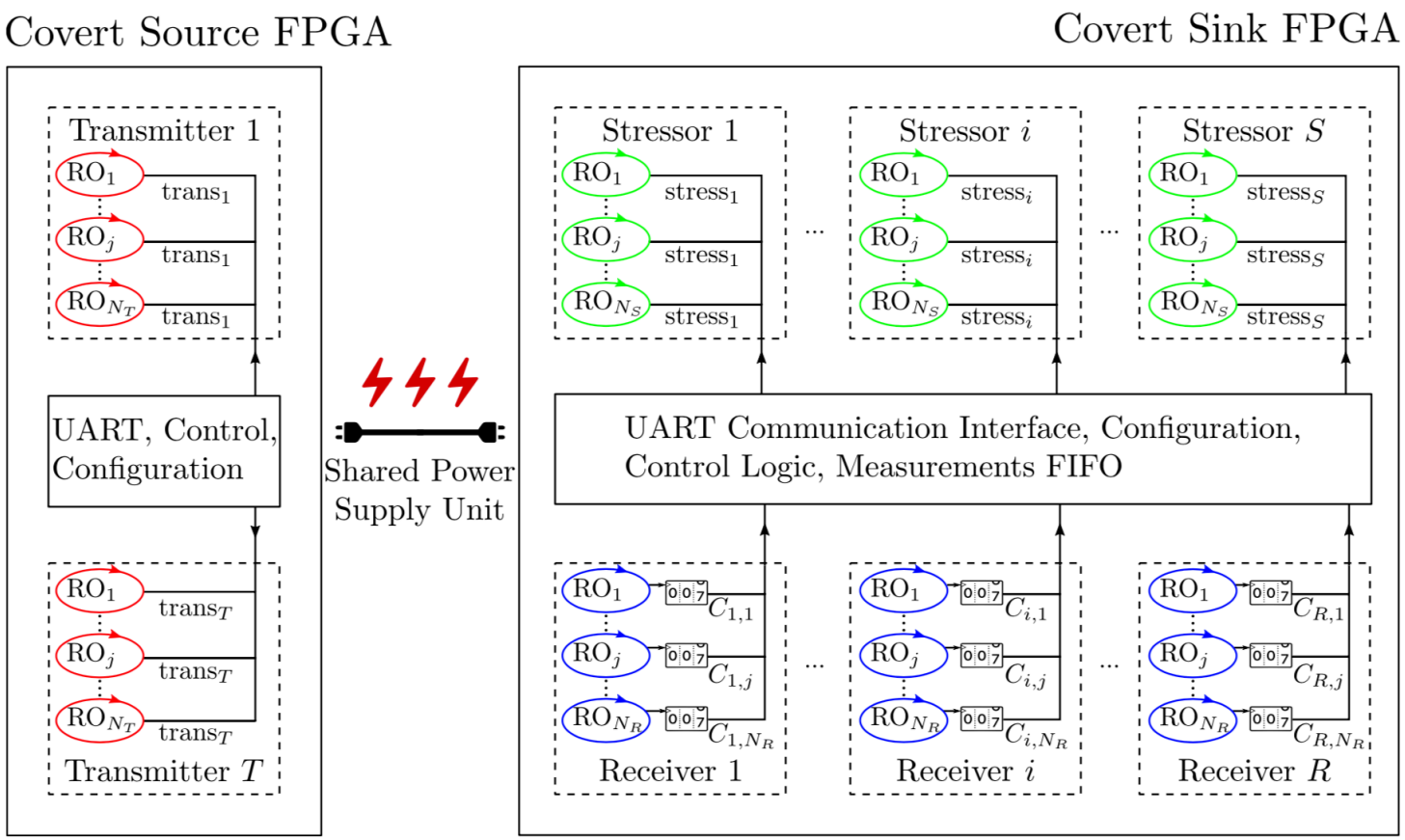
- FPGAs in Cloud FPGAs share much of the infrastructure
  - Share power supply within server
  - Share PCIe bus within server
  - Servers share power, cooling, networking, etc. with other servers

- We present a new type of information through a shared power supply



Cross-FPGA Communication Using Power Supplies

Sender FPGA uses array of ROs to consume a lot of power, and stress the power supply

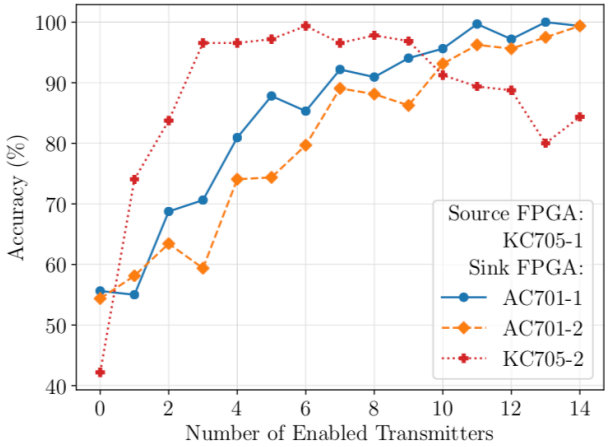


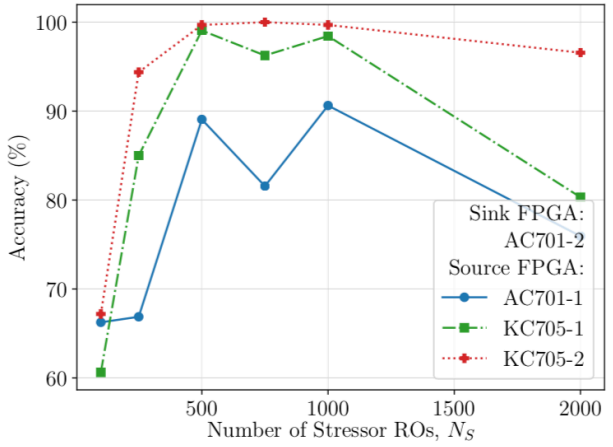
Stressor ROs needed to cause local voltage regulator to be over stressed

Receiver FPGA has both stressor ROs and measurement ROs

Analyzing Cross-FPGA Data Transmission Parameters

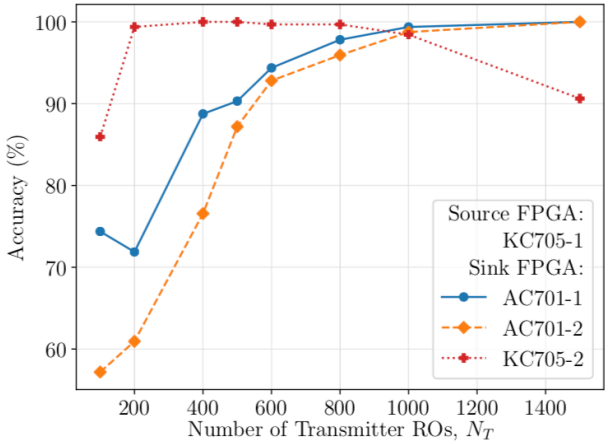
Accuracy vs. Number of Transmitters

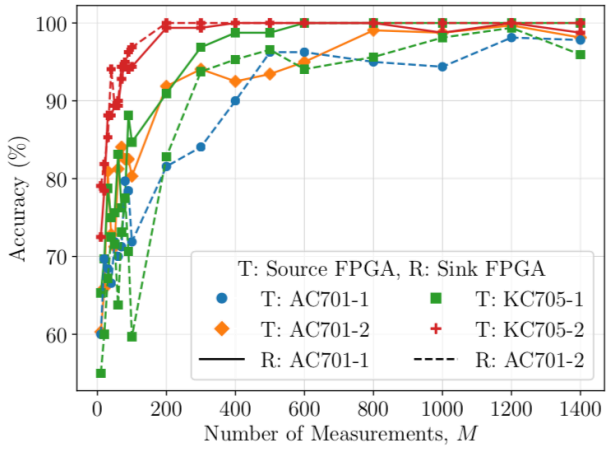




Accuracy vs. Number of Stressors (more is not always better)

Accuracy vs. Number of Transmitter ROs





Accuracy vs. Number of Measurements

Cross-FPGA Data Transmission

- Using the presented design, communication between different Artix and Kintex FPGAs was achieved
  - Few bits per second
  - High reliability
  - Slow, but reliable channel for leaking cryptographic keys, for example

Property	Artix 7	Kintex 7
Transmitter ROs, N_T	1,000	1,000
Enabled Transmitters	10	14
Transmitted Pattern	`0xf3ed1`	`0xf3ed1`
Transmitter Types	LUT-RO	LUT-RO
Stressor ROs, N_S	500	500
Enabled Stressors	1	5
Stressor & Receiver Types	LUT-RO	LUT-RO
Measurement Cycles, 2^t	2^{15}	2^{21}
Repetitions per Bit, M	500	500
Channel Bandwidth b (bps)	6.1	0.1

PSU	↓ T → R	AC701-1	AC701-2	KC705-1	KC705-2
A	AC701-1	-	79%	92%	100%
A	AC701-2	99%	-	93%	100%
A	KC705-1	100%	86%	-	100%
A	KC705-2	100%	98%	99%	-
B	AC701-1	-	100%	98%	100%
B	AC701-2	100%	-	99%	100%
B	KC705-1	100%	95%	-	100%
B	KC705-2	100%	100%	98%	-

Cloud FPGA Fingerprinting

Fingerprinting Cloud FPGAs

Shanquan Tian, Wenjie Xiong, Ilias Giechaskiel, Kasper Rasmussen, and Jakub Szefer, "Fingerprinting Cloud FPGA Infrastructures", in Proceedings of the International Symposium on Field-Programmable Gate Arrays (FPGA), February 2020.

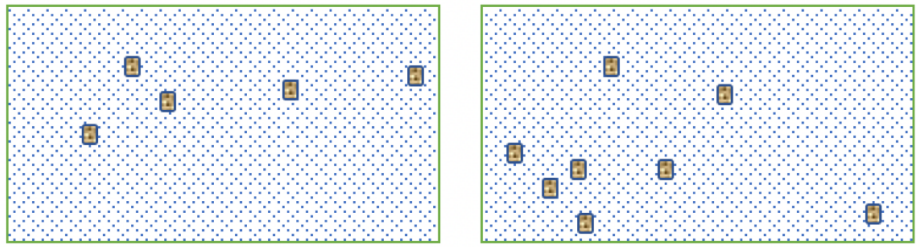
- Being able to identify FPGA instances can allow for improved security, but also for potential new attacks
  - Currently FPGAs in the cloud do not expose IDs or serial numbers
  - Fingerprinting using Physically Uncloneable Functions (PUFs) can be used to identify the FPGAs

- Improve security: identify FPGAs to ensure different ones are used for reliability or fault tolerance
- Potential attacks: fingerprint whole Cloud FPGA infrastructures

- Fingerprinting can be done using FPGA on-chip resources, or other components of the FPGA board, such as DRAM modules

DRAM PUFs leverage decay of DRAM cells to uniquely identify the DRAM module

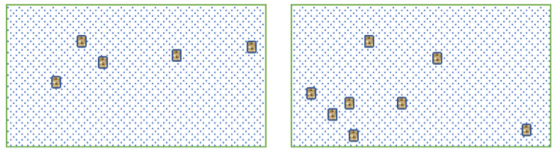
Each DRAM has unique pattern, stable over time, that can be adjusted for temperature changes

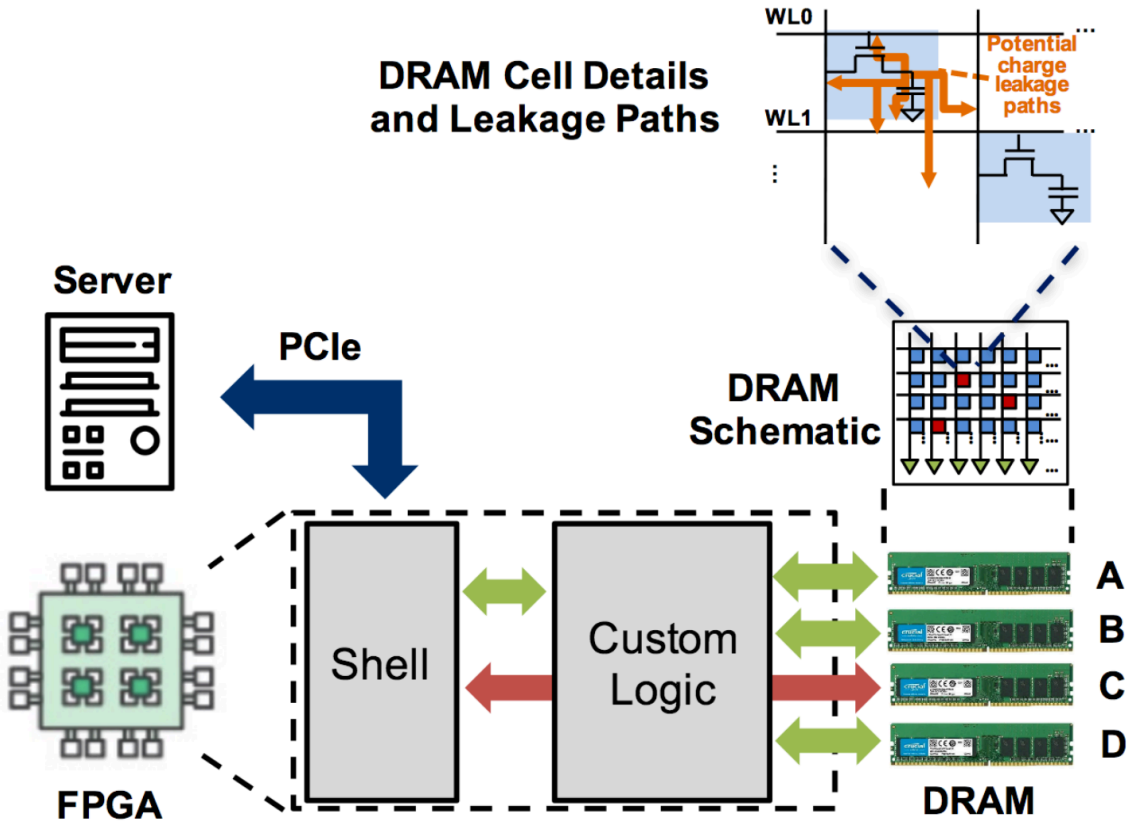


Can leverage DRAM modules in Cloud FPGAs to fingerprint the FPGA instances!

DRAM Modules in Cloud FPGAs

- Each FPGA board is populated with multiple DRAM modules
- Can assume that in most cases, the DRAMs are not physically moved between FPGA boards
  - **Fingerprinting DRAM module equals fingerprinting an FPGA board**
- Fingerprinting can be done using DRAM PUFs:
  1. Charge DRAM cells (capacitors)
  2. Let DRAM cells decay
  3. Read back DRAM to see which cells decayed

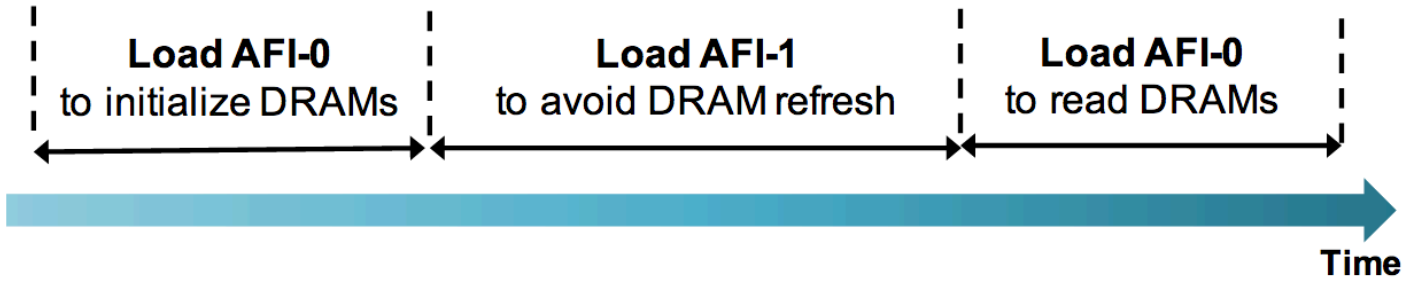




Fingerprinting Approach

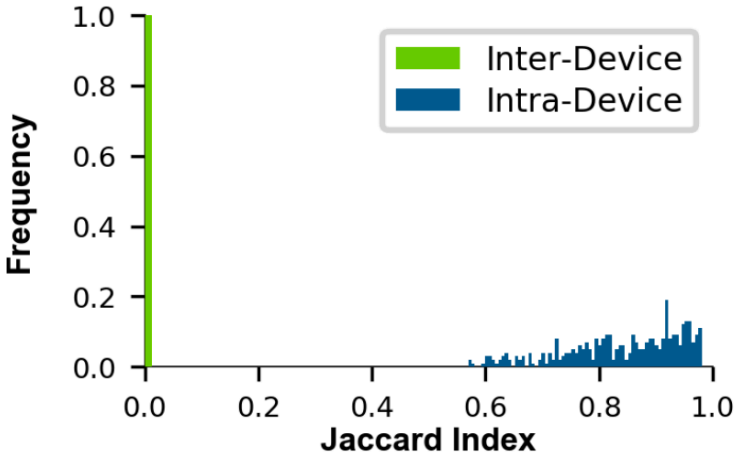
- Simply disabling DRAM refresh is not possible in Cloud FPGAs
  - One DRAM (DRAM C) fully controlled by the *shell*
  - Other DRAMs controlled by users, but DRAM controller is encrypted IP from Xilinx

- Need a work-around to disable DRAM refresh: use two AFIs (bitstreams) with and without DRAM controller:

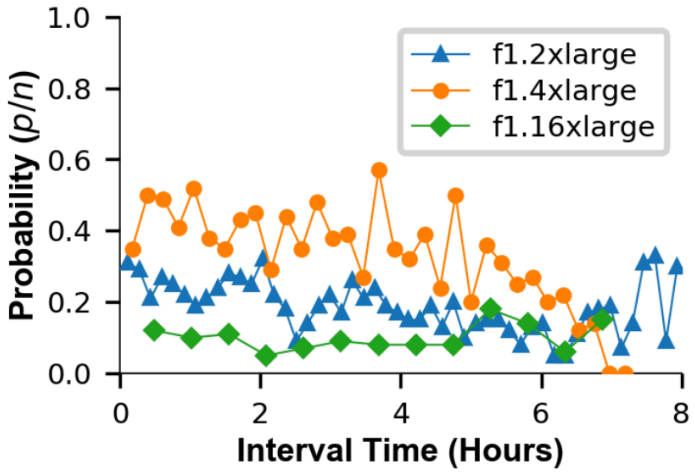


Some Fingerprinting Results

- Fingerprinting can reliably distinguish FPGA instances based on the DRAM PUFs
  - Inter- and intra-device Jaccard Index shows clear separation of the fingerprints
  - Use three DRAMs to further increase accuracy



- Can learn probability of getting same FPGA instance over time



- *Other interesting insights are in our paper!*

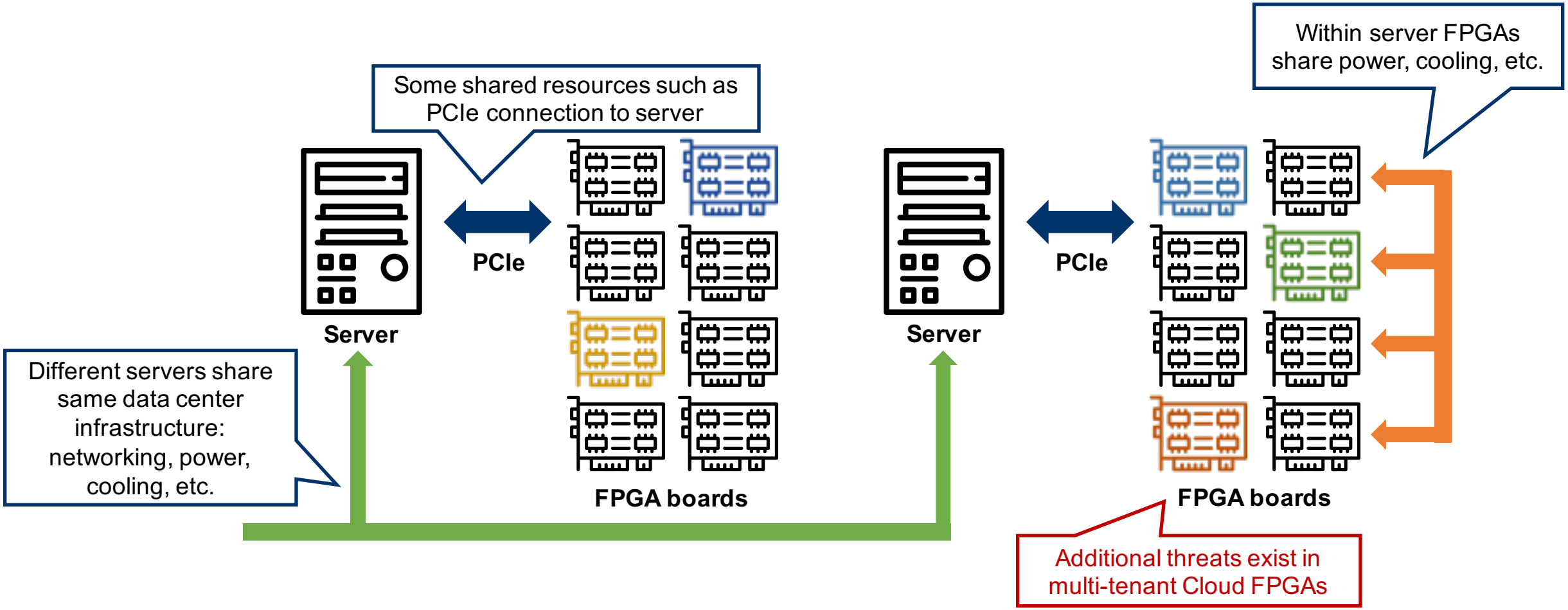
Security Challenges in Single-Tenant Cloud FPGAs

Security Challenges

- Users could load potential malicious FPGA bitstreams (AFI) to attack or leak information in Cloud FPGAs, even without multiple users sharing the same FPGA:
  - Leak information from one FPGA to another (covert channel)
  - Steal information from another FPGA (side channel)
  - Steal information from the shell (side channel)
  - Reverse engineer Cloud FPGA infrastructure
  - Induce faults, waste power, waste resources (e.g. generate PCI traffic that blocks others form accessing FPGA)
  - For all, use: thermal, cross-talk, or power attacks

- Recent research has focused on attacks – more defenses need to be deployed to prevent these from happening in practice
- A new type of security threat, compared to CPUs or GPUs in the cloud

Security Challenges: Many Side Channels Possible



Thank You!

https://caslab.csl.yale.edu/