Safe and Secure Real-Time Computing Infrastructure for Intelligent Cyber-Physical Systems

Heechul Yun

Associate Professor, University of Kansas

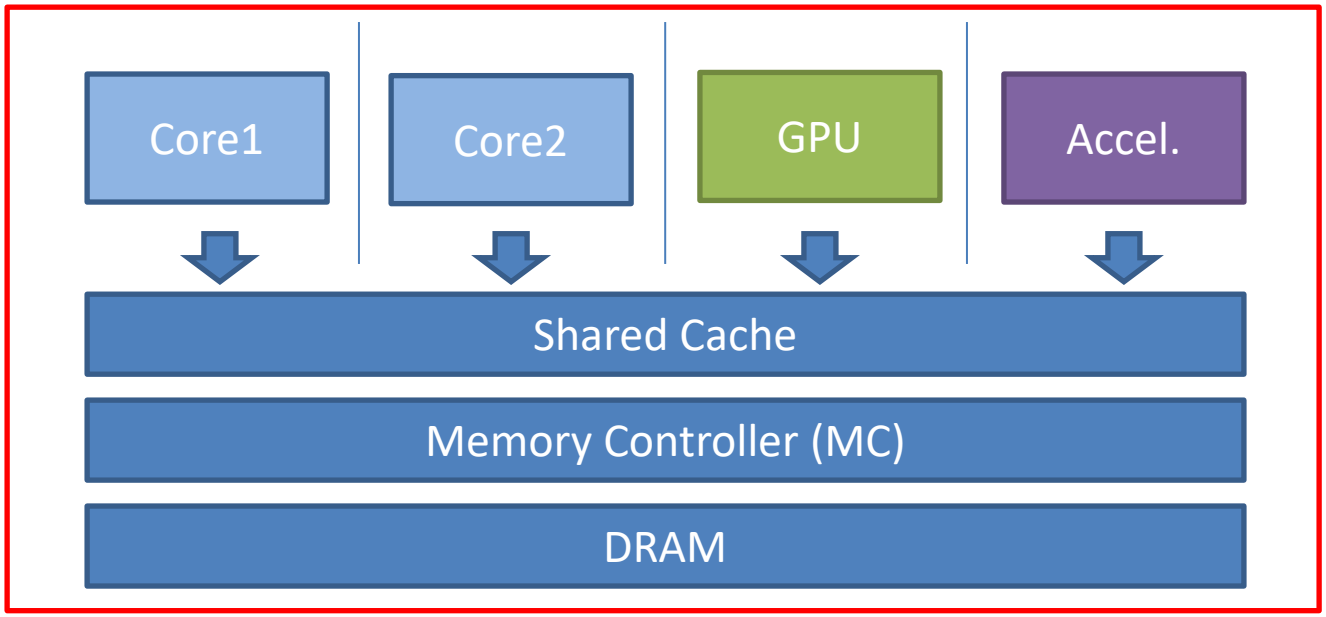
March, 2021

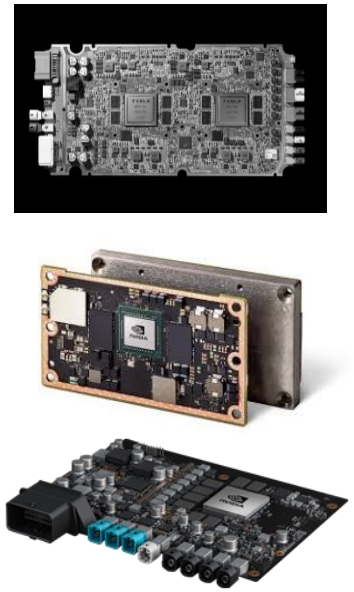
KU THE UNIVERSITY OF KANSAS

Modern Cyber-Physical Systems

- Cyber Physical Systems (CPS)
  - Cyber (Computer) + Physical (Plant)
- Real-time
  - Control physical process in real-time
- Safety-critical
  - Can harm people/things
- **Intelligent**
  - Can function autonomously

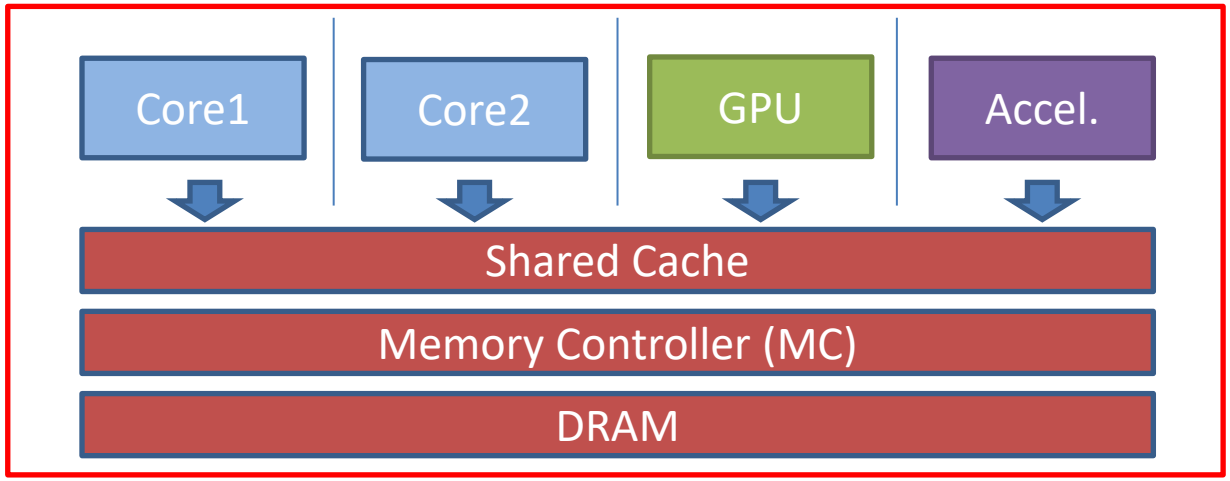
Modern System-on-Chip (SoC)





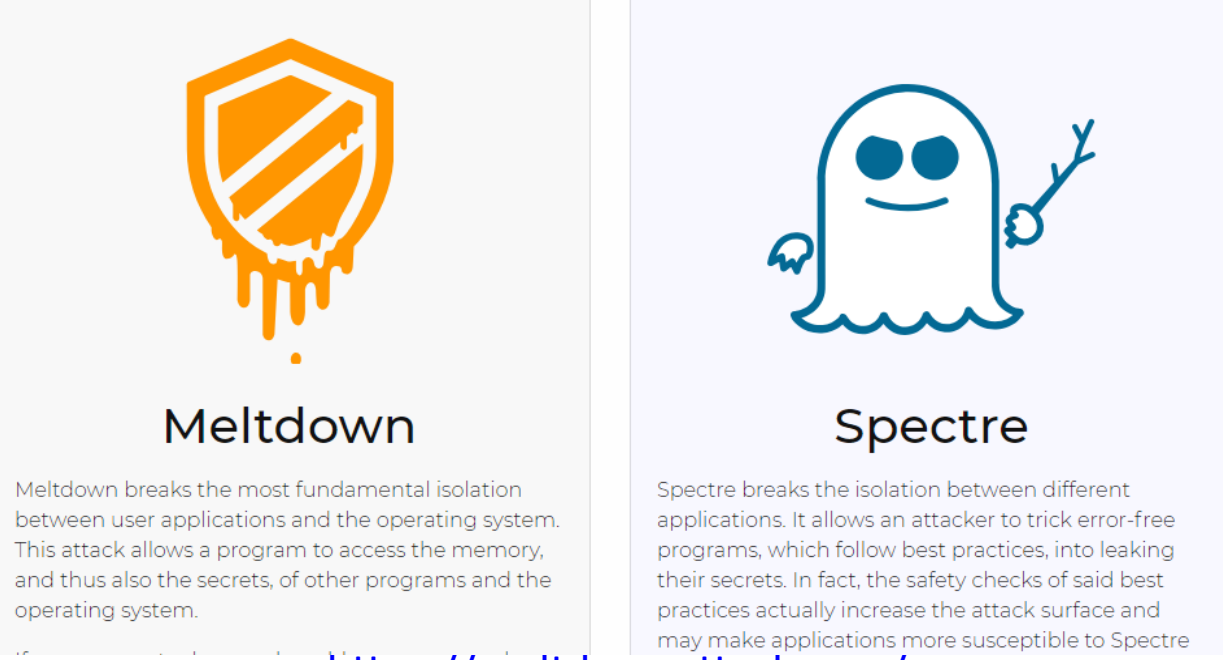
- Integrate multiple cores, GPU, accelerators
- Good performance, cost, size, weight, power
- **But …**

Time Predictability and Safety



- Many important hardware resources are *shared*
- Difficult to guarantee predictable timing
- A **safety problem** in CPS

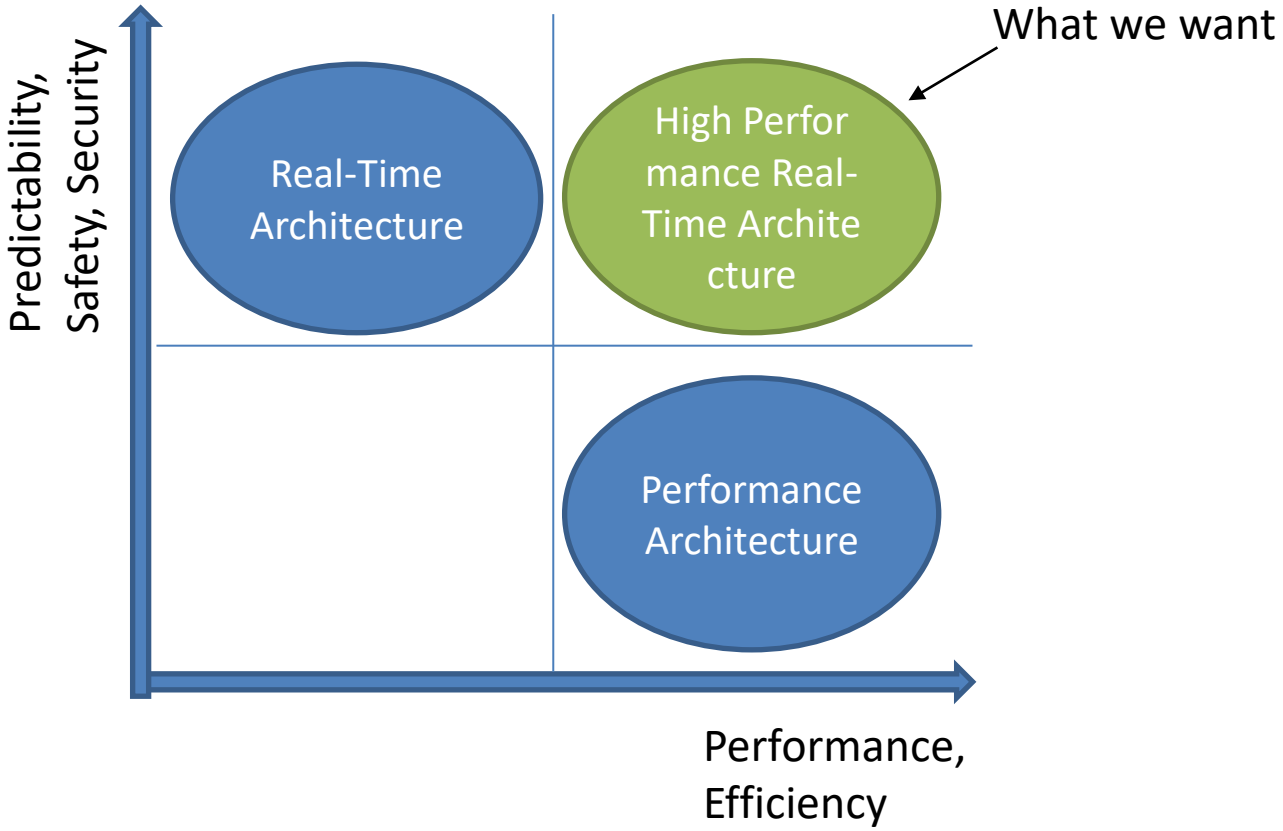
Timing Channels and Security



https://meltdownattack.com/

- Measurable timing differences in accessing *shared* hardware resources can leak secret: a **security problem**

Computing Infrastructure for CPS



- Predictable, secure, and high-performance computing

Our Research

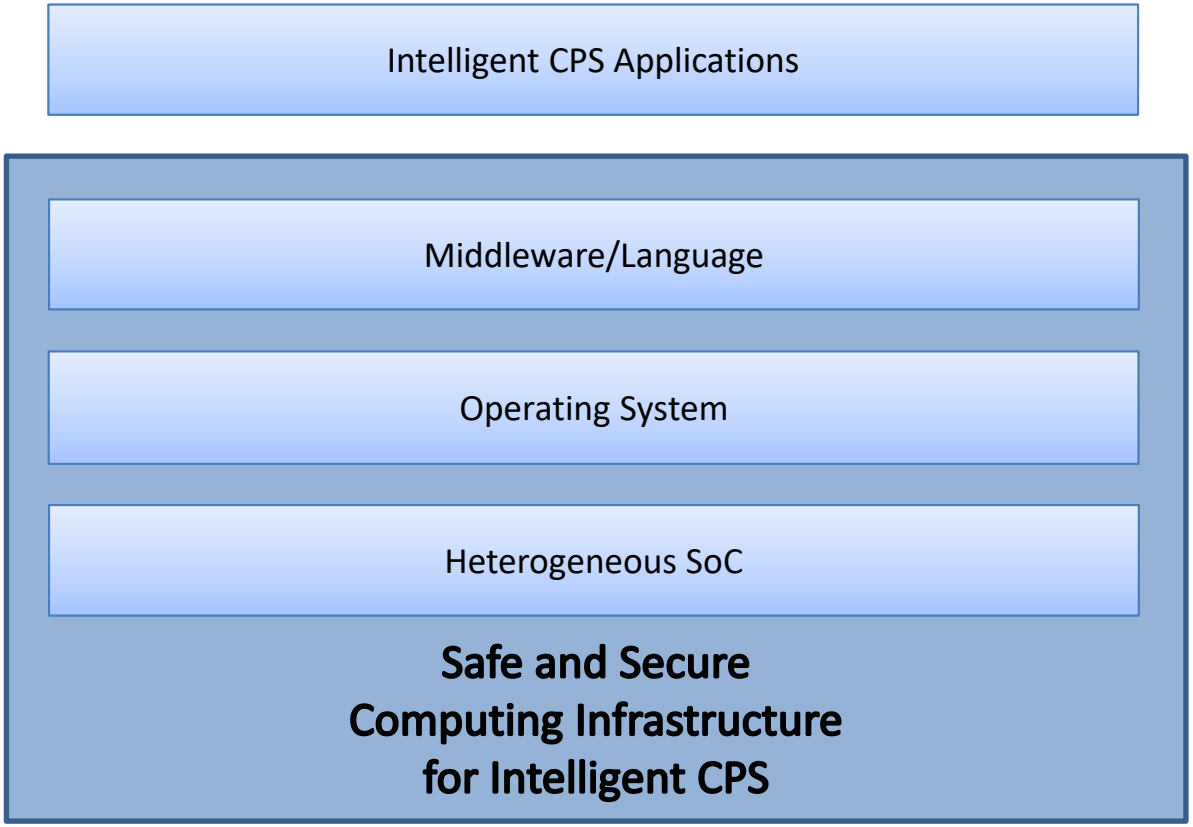
- Build **safe and secure real-time computing infrastructure** for the next generation of intelligent Cyber Physical Systems (CPS).

Research Sponsors

Equipment Sponsors

KU THE UNIVERSITY OF KANSAS

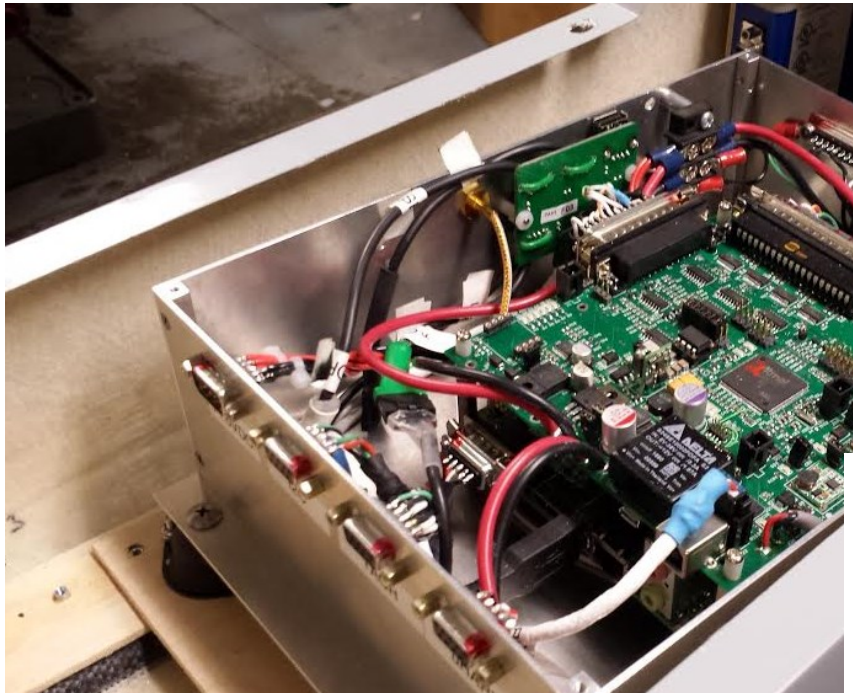
Our Approach

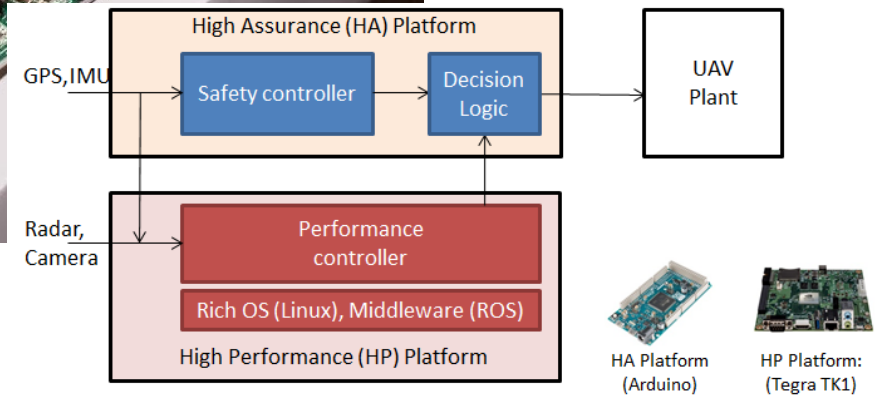


- Holistic, cross-layer approach for safe and secure CPS

KU THE UNIVERSITY OF KANSAS

KU Fixed-wing UAV





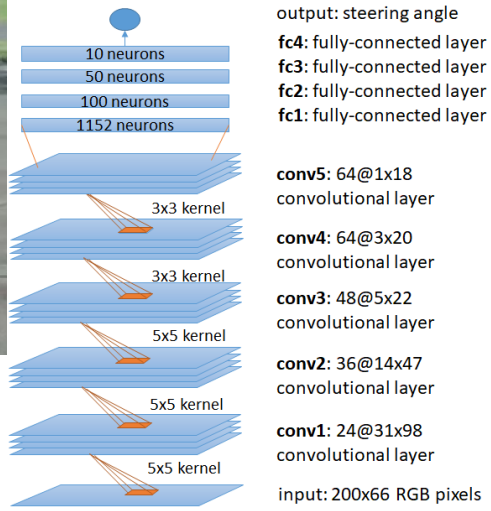
Safe and Intelligent UAV architecture

(*) Prasanth Vivekanandan, Gonzalo Garcia, Heechul Yun, Shawn Keshmiri. A Simplex Architecture for Intelligent and Safe Unmanned Aerial Vehicles. In *RTCSA*, IEEE, 2016. **Best Student Paper Nominee**

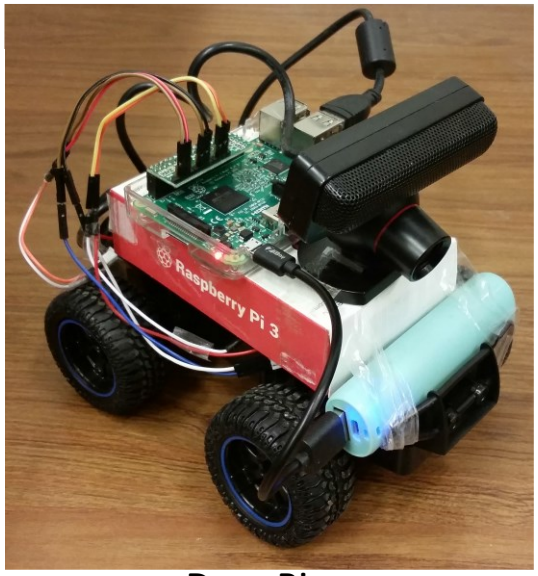
DeepPicar

- End-to-end deep learning: *pixels* to *steering*
- Using **identical** DNN with NVIDIA's DAVE-2

DAVE-2



Same DNN

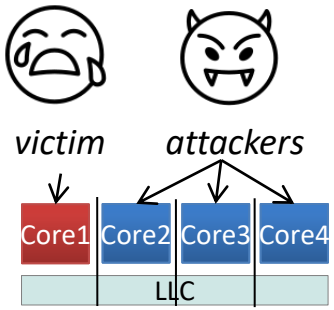


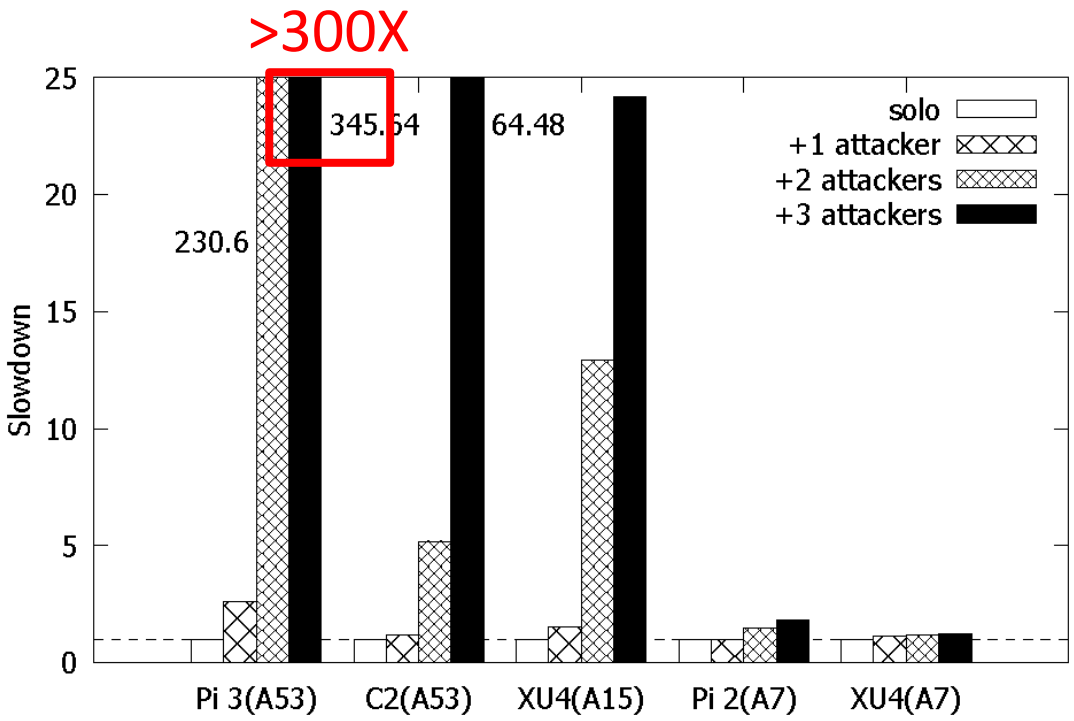
DeepPicar

DNN based real-time control in embedded multicore CPU

Michael G. Bechtel, Elise McEllhiney, Minje Kim, Heechul Yun. "DeepPicar: A Low-cost Deep Neural Network-based Autonomous Car." In *RTCSA*, 2018.

Cache Denial-of-Service Attacks



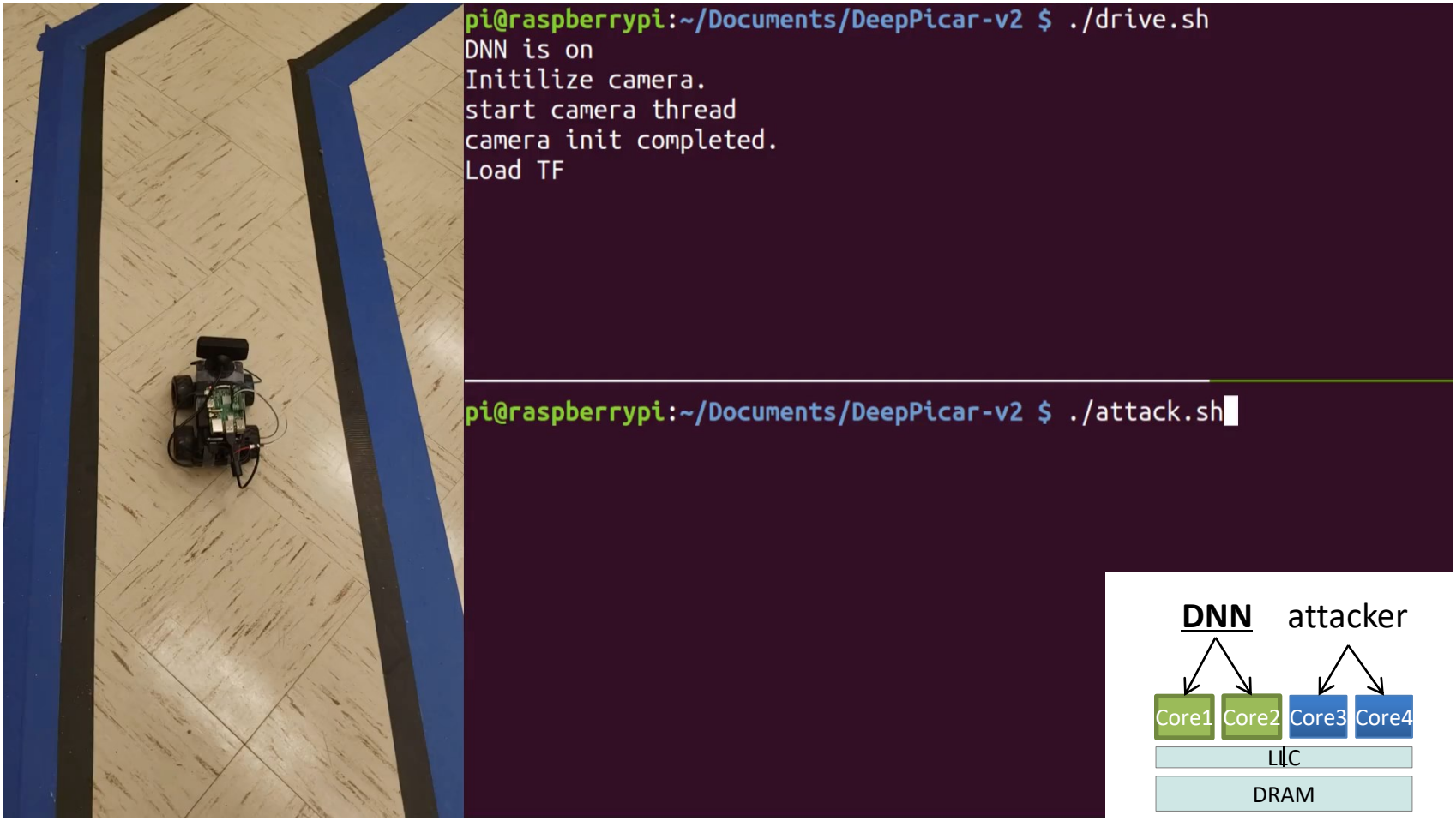


- Observed worst-case: >300X (times) slowdown
  - On popular in-order multicore processors

Found serious DoS attack vulnerability in COTS multicore

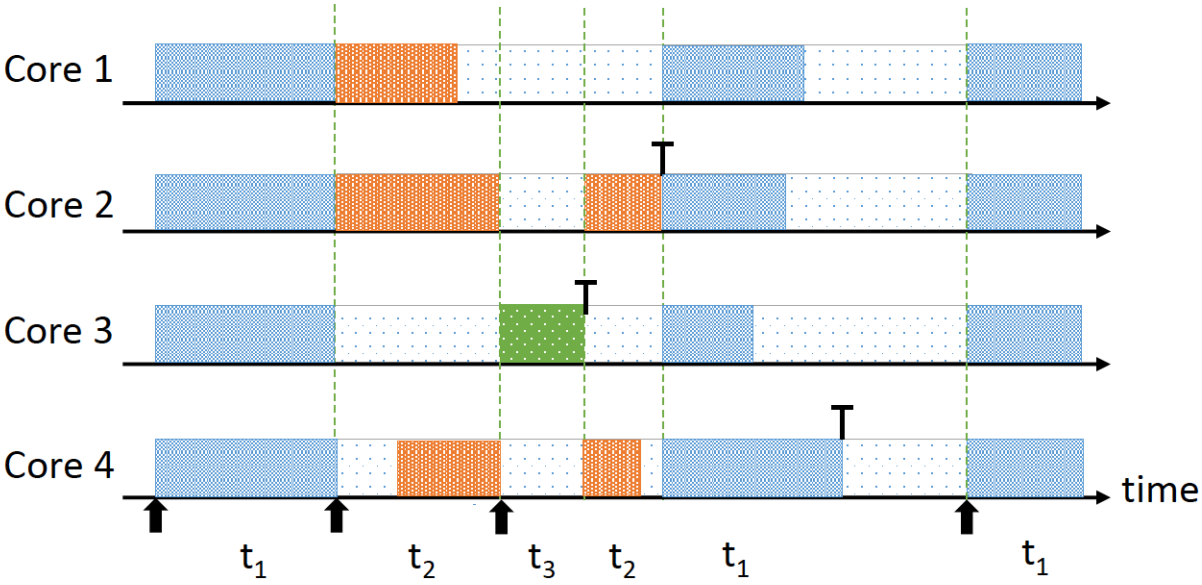
M. G. Bechtel and H. Yun. "Denial-of-Service Attacks on Shared Cache in Multicore: Analysis and Prevention." In *RTAS*, 2019
Outstanding Paper Award

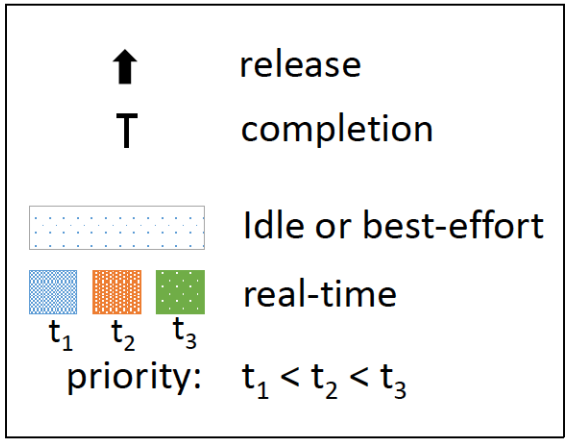
Effect of Cache DoS Attack



https://youtu.be/Jm6KSDqlqiU

RT-Gang



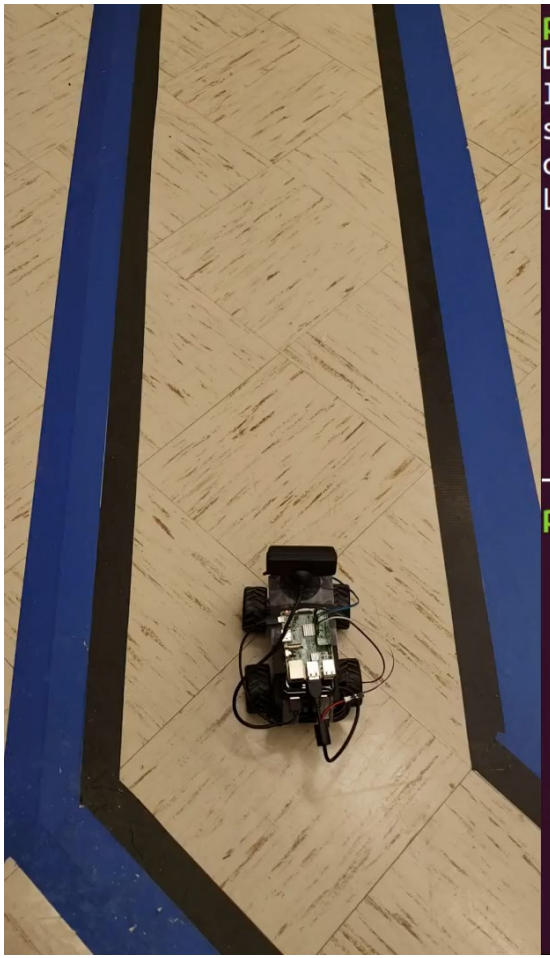


- **One parallel real-time task---a gang---at a time**
  - Eliminate inter-task interference by construction
- Schedule best-effort tasks during slacks **w/ throttling**

OS solution for predictable real-time computing on COTS

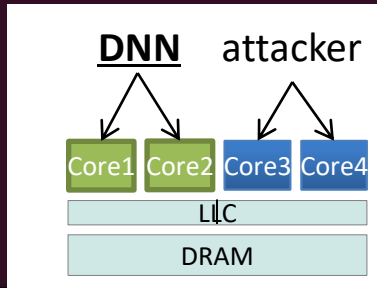
W. Ali and H. Yun. "RT-Gang: Real-Time Gang Scheduling Framework for Safety-Critical Systems." In *RTAS*, 2019

Effect of RT-Gang



```
pi@raspberrypi:~/Documents/DeepPicar-v2 $ ./drive.sh
DNN is on
Initilize camera.
start camera thread
camera init completed.
Load TF
```

```
pi@raspberrypi:~/Documents/DeepPicar-v2 $ ./attack.sh
```

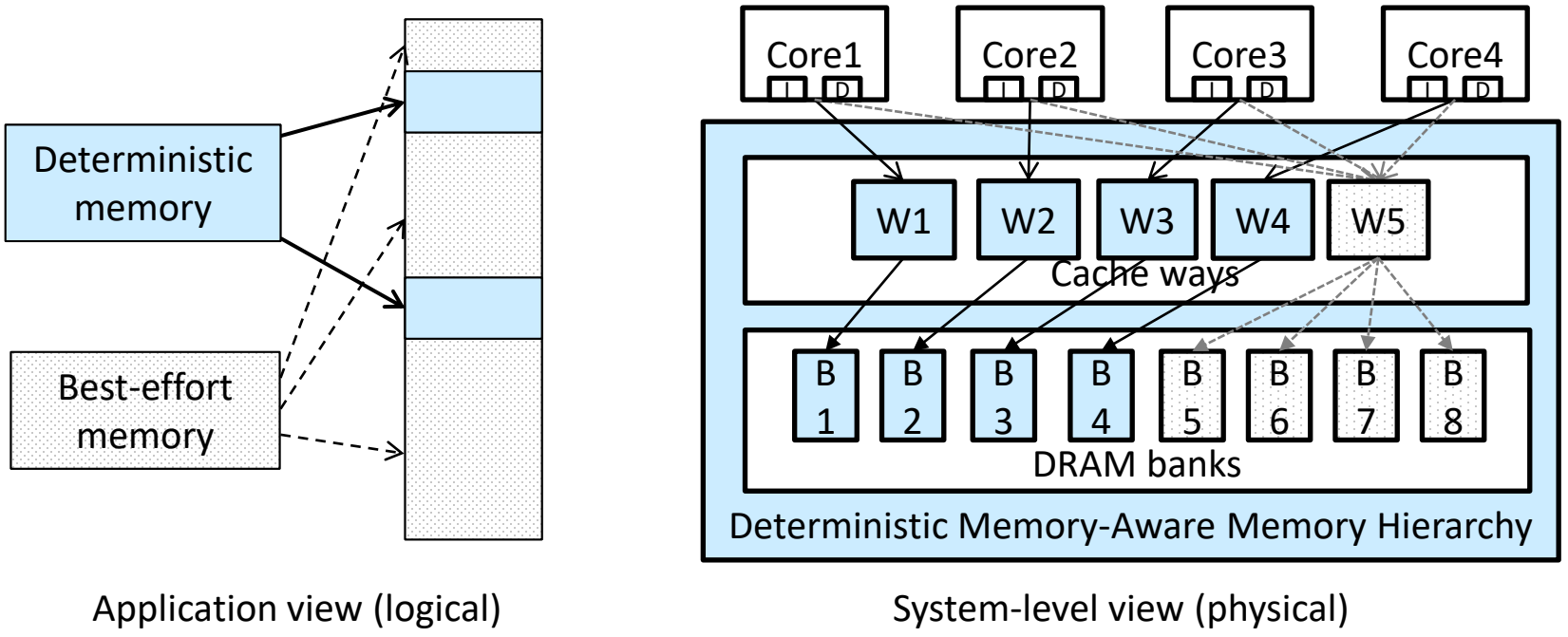


https://youtu.be/pk0j063cUAs

W. Ali, M. G. Bechtel and H. Yun. "Analyzable and Practical Real-Time Gang Scheduling on Multicore Using RT-Gang." In *OSPERT*, 2019

Deterministic Memory

- Declare all or part of address space as deterministic memory
- DM-aware **end-to-end resource management**

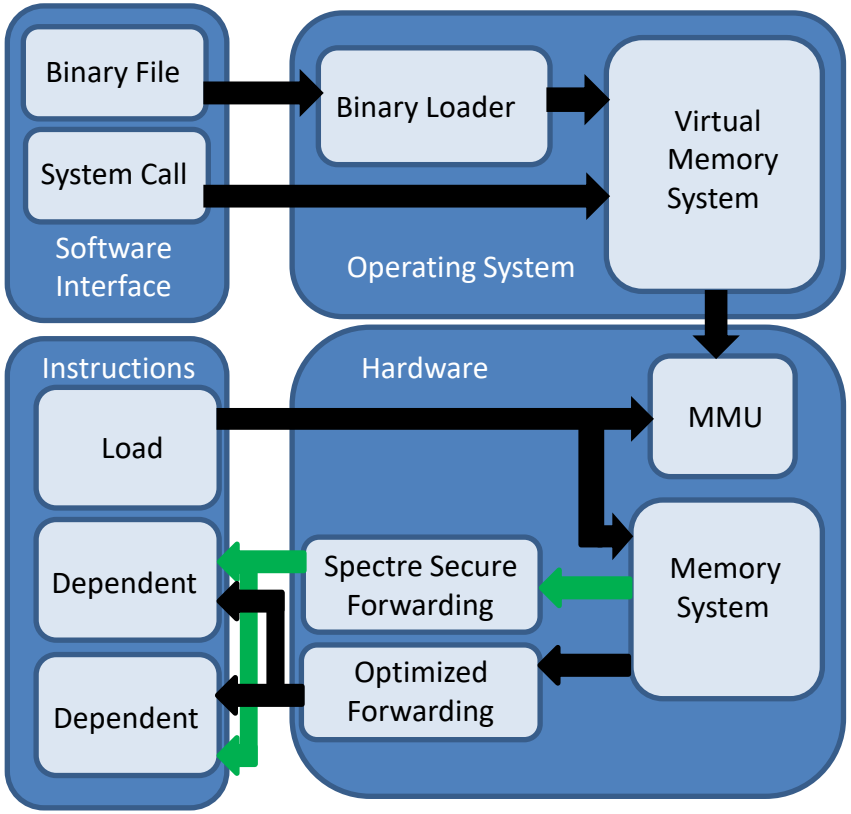


Data-centric cross-layer approach for **real-time**

F. Farshchi, P. K. Valsan, H. Yun. "Deterministic memory abstraction and supporting multicore system architecture." In *ECRTS*, 2018

SpectreGuard

- Step 1: Software tells OS what **data** is secret
- Step 2: OS updates the page table entries
- Step 3: Load of the secret data is identified by MMU
- Step 4: secret data forwarding is **delayed** until safe

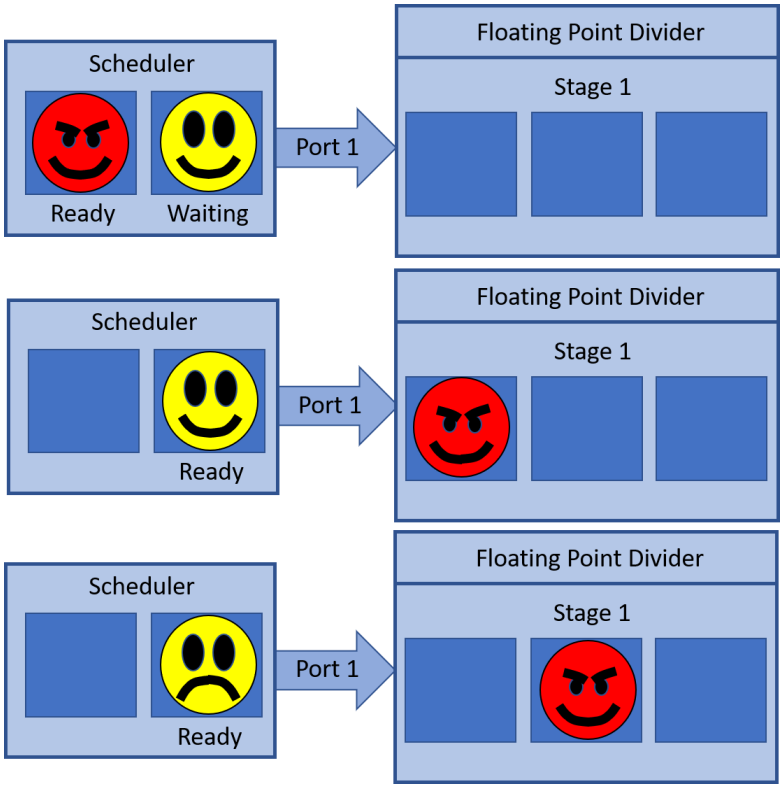


Data-centric cross-layer approach for **security**

J. Fustos, F. Farshchi, H. Yun. "SpectreGuard: An Efficient Data-centric Defense Mechanism against Spectre Attacks.." In *DAC*, 2019

SpectreRewind

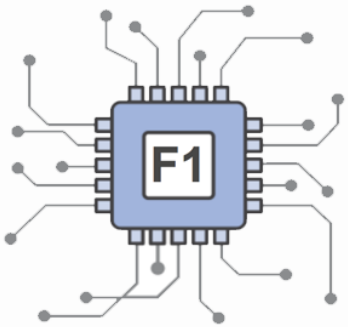
- New covert channel for Spectre-like speculative execution attacks
- Exploit contention on non-pipelined functional units
- Leak secret to past instructions, bypassing mitigation techniques
- High performance, low noise



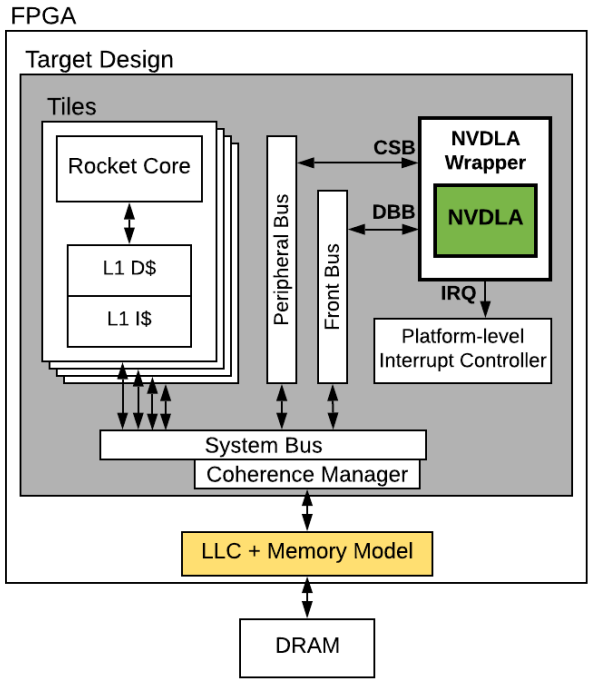
Discovered new speculative contention-based attacks

J. Fustos, M. Bechtel and H. Yun. "SpectreRewind: Leaking Secrets to Past Instructions." In *ASHES*, 2020

RISC-V + NVDLA SoC Platform



amazon web services



- Full-featured quad-core SoC with hardware DNN accelerator on Amazon FPGA cloud
  - Run Linux, YOLO v3 object detection

Open-source hardware for research

F. Farshchi, Q Huang, H. Yun. "Integrating NVIDIA Deep Learning Accelerator (NVDLA) with RISC-V SoC on FireSim." In *EMC^2*, 2019

https://devblogs.nvidia.com/nvdla/

NVIDIA. DEVELOPER

Join Login

NVIDIA Developer Blog

AI / DEEP LEARNING AUTONOMOUS MACHINES AUTONOMOUS VEHICLES DATA SCIENCE FEATURES GRAPHICS / SIMULATION HPC SMART CITIES

AI / DEEP LEARNING

NVDLA Deep Learning Inference Compiler is Now Open Source

By Rekha Mukund, Prashant Gaikwad and Mitch Harwell | September 11, 2019

Tags: compilers, Inference, Jetson, Machine Learning and AI, Object Detection

Designing new custom hardware ac
performance and efficiency with a n

Two years ago, NVIDIA opened the s
help advance the adoption of efficie
NVIDIA Jetson AGX Xavier Develope
open-source release of NVDLA's opt
point with the complete source for t

In this blog we'll explain the role tha
purpose-built hardware accelerator
NVDLA software and hardware desi

"

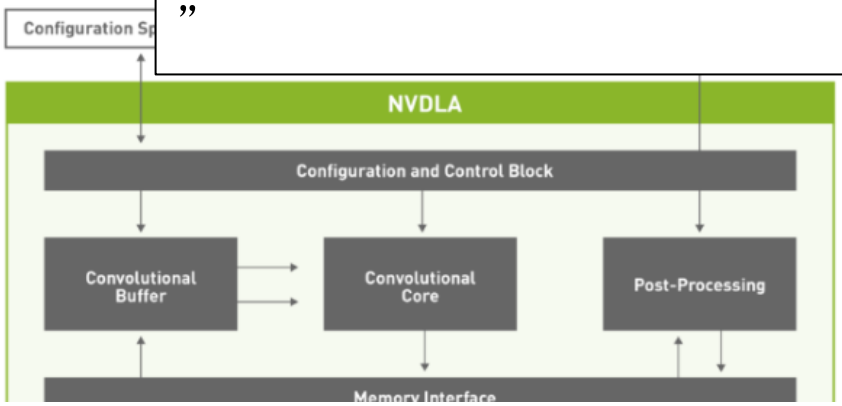
One of the best ways to get started is to dive right in with object detection using YOLOv3 on NVDLA with RISC-V and FireSim in the cloud.

...

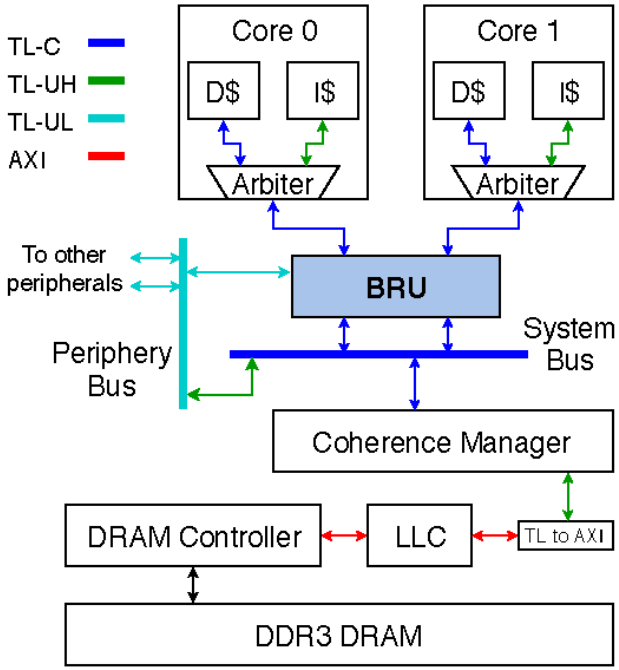
git clone https://github.com/CSL-KU/firesim-nvdla

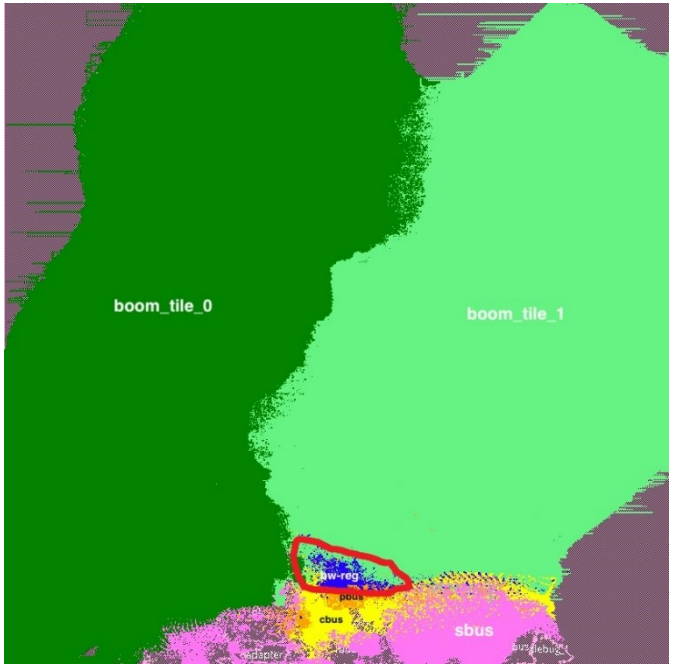
"

Star 108



Bandwidth Regulation Unit (BRU)





- Regulate per-core/group memory bandwidth
  - Drop-in addition to existing processor design

Real hardware-based resource control mechanism

F. Farshchi, Q. Huang, and H. Yun. BRU: Bandwidth Regulation Unit for Real-Time Multicore Processors, in *RTAS*, 2020.

Conclusion

- Intelligent cyber-physical systems need safe and secure computing infrastructure.
- Today's computing infrastructure is inadequate for safe and secure CPS
- Our research develops fundamental computing infrastructure technologies to enable safe and secure computing for intelligent CPS

KU THE UNIVERSITY OF KANSAS

Recent Publications

1. [**DATE'21**] Waqar Ali, Rodolfo Pellizzoni, Heechul Yun. Virtual Gang based Scheduling of Parallel Real-Time Tasks (to appear)
2. [**RTSS'20**] Shengzhong Liu, Shuochao Yao, Xinzhe Fu, Rohan Tabish, Simon Yu, Ayoosh Bansal, Heechul Yun, Lui Sha and Tarek Abdelzaher. On Removing Algorithmic Priority Inversion from Mission-critical Machine Inference Pipelines.
   Best Paper Award
3. [**ASHES'20**] Jacob Fustos, Michael Garrett Bechtel, Heechul Yun. SpectreRewind: Leaking Secrets to Past Instructions.
4. [**EMSOFT'20**] Homa Aghilinasab, Waqar Ali, Heechul Yun, Rodolfo Pellizzoni. Dynamic Memory Bandwidth Allocation for Real-Time GPU-Based SoC Platforms.
5. [**RTAS'20**] Farzad Farshchi, Qijing Huang, and Heechul Yun. BRU: Bandwidth Regulation Unit for Real-Time Multicore Processors.
6. [**DAC'19**] Jacob Fustos, Farzad Farshchi, and Heechul Yun. SpectreGuard: An Efficient Data-centric Defense Mechanism against Spectre Attacks.
7. [**ECRTS'19**] Renato Mancuso, Heechul Yun, Isabelle Puaut. Impact of DM-LRU on WCET: a Static Analysis Approach.
8. [**RTAS'19-2**] Waqar Ali and Heechul Yun. RT-Gang: Real-Time Gang Scheduling Framework for Safety-Critical Systems.
9. [**RTAS'19-1**] Michael Bechtel and Heechul Yun. Denial-of-Service Attacks on Shared Cache in Multicore: Analysis and Prevention.
   Outstanding Paper Award
10. [**EMC2'19**] Farzad Farshchi, Qijing Huang, and Heechul Yun. Integrating NVIDIA Deep Learning Accelerator (NVDLA) with RISC-V SoC on FireSim.
11. [**RTCSA'18**] Michael Bechtel, Elise McEllhiney, Minje Kim, Heechul Yun. DeepPicar: A Low-cost Deep Neural Network-based Autonomous Car.
12. [**ECRTS'18-2**] Waqar Ali, Heechul Yun. Protecting Real-Time GPU Applications on Integrated CPU-GPU SoC Platforms.
13. [**ECRTS'18-1**] Farzad Farshchi, Prathap Kumar Valsan, Renato Mancuso, Heechul Yun. Deterministic Memory Abstraction and Supporting Multicore System Architecture.
14. [**RTSJ'17**] Prathap Valsan, Heechul Yun, Farzad Farshchi. Addressing Isolation Challenges of Non-blocking Caches for Multicore Real-Time Systems.
15. [**TC'17**] Heechul Yun, Waqar Ali, Santosh Gondi, Siddhartha Biswas. BWLOCK: A Dynamic Memory Access Control Framework for Soft Real-Time Applications on Multicore Platforms.
16. [**RTCSA'16**] Prasanth Vivekanandan, Gonzalo Garcia, Heechul Yun, Shawn Keshmiri. A Simplex Architecture for Intelligent and Safe Unmanned Aerial Vehicles.
    Best Student Paper Nomination
17. [**RTAS'16**] Prathap Valsan, Heechul Yun, Farzad Farshchi . Taming Non-blocking Caches to Improve Isolation in Multicore Real-Time Systems.
    Best Paper Award
18. [**TC'16**] Heechul Yun, Gang Yao, Rodolfo Pellizzoni, Marco Caccamo, and Lui Sha. Memory Bandwidth Management for Efficient Performance Isolation in Multi-core Platforms.
    Editor's Pick of the Year

KU THE UNIVERSITY OF KANSAS

Full List: http://www.ittc.ku.edu/~heechul/pub.html

Thank You!

Acknowledgement:

This research has been supported in part by NSA Science of Security initiative contract #H98230-18-D-0009; NSF CNS 1718880, 1815959; NASA SBIR