CPS: Medium: GOALI: Enabling Scalable Real-Time Certification for AI-Oriented Safety-Critical Systems

PI: Jim Anderson (UNC). Co-PIs: F. Don Smith (UNC), Ron Alterovitz (UNC), and Prakash Sarathy (Northrop Grumman)

Students: Tanya Amert, Angelos Angelopoulos, Joshua Bakita, Sandeep Kumar, Zelin Tong, and Sergey Voronov

https://www.cs.unc.edu/~anderson/projects/rtai.html

Motivation

There is an evolution looming towards highly intelligent systems in our everyday lives, and some of the most compelling use cases—such as autonomous aircraft and automobiles—fall within safety-critical domains for which certification is essential.

The AI-oriented workloads that must be supported are very complex, and the hardware needed to support them—typically multicore machines augmented with accelerators—is much more complex as well.

How should certification processes in safety-critical domains evolve to address these complexities? Can these processes be made to scale to large systems that may be subject to partial redesigns during their lifetime?

Problem

Research Overview: Can a time and space partitioning solution be found for AI-oriented avionics workloads hosted on multicore+accelerator hardware that enables the real-time correctness of an overall system to be certified with high confidence while making efficient usage of hardware resources?

Our Solution: We decompose AI-oriented workloads into manageable components that are isolated from one another, so that intra-component timing constraints can be verified independently.

Objectives

- Isolate system components despite non-preemptive accelerator accesses.
- Provide a response-time analysis for isolated components with accelerator accesses.
- Enable many components on memory-scarce multicore+ accelerator platforms.
- Create methods for component-wise AI.

Activities

- Summer internships at Northrop Grumman, 2021, 2022.
- Bi-weekly meetings with DoD partners on evolving avionics certification.
- Outreach efforts to obtain industry perspectives from key players (e.g., Bosch, Apex.AI).

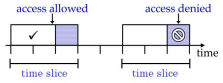
Cross-Component Temporal Isolation with TimeWall

Goal: Isolate system components despite non-preemptive accelerator accesses.

We introduce TimeWall, which enforces temporal isolation by assigning time slices to each component. During a timeslice, a component has exclusive access to all its assigned system resources.

To ensure non-preemptive GPU accesses do not cross time-slice boundaries, TimeWall utilizes forbidden zones. Accesses that cannot complete by the end of a time slice will be delayed until the component's next assigned time slice.

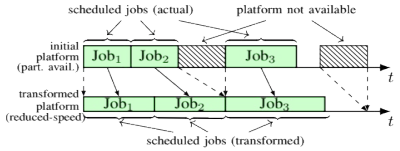
To prevent anomalies where a GPU access begins long after it is requested, TimeWall preempts all jobs requesting GPU access in the forbidden zone.



Criticality-Aware Timing Analysis for Multicore+Accelerator

Goal: Provide a response-time analysis for isolated components with accelerator accesses, while allowing jobs of the same task to be scheduled in parallel.

Existing response-time analysis is inapplicable, as component isolation makes the computing platform partially available.



We solve this via a new framework of schedule and platform transformations which allows derivation of the response-time bound of the partially available system from that of the transformed one. This allows most existing analysis to be reused.

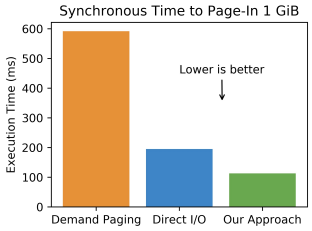
Our system supports non-preemptive accelerator accesses via a locking protocol, and led to findings that partitioning improves the response time of "heavy" components.

Enabling DRAM Oversubscription Among Components

Goal: Enable many components on memory-scarce multicore+ accelerator platforms.

We develop a low-overhead approach to DRAM oversubscription by leveraging developments in common-off-the-self SSDs and embedded multicore+GPU platforms.

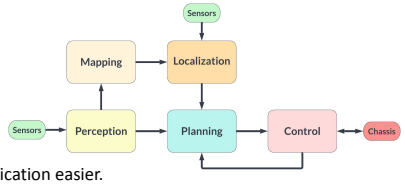
Implementation is 4x faster than Linux's default, with 80% less overhead.



Component-Wise AI

Goal: Refactor the AI software of an autonomous system (e.g., drone) into components amenable to real-time certification.

All components of an autonomous system can use AI and there are complex dependencies. There is major competition for resources and scheduling is hard. Altering the components into ones that are simpler to schedule can make certification easier.



To incorporate real-time constraints and multiple criticality levels, we propose to use a time-horizon-based planning and control module:

- Short-term (~1s): Avoid collisions with high-speed obstacles.
- Medium-term (~10s): Efficient motion planning in cluttered environments.
- Long-term (~100s): Mission goals, e.g., reaching a distant destination.

Broader Impacts

- *Society*: Safe autonomy will never happen without certification.
- *Industry*: Cross pollination via internships at Northrop Grumman.
- *Mentoring*: Weekly reading group for undergraduate women in CS.

Broadening Participation in Computing

TOPICS (Talking Over Papers In Computer Science)

CS Undergrad Women's Reading Group

- We read papers out loud (you read that right) and discuss them.
- We've read papers on everything from quantum computing to AI to computer security to Turing Award lectures.
- We also talk about writing tips, applying to grad school, and other things.
- We have only two rules:
  - There's no such thing as a dumb question.
  - We do absolutely no work outside of our one hour per week.
- It's a fun group with interesting discussions.

2022 NSF Cyber-Physical Systems Principal Investigators' Meeting
November 8-9, 2022

Award ID#: 2038855