

Edge Intelligence for Structural Health Monitoring: Algorithms, Hardware, and Timing-Aware Co-Design

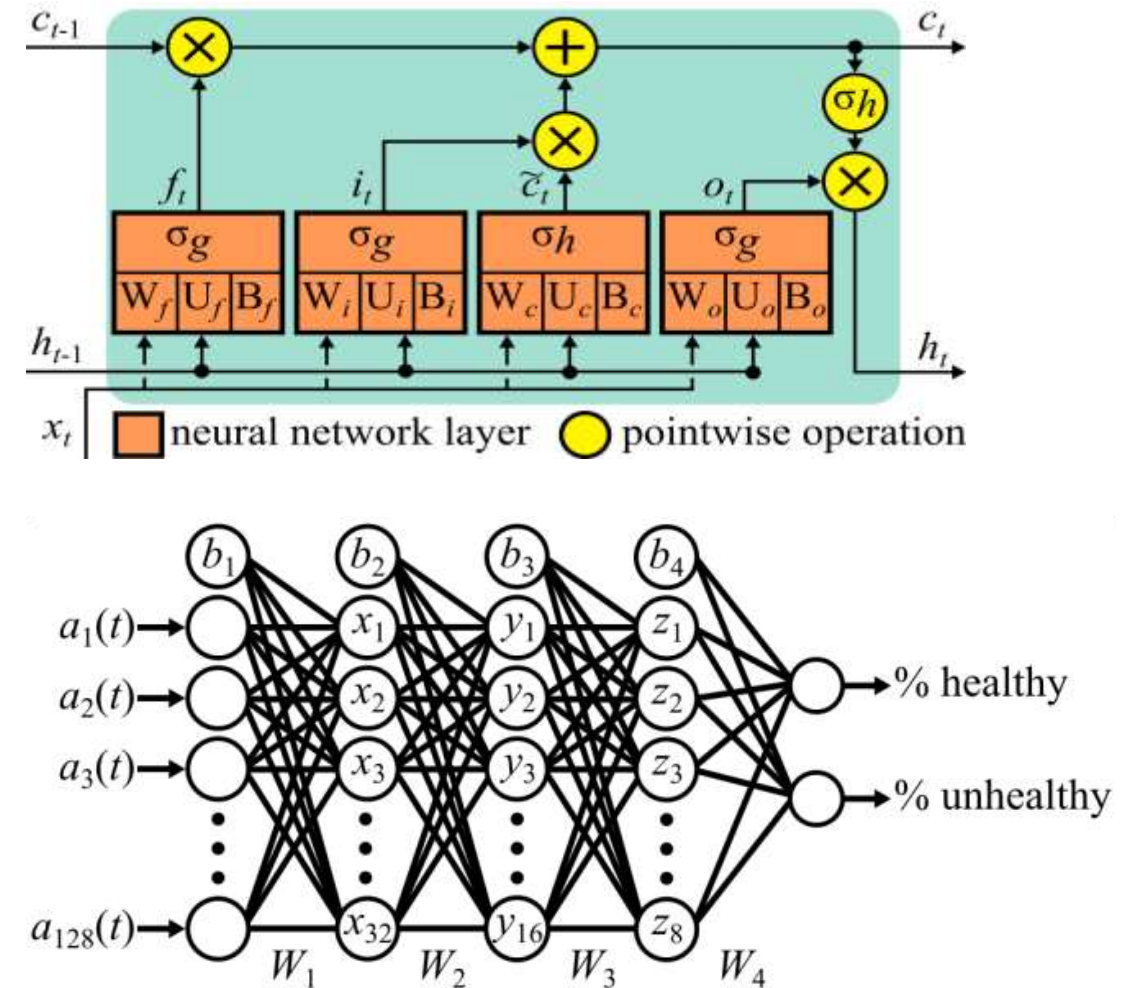
Columbia Civil Engineering & Engineering Mechanics Seminar Series

Austin R.J. Downey
Associate Professor
Mechanical Engineering
Civil and Environmental Engineering



Introduction

- Part 1: The ARTS-Lab.
- Part 2: The Challenge: Hardware–Software Co-Design.
- Part 3: Case Studies
 - Case Study 1: Real-time model updating on microprocessors
 - Case Study 2: On-device LSTM compensation on microcontrollers
 - Case Study 3: Microsecond state estimation with MLPs on FPGAs
- Part 4: Closing Remarks.



A Quick Note on Format

- There are a lot of slides (50!). I'll move through them quickly.
- The goal is to raise ideas, not cover every detail.
- Interrupt anytime with questions or comments.
- I'll save the second half for open discussion.
- Think of this as a conversation, not a lecture.



Part 1: Autonomous Real-Time Systems Laboratory (ARTS-Lab)

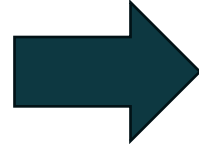
An intelligent controls group working at USC

How We See Ourselves

We use
foundational science



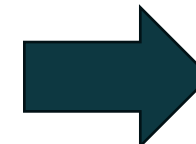
Day School



to develop
essential tools



Dan Thompson



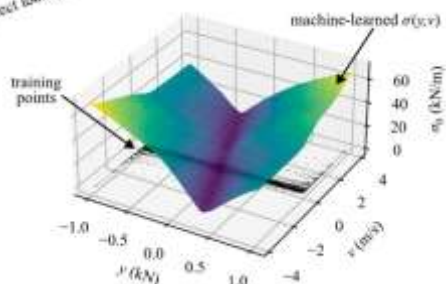
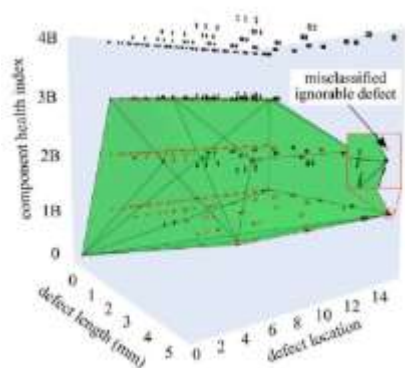
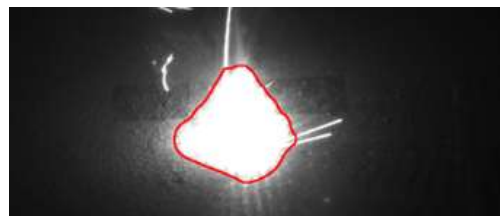
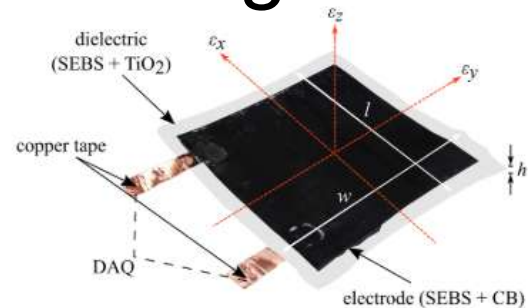
to solve real-world
problems



public domain

**We are Engineers
(mostly)**

Sensing

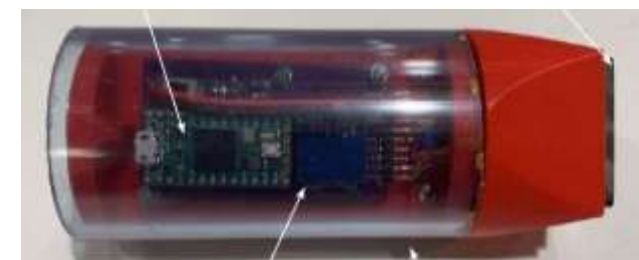


AI/ML

6



Data Assimilation

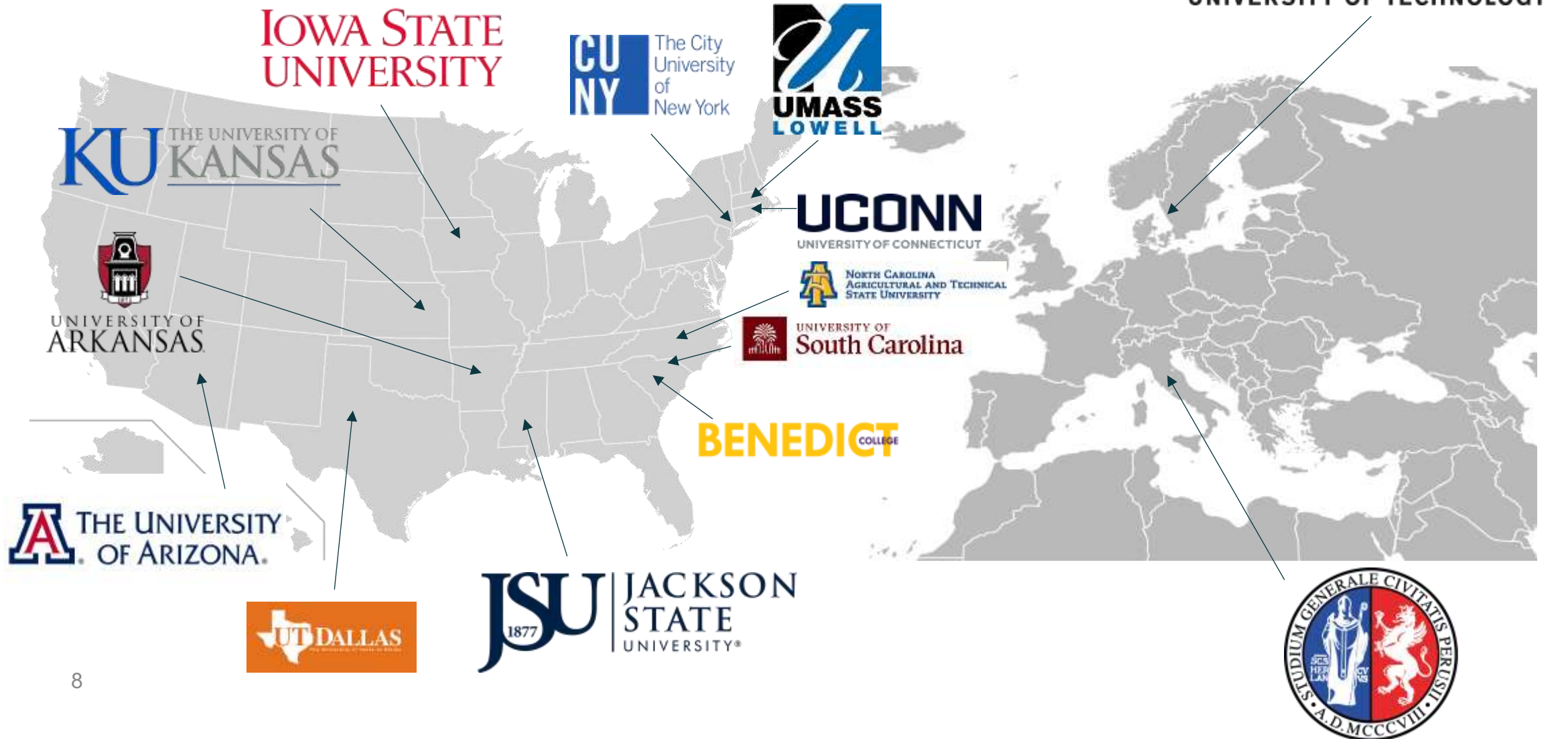


Embedded Systems

Supporting Agencies, Companies, and Partners



Out Academic Partners



Edge Sensing Systems in the ARTS-Lab

- UAV-deployable sensors for water quality, water level, soil moisture, structural vibration, and imaging.
- Autonomous deployment and retrieval mechanisms.
- Embedded systems for shock and vibration in mechanical structures.
- Power budgets in milliwatts.
- Packaging, computing, energy storage, sensing, and algorithms are co-designed as one system.
- Today's talk covers a small subset of this work.



Part 2: The Challenge: Hardware– Software Co-Design

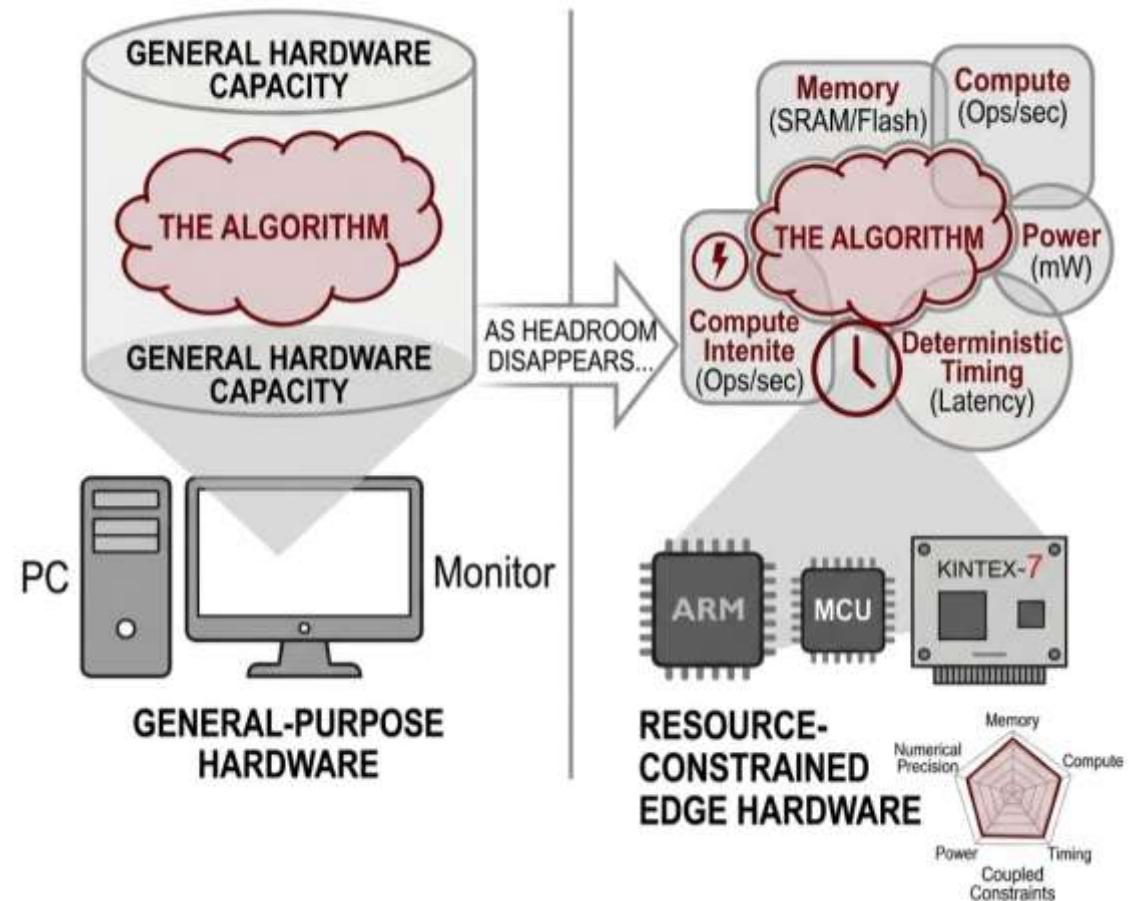
The Case for Intelligence at the Edge

- Sensing is cheap. The value comes from turning data into decisions where it's collected.
- Speed: edge decisions meet windows from minutes (infrastructure) down to microseconds (shock).
- Energy: sending a result costs far less than streaming raw data, because the radio often outweighs the compute.
- Autonomy: remote, mobile, and comms-denied systems keep working with no link home.
- Payoff: a shorter path from measurement to assessment.



Co-Design Begins Where Headroom Ends

- General-purpose hardware: the algorithm fits with room to spare, so the code just runs.
- Edge hardware: the headroom disappears, and memory, compute, power, timing, and precision press in on the algorithm.
- These constraints are coupled. Lower bit-width saves memory and time, but costs accuracy.
- Co-design: the algorithm and the hardware are designed together, just to make it work.



Selecting Edge Computing Hardware

Processing hardware comes in all shapes and sizes!

- Microprocessor
 - Programming: easy, with a full OS and software ecosystem
 - Peripherals: reached through OS drivers, less directly
 - Execution: multithreaded, but timing isn't guaranteed
- Microcontroller
 - Programming: relatively easy, bare-metal C/C++
 - Peripherals: direct access to sensors over I²C, SPI, and ADCs
 - Execution: single-threaded with predictable timing
- FPGA
 - Programming: hard, with circuits designed in HDL or HLS
 - Peripherals: any interface you build in logic
 - Execution: truly parallel and fully deterministic



Micro-processor



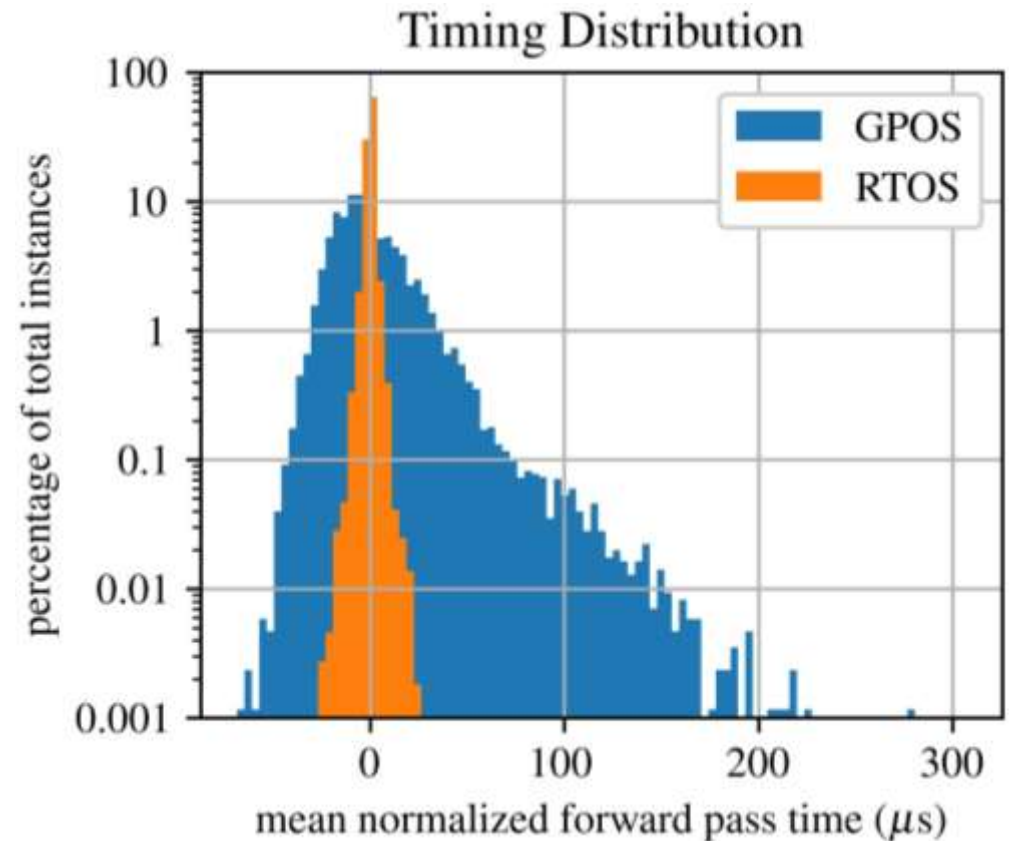
Micro-controller



FPGA

Latency and Deterministic Timing

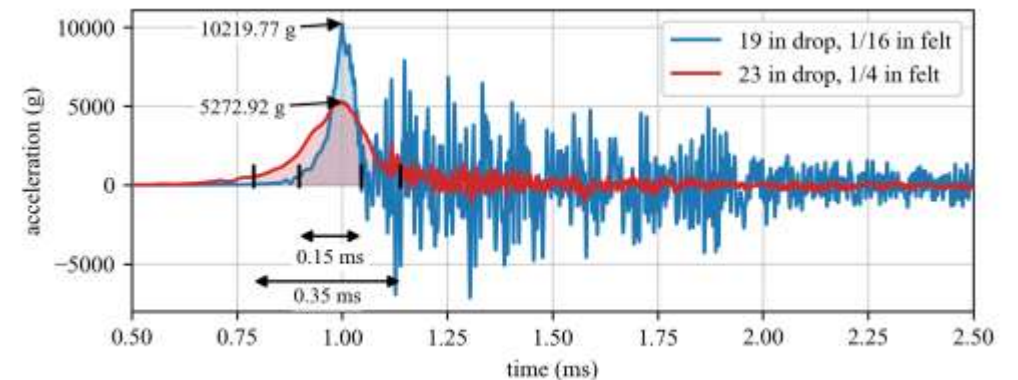
- Determinism: how repeatable latency is from run to run.
- Deadlines are met by the slowest run. In closed-loop control, a late answer is a wrong answer.
- General-purpose OS (GPOS, e.g., Windows, Linux): shares the CPU across many tasks. Fast on average, with a tail past +250 μ s.
- Real-time OS (RTOS, e.g., FreeRTOS, NI Linux Real-Time): schedules by priority. No faster, but more consistent.
- FPGA: no OS at all, so latency is fixed by the circuit and identical every run.



Case Study 1: Real-time model updating (LEMP) on microprocessors

High-Rate Dynamics

- High-rate events last under 100 ms and exceed $100 g_n$: blasts, impacts, crashes.
- They bring large load uncertainty, strong non-stationarity, and unmodeled dynamics.
- Latency targets from event detection to decision:
 - High-rate: 1 ms
 - Very high-rate: 100 μs
 - Ultra high-rate: 1 μs
- The grand challenges are adequate sensing, limited system knowledge, high variability in loading, and limited compute and power for real-time algorithms.

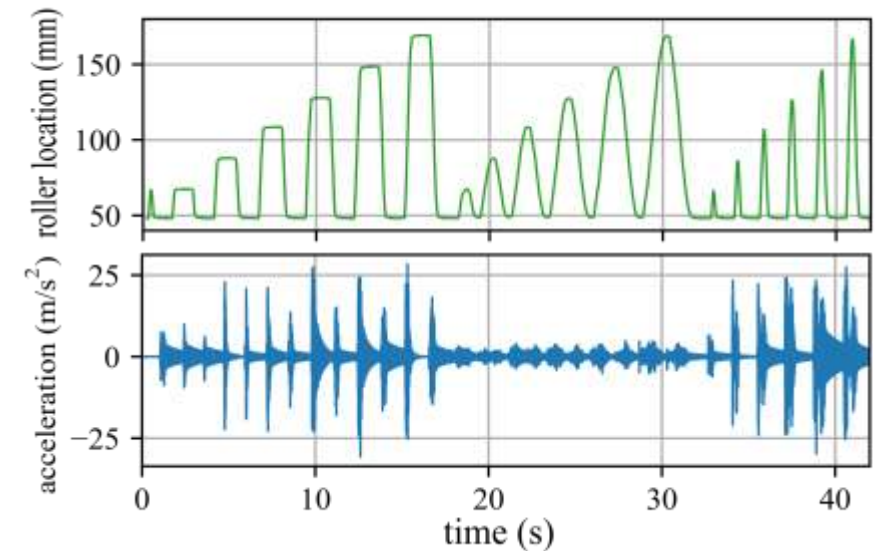


(Downey and Bakos, 2023)

Experimental System used for this work

DROPBEAR experimental testbed:

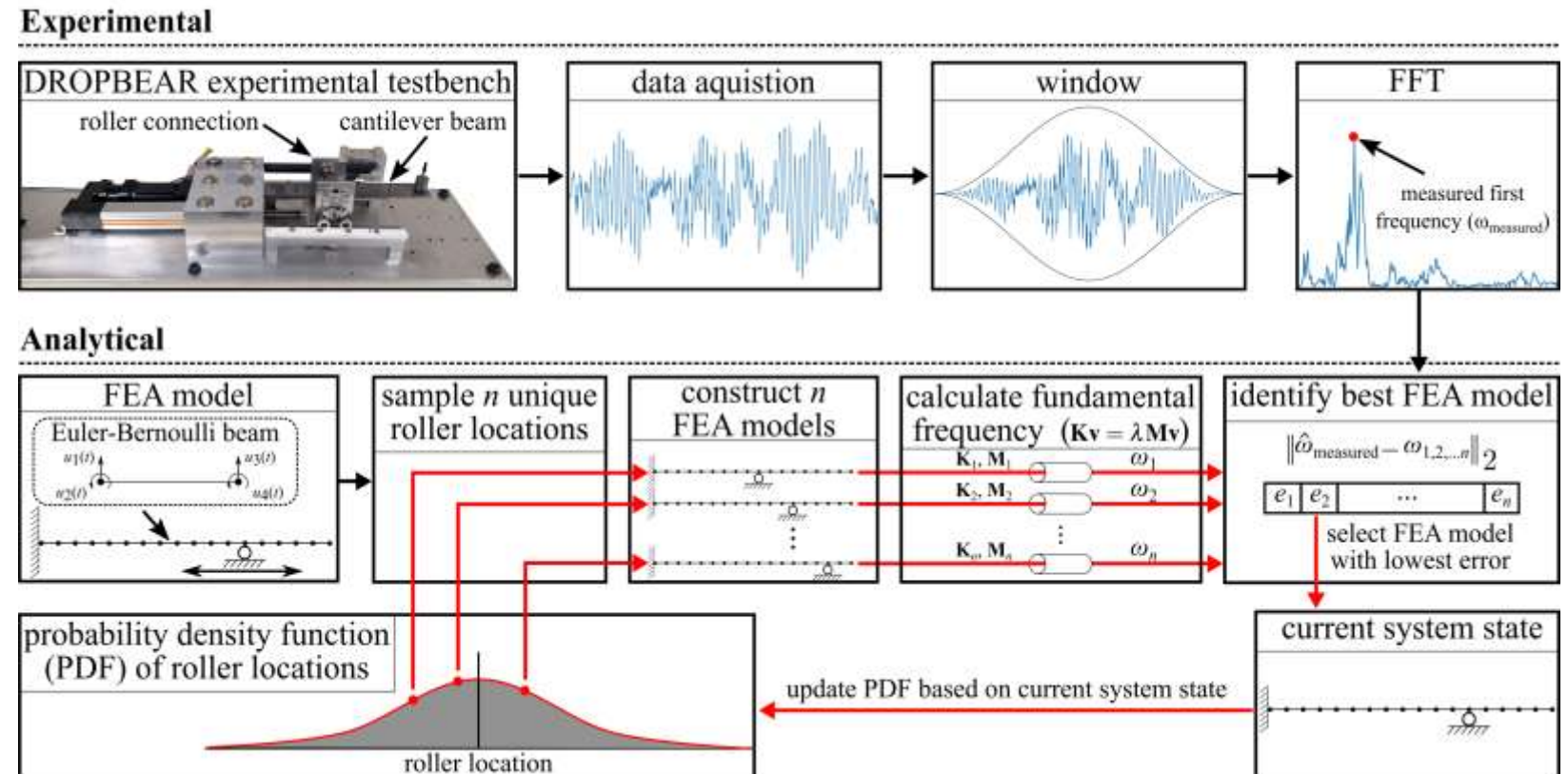
- The Dynamic Reproduction of Projectiles in Ballistic Environments for Advanced Research (DROPBEAR) was used to generate the experimental data in this work.
- Cantilever beam with a controllable roller to alter the state.
- Acceleration and pin location are recorded.
- Dataset available on GitHub at:
<https://github.com/High-Rate-SHM-Working-Group/Dataset-2-DROPBEAR-Acceleration-vs-Roller-Displacement>



(Nelson et al., 2022)

Real-Time Model Updating Through Error Minimization

- A frequency-based model updating technique was developed to update an FEA model of the system.



Eigen Value Problem - Can we just solve it faster?

- Hardware accelerators do exist, but their throughput is limited by communication bandwidth.

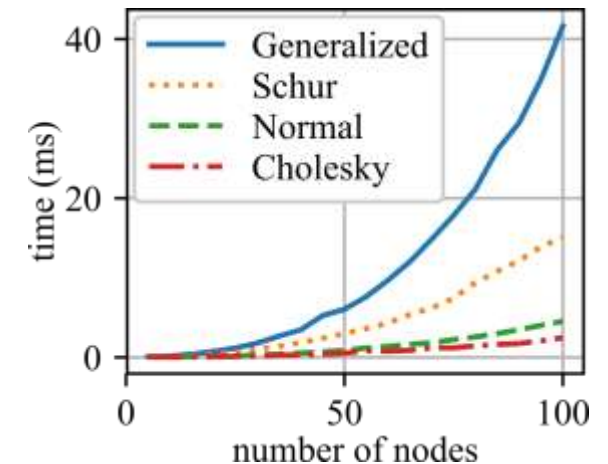
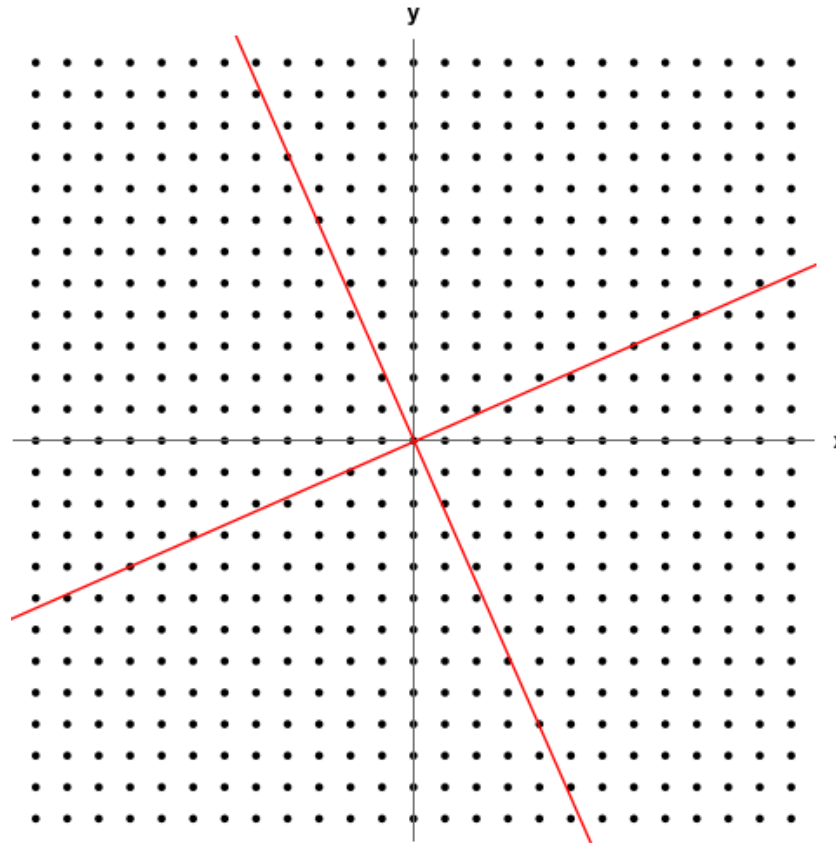
$$\mathbf{M}_1 \ddot{x} + \mathbf{K}_1 x = 0$$

$$\det[\mathbf{K}_1 - \lambda \mathbf{M}_1] = 0$$

$$[\mathbf{K}_1 - \lambda \mathbf{M}_1] \mathbf{U}_1 = 0$$

$$\lambda = \begin{bmatrix} \omega_1^2 & 0 & 0 & 0 \\ 0 & \omega_2^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \omega_n^2 \end{bmatrix}$$

$$\mathbf{U}_1 = [\vec{u}_1^1 \quad \vec{u}_2^1 \quad \cdots \quad \vec{u}_n^1]$$



Developed a specialized Cholesky-Jacobi method formulated specifically for this challenge; a 66 node FEA model can be solved for within the 1 ms.

LEMP usage in the '70s and '80s

- LEMP enabled these calculations to be done very efficiently on very slow desktop computers.
- Structural Measurement Systems (SMS) sold a custom hardware and software setup.
- This was before the “personal computer” stage.



HP1000/A700 w/DIFA modal analysis system



20 HP5423 first dedicated
FFT/Modal system - 1979

*SMS modal
software
called SDM
used LEMP*

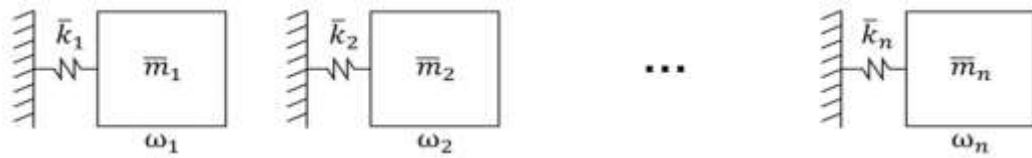


HP3000 desktop running “Rocky Mountain BASIC”

All images and knowledge courtesy of Peter Avitabile Professor Emeritus, Co-Director - Structural Dynamics & Acoustic Systems Laboratory at the University of Massachusetts Lowell

Local Eigenvalue Modification Procedure

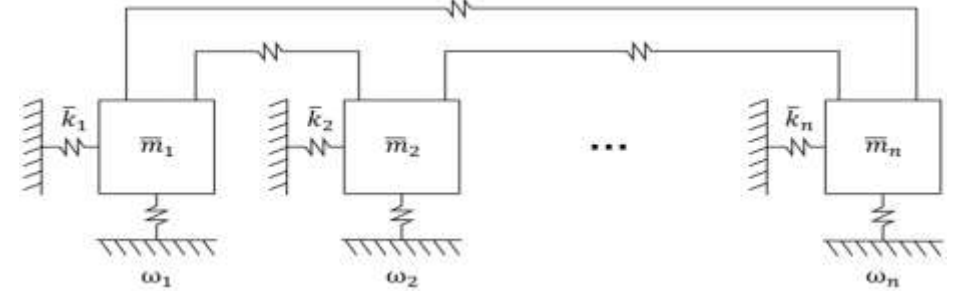
n independent single DOF systems representing the initial state



Initial State

Modification

Coupled single DOF systems representing the altered state



Altered State

Physical Space

$$[\mathbf{M}_1], [\mathbf{K}_1]$$

$$[\Delta \mathbf{M}_{12}], [\Delta \mathbf{K}_{12}]$$

$$[\mathbf{M}_2], [\mathbf{K}_2]$$

'n'
Physical DOF

Modal Transformation

$$\{x\} = [U_1]\{p_1\}$$

$$\{x\} = [U_2]\{p_2\}$$

$m \ll n$

Modal Space

$$[\omega_1^2], [U_1]$$

$$\{p_1\} = [U_{12}]\{p_2\}$$

$$[\Omega_2^2], [U_2]$$

'm'
Modal DOF

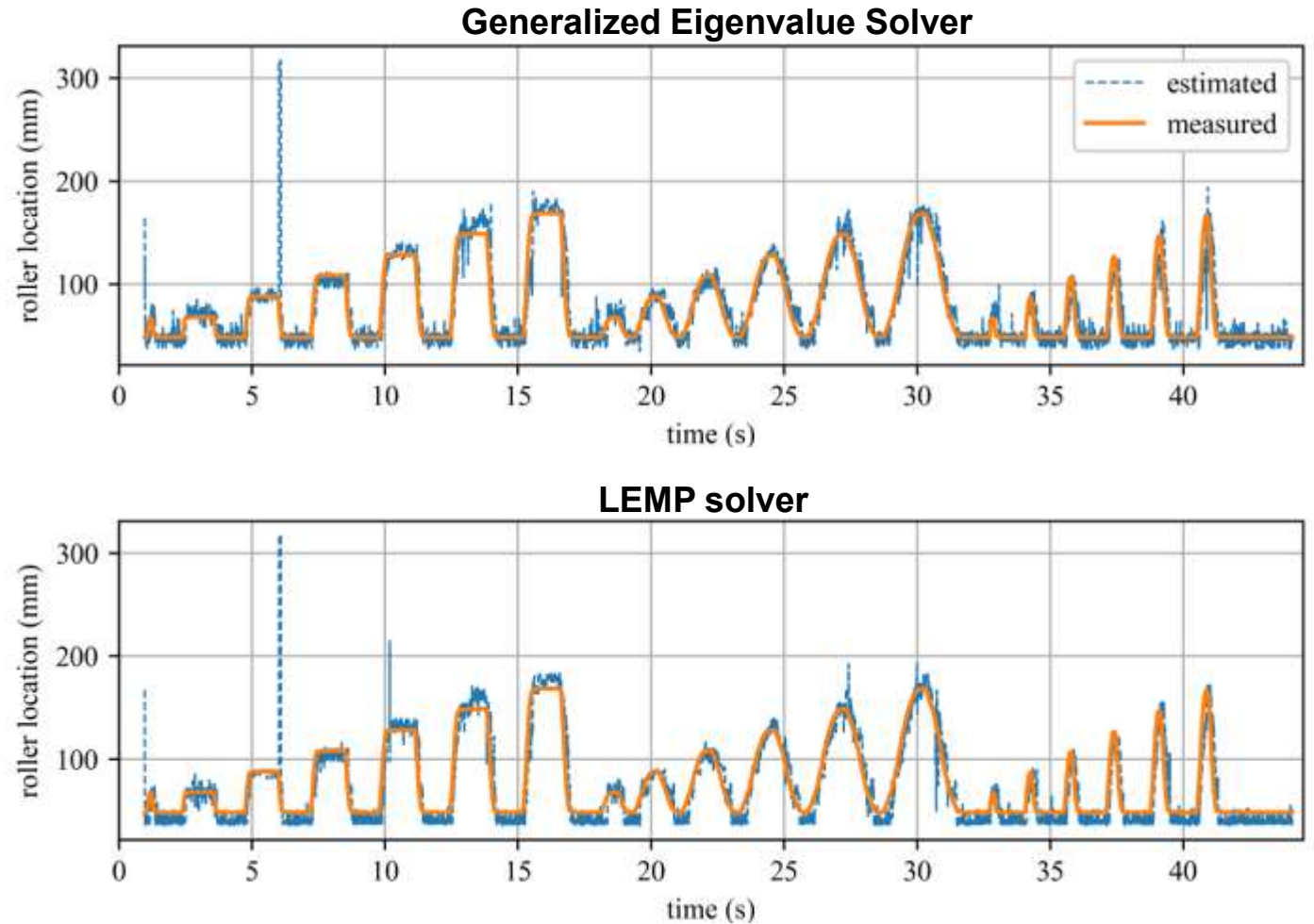
$$\frac{-1}{\alpha} = \sum_{r=1}^m \frac{v_r^2}{\omega_r^2 - \Omega_r^2}$$

Solved using Divide and Conquer method

LEMP State Estimation Results

LEMP has been shown to:

- Have similar accuracy to the generalized eigenvalue solver
- Be robust to sensor noise
- Be capable of working with various error estimators developed for the project

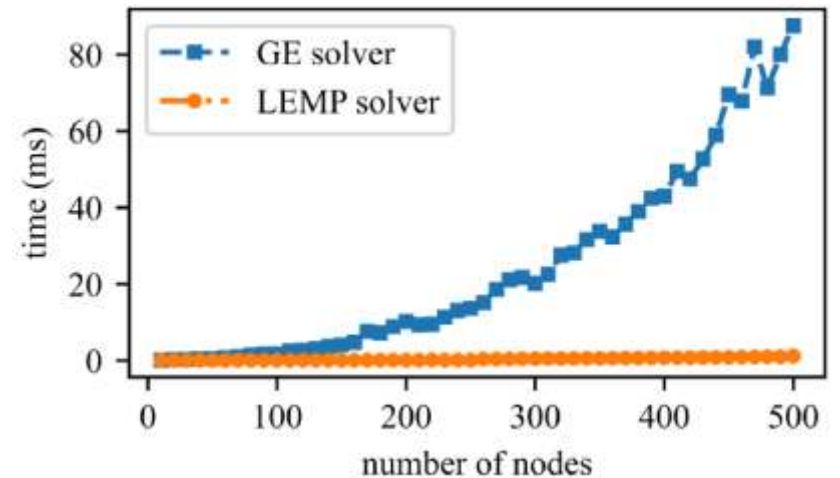


(Ogunniyi et al. 2023)

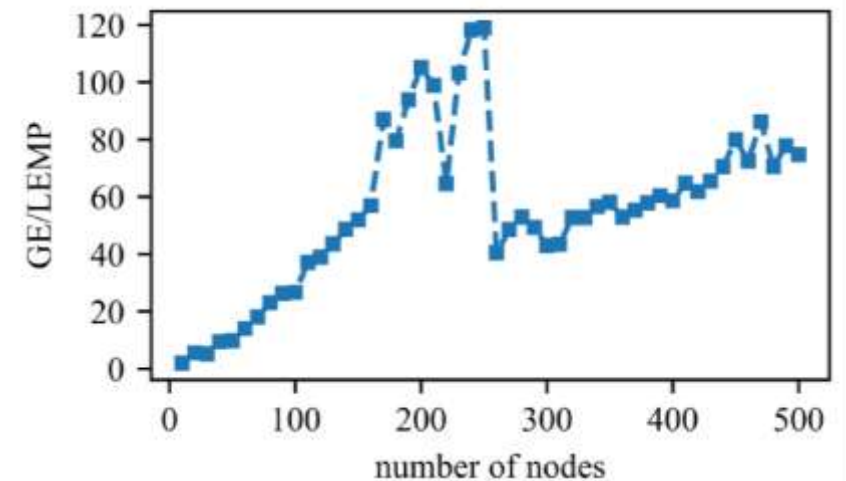
LEMP Scales Well

- The generalized eigenvalue solver grows to ~90 ms at 500 nodes.
- LEMP stays near 1 ms across the same range.
- Speed-up climbs steadily to ~120× near 250 nodes.
- On a 2D plate, the $O(n)$ formulation reaches 0.46 ms vs. 9.03 ms at 25 nodes, stays under 2 ms at 169 nodes, and exceeds 2500× speed-up.

Algorithm Timing

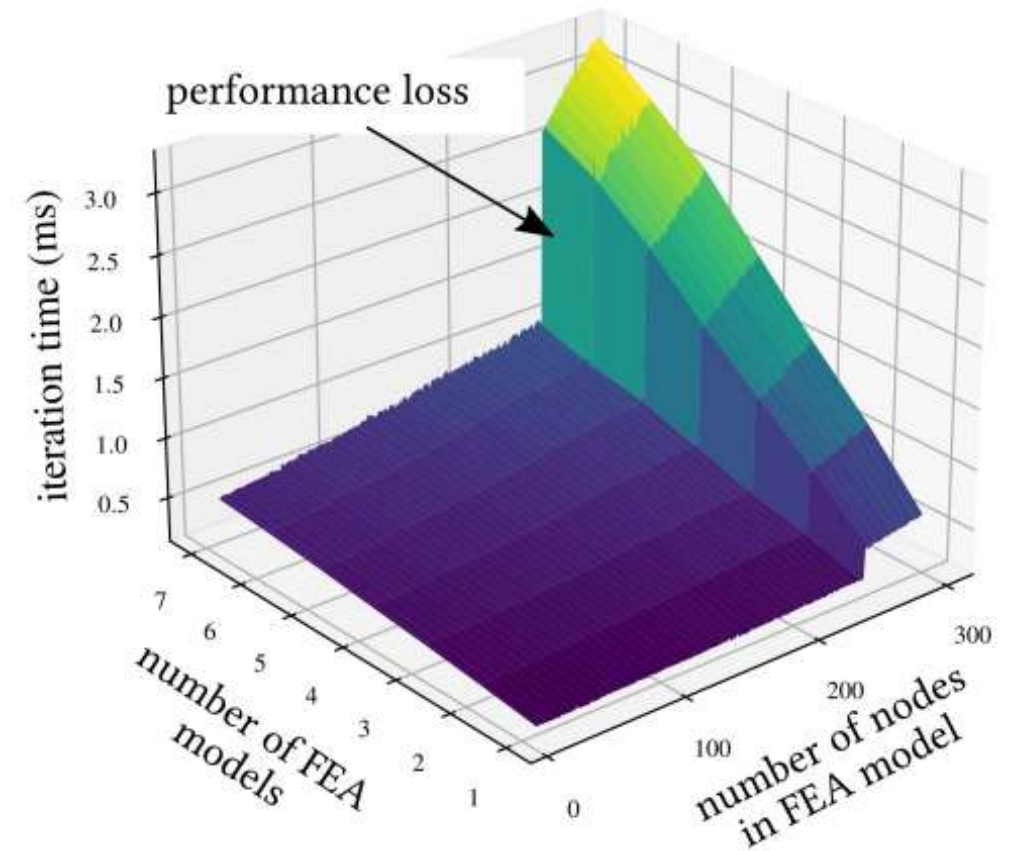


Speed up Factor



The Memory Wall in Hardware–Software Co-Design

- Past ~250 nodes, speed-up falls abruptly from ~120× to ~40×.
- Iteration time jumps from ~0.5 ms to ~3 ms at the same point.
- The algorithm didn't change. The model outgrew the cache and spilled into main memory.
- Algorithmic scaling and hardware scaling are not the same thing.
- No device has unlimited high-speed memory.

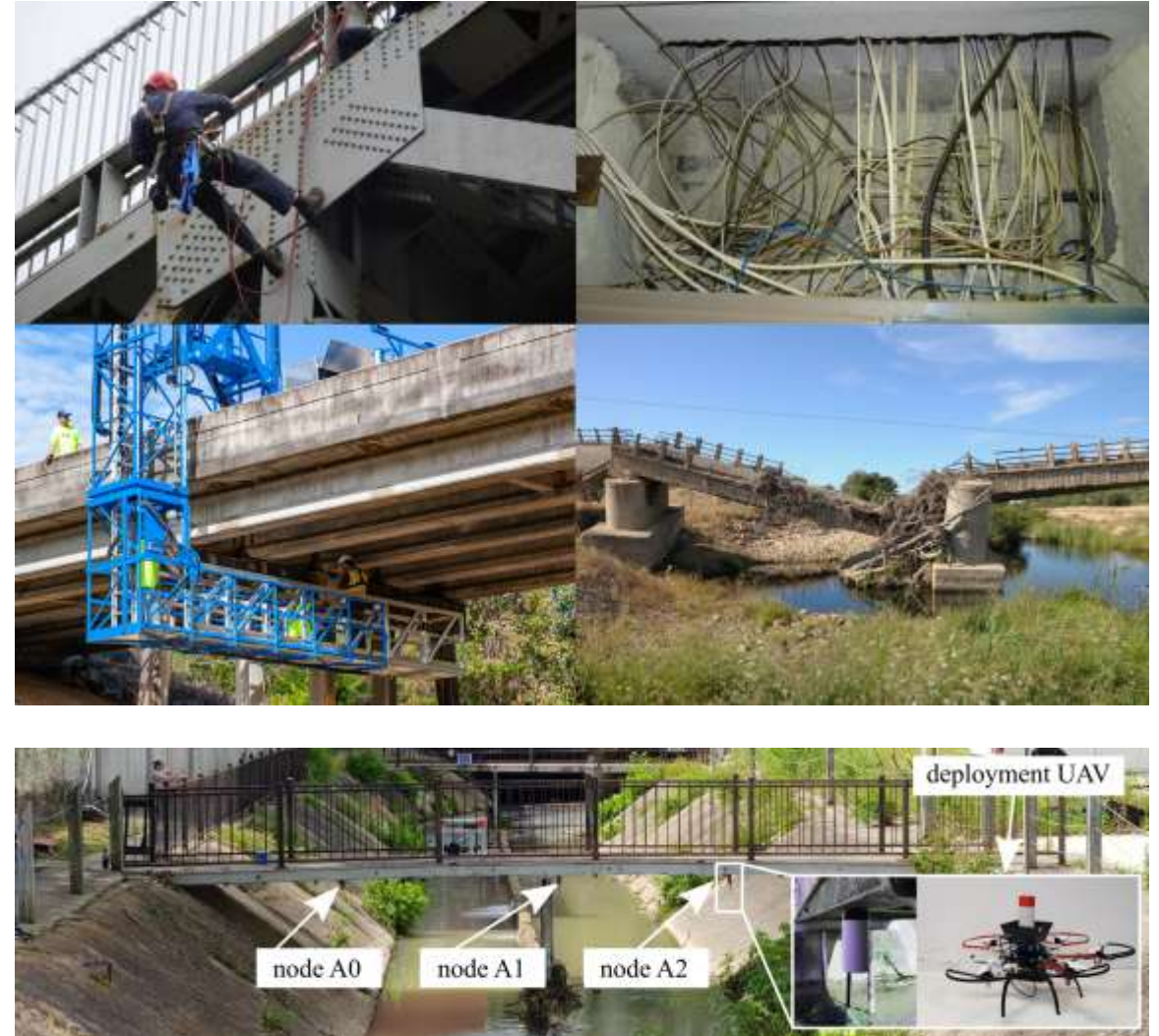


(Downey and Bakos, 2023)

Case Study 2: On-device LSTM compensation on microcontrollers

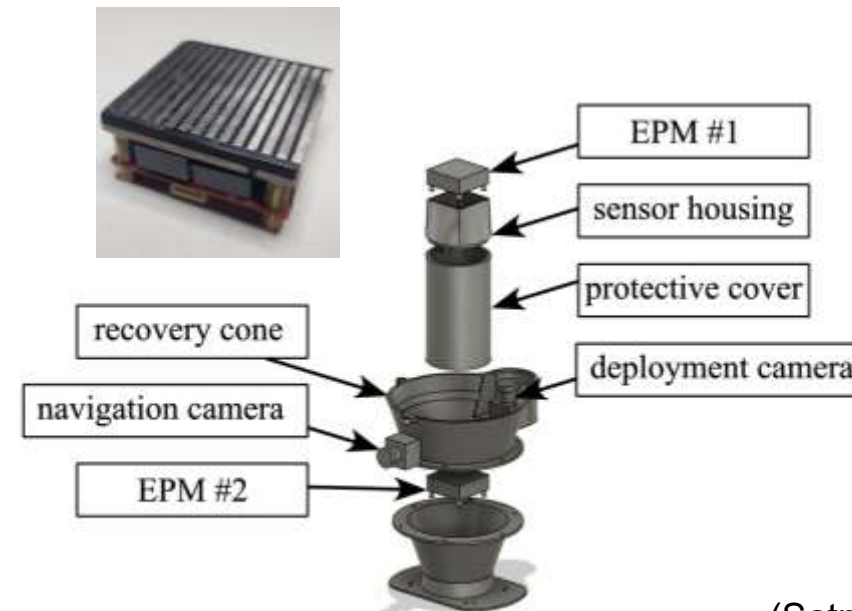
Rapid Assessment After Extreme Events

- After earthquakes, floods, and hurricanes, large inventories of structures need fast assessment.
- Manual inspection is slow and costly, and it puts people on unstable structures.
- UAVs deliver and retrieve sensors without anyone on the structure.
- Goal: minimize the time from acquisition to structural assessment.



DOVE: A UAV-Deployable Sensor Package

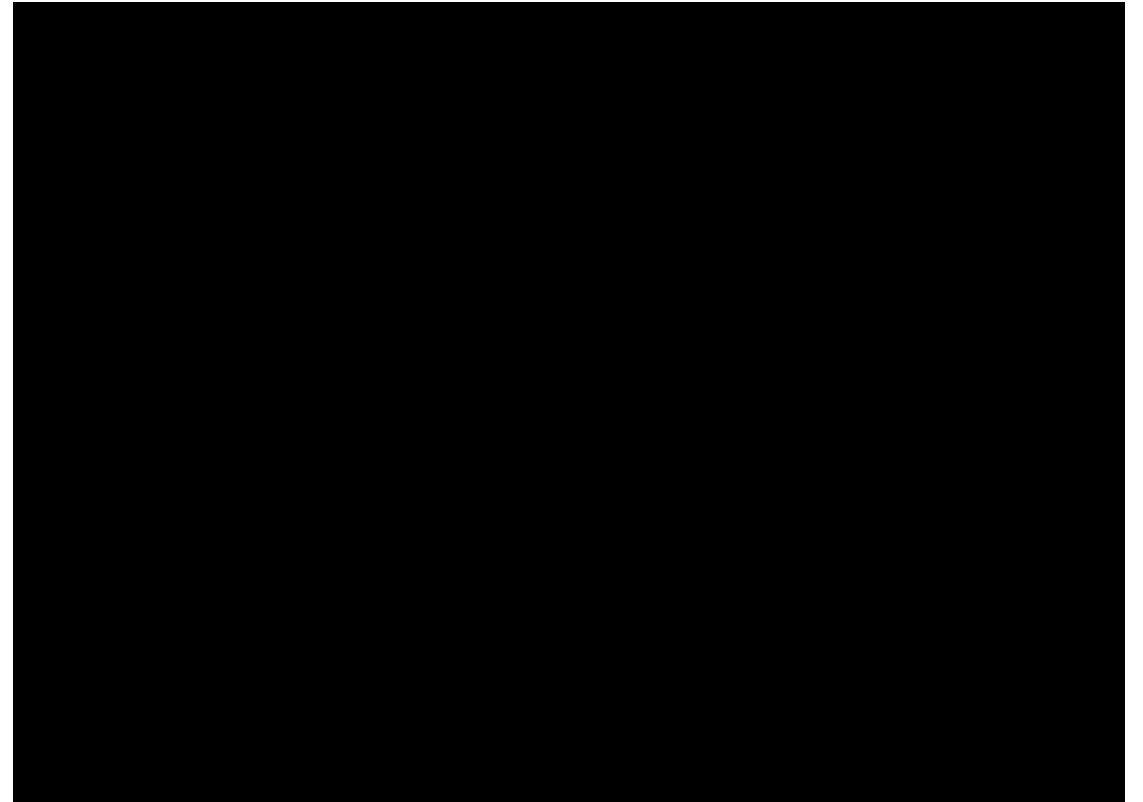
- Dynamics Observation and Vibration at the Edge.
- Electropermanent magnets hold the package to steel at zero power.
- A camera-guided recovery cone.
- An ARM Cortex-M7 microcontroller runs at 600 MHz with no OS.
- MEMS accelerometer up to 28 kS/s, SD storage, 2.4 GHz radio.
- Open source: github.com/ARTS-Laboratory/DOVE



(Satme et al., 2024)

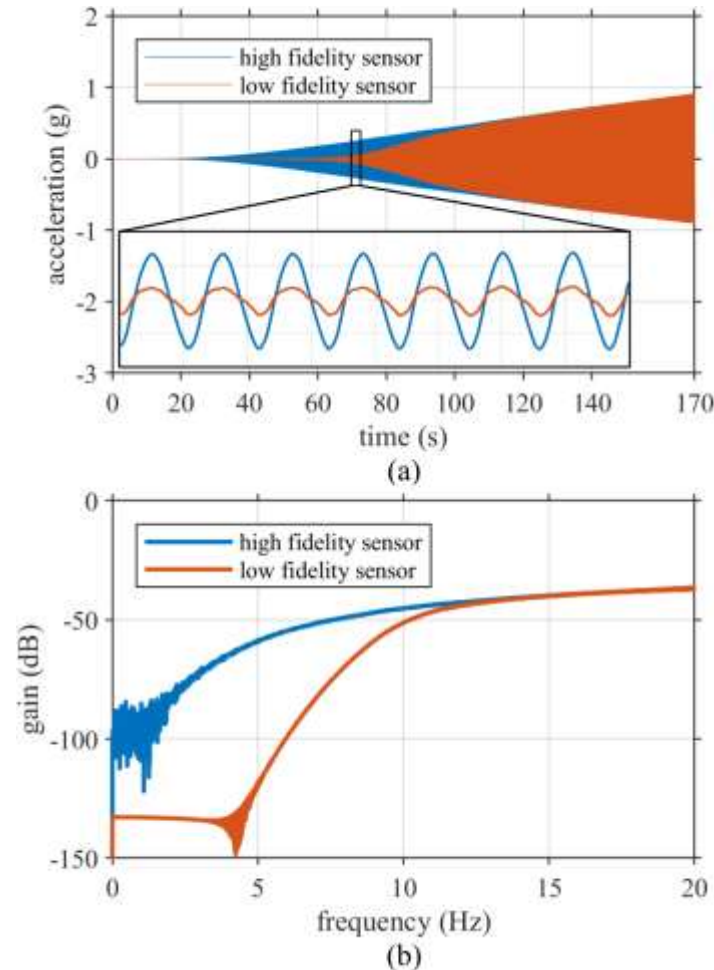
The Cost of a Deployable Package

- Limits on payload, power, and size force the use of low-cost MEMS accelerometers.
- MEMS sensors lose sensitivity and add nonlinear error below 20 Hz, where civil structures respond.
- The package and its magnetic mount attenuate and distort the transmitted response.
- Classical transfer-function correction misses the nonlinear behavior.



Why Compensation Is Needed

- Deployable packages favor low-cost, low-power, lightweight MEMS accelerometers.
- Both sensor and package mounts attenuate low-frequency content.
- The loss is concentrated below 10 Hz, where many civil structures respond.
- Every sensing system has a transfer function, and some correction is always needed.
- Learned compensation recovers what the hardware gives up.



(Satme et al., 2023)

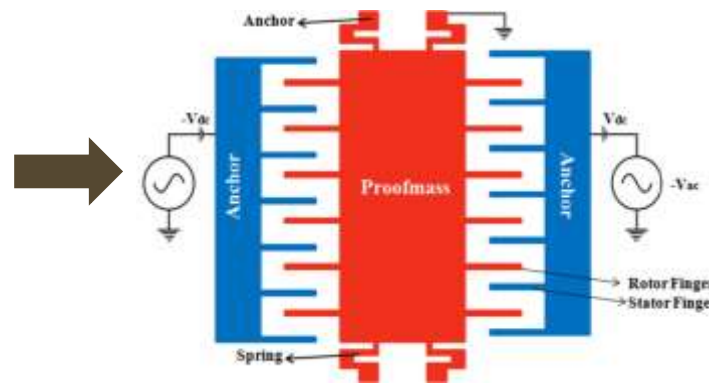
Learning the Compensation

- An LSTM learns the mapping from degraded measurements to the true response.
- The degraded input is a 10 Hz high-pass-filtered copy of the reference.
- Training records were 0–20 Hz chirps at 170 s long at 400 S/s.
- The model is sequence-to-one: each 400-sample window produces one compensated sample, with a stride of one.

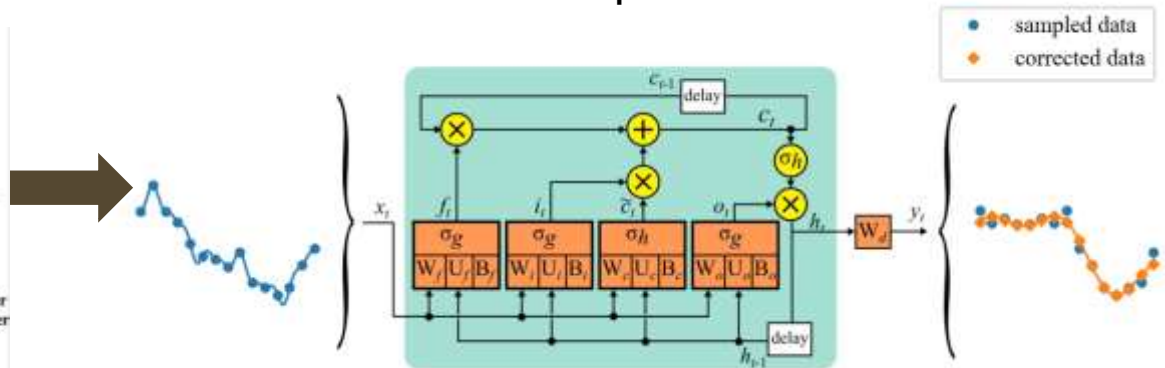
sensor deployment



accelerometer measurement

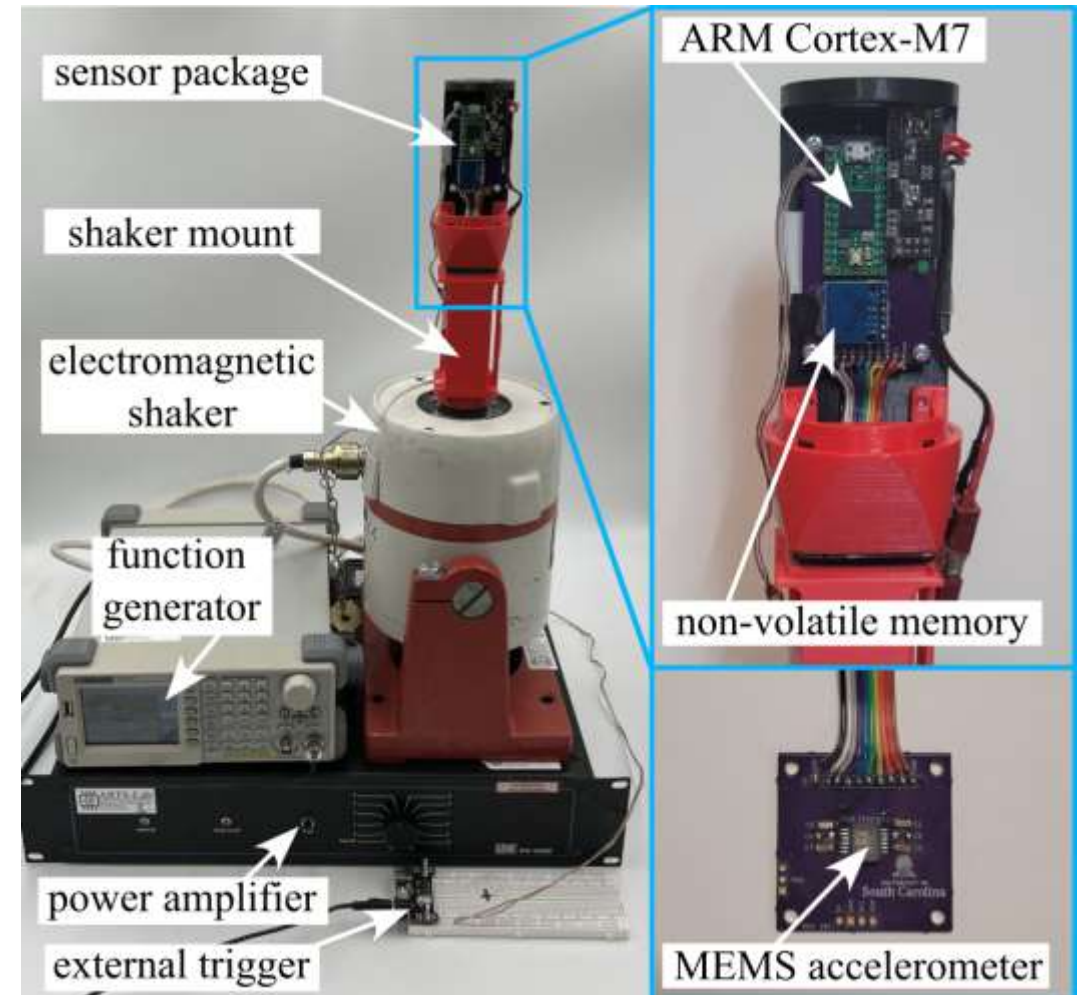


error compensation



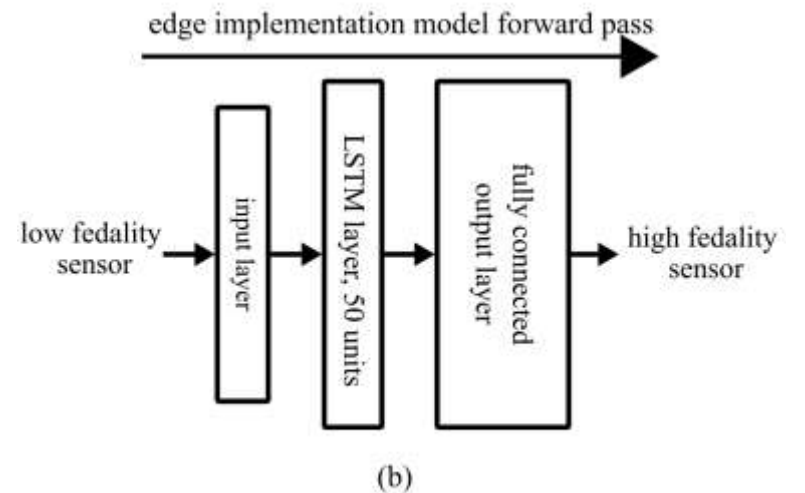
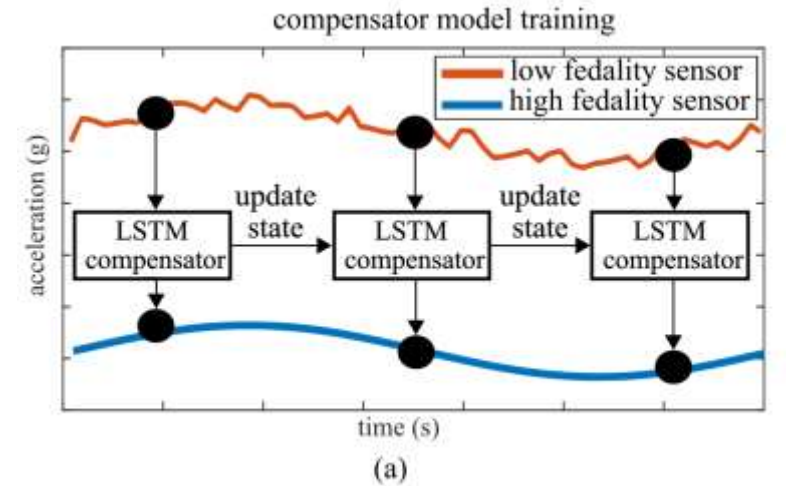
A Shaker-Based Training Testbed

- An electromagnetic shaker drives DOVE with a known control signal.
- A digital trigger synchronizes the excitation and the measurements.
- The onboard MEMS accelerometer provides the high-fidelity reference.
- A 5th-order, 10 Hz Butterworth high-pass filter creates the degraded input.
- Chirp and single-frequency excitation span 0.1–20.9 Hz.



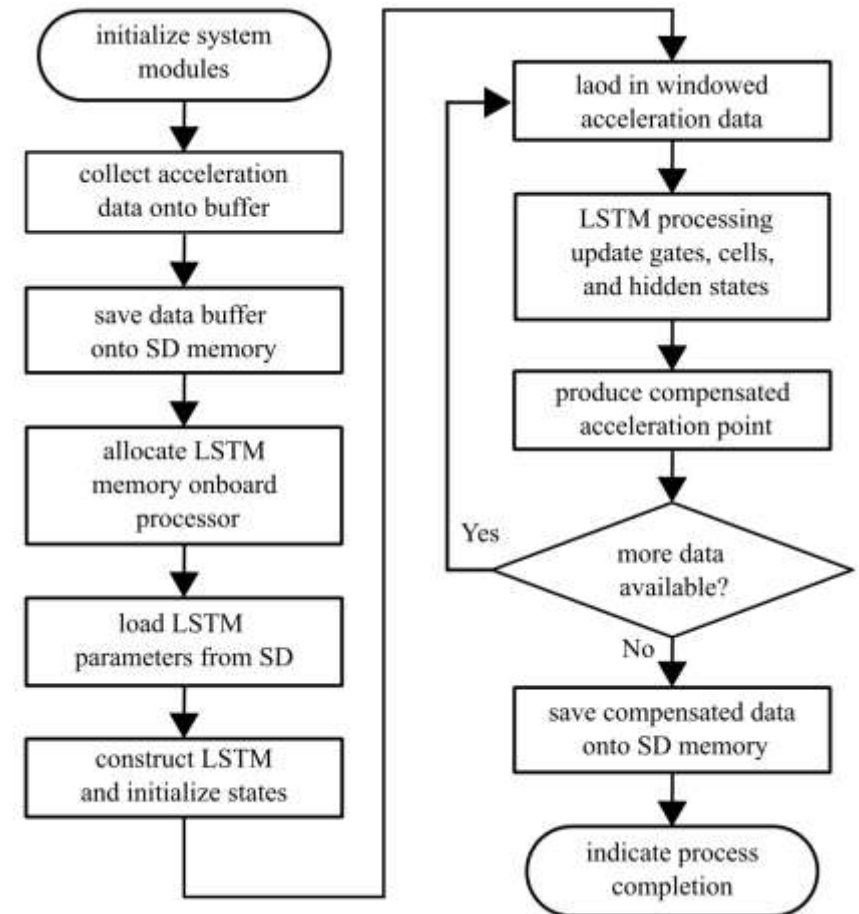
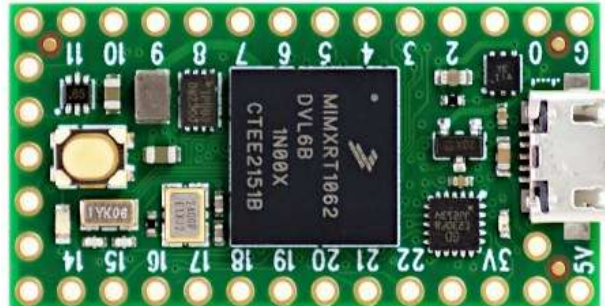
An LSTM Compensator at the Edge

- An LSTM maps the low-fidelity signal back to the high-fidelity response.
- The goal is to restore low-frequency content at the correct amplitude and phase.
- The recurrent state carries signal history, which captures nonlinear, time-dependent distortion.
- The model is sequence-to-one: each 400-sample window yields one compensated sample.
- It is trained offline on paired signals, then runs as a single forward pass on the microcontroller.



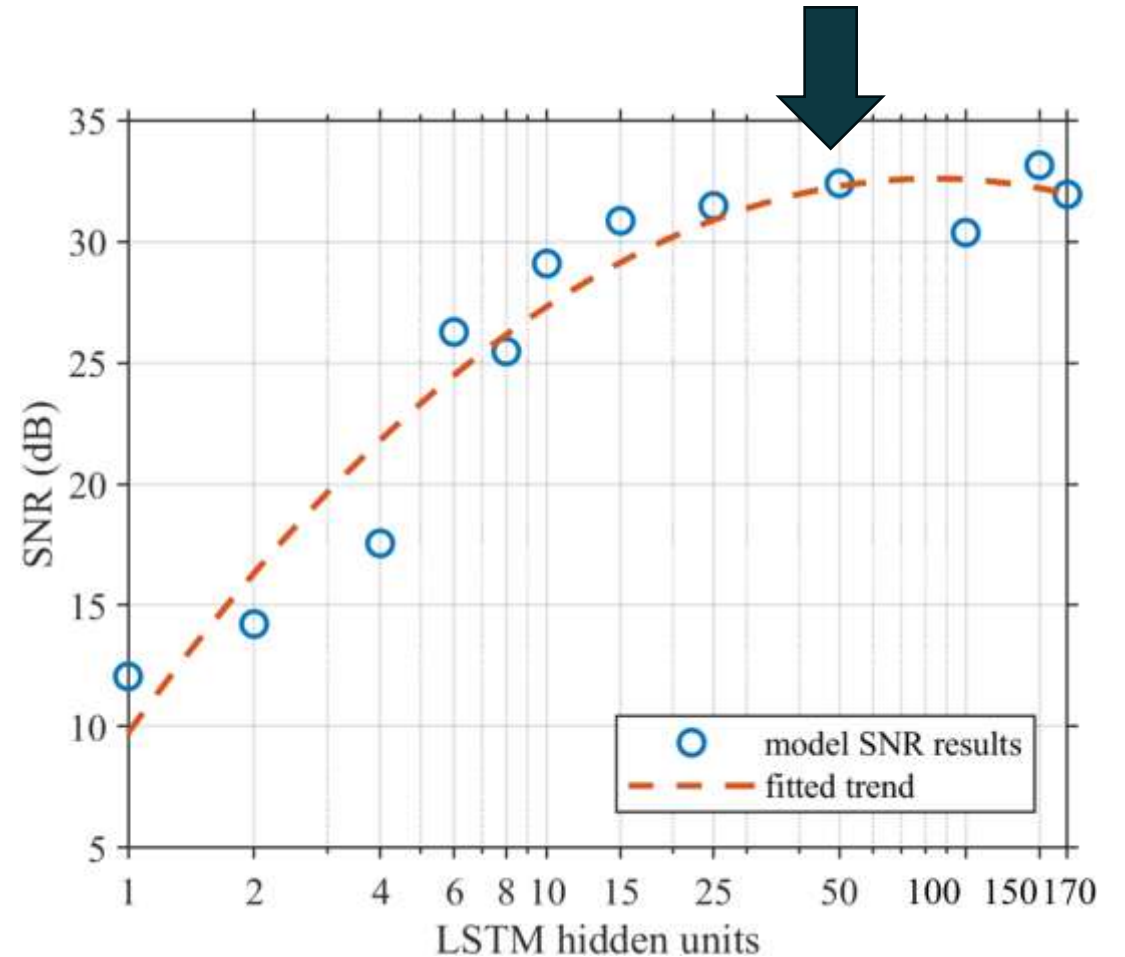
Moving Data Through Limited Memory

- 69,000 16-bit samples are acquired into on-chip RAM.
- The raw data is offloaded to the SD card.
- 400-sample windows are streamed from SD through the network written back.
- SD card read/write adds significant power cost.
- Could a custom board with dedicated memory over I²C close this gap?



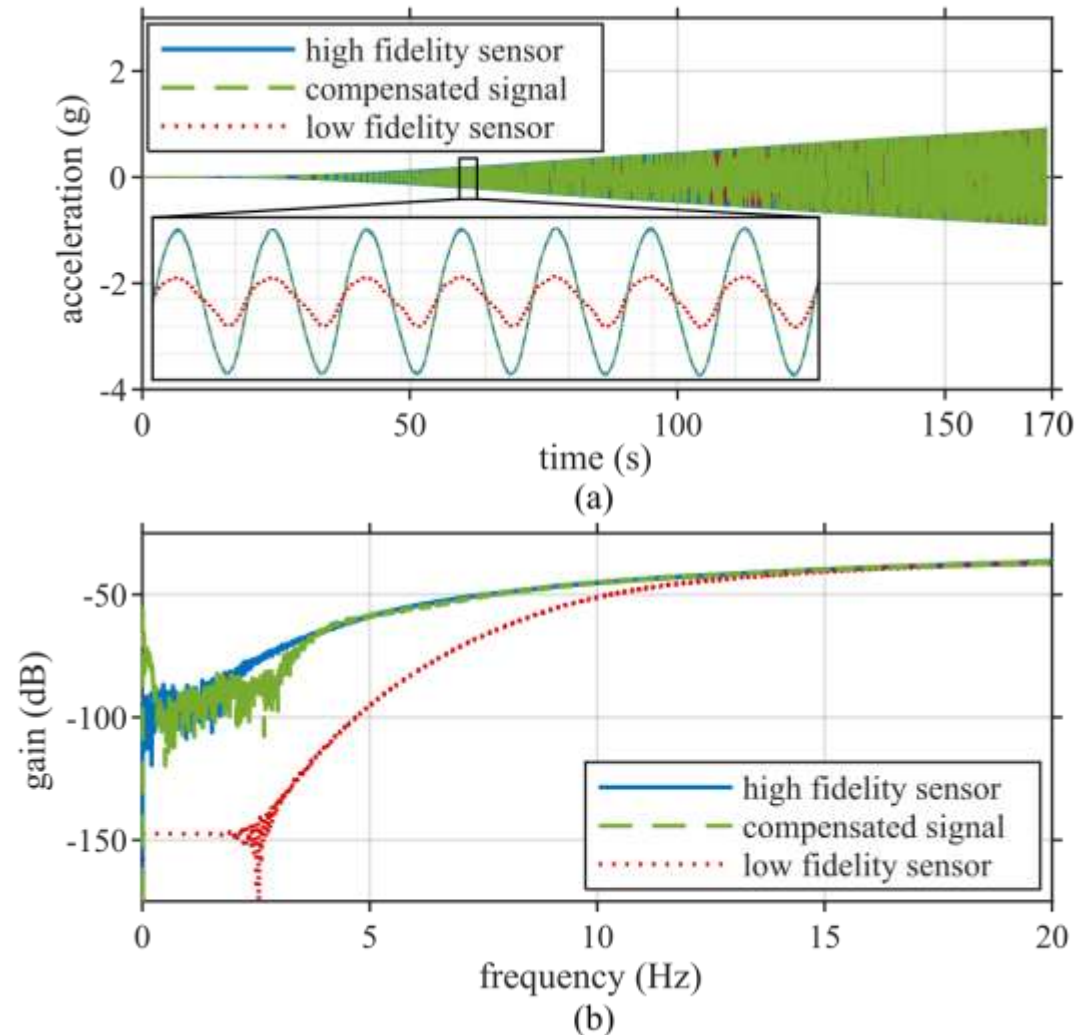
The Selected 50-Unit Model

- Best balance of reconstruction accuracy and hardware cost.
- SNR of 32.39 dB, a 17.94 dB improvement.
- RMSE of 0.0082 g, an 87.32% reduction.
- TRAC of 0.9994.
- 106.69 ms per 400-sample forward pass.



Recovering the Low-Frequency Response

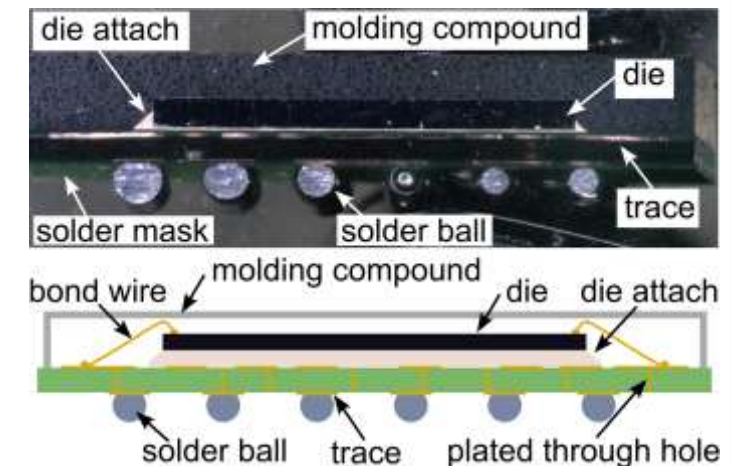
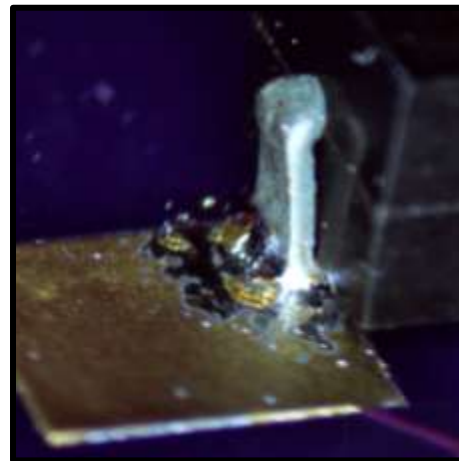
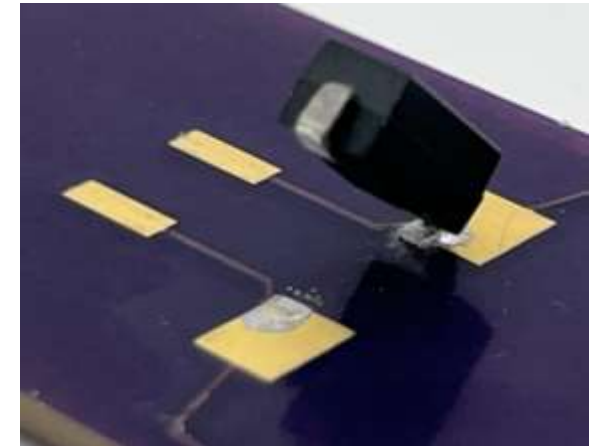
- The LSTM restores signal content attenuated below 10 Hz.
- The compensated FRF approaches unity across 0–20 Hz.
- The compensated PSD tracks the high-fidelity reference.
- Compensation runs entirely on-device, with no external post-processing.



Case Study 3: Microsecond state estimation with MLPs on FPGAs

Electronics Under Shock

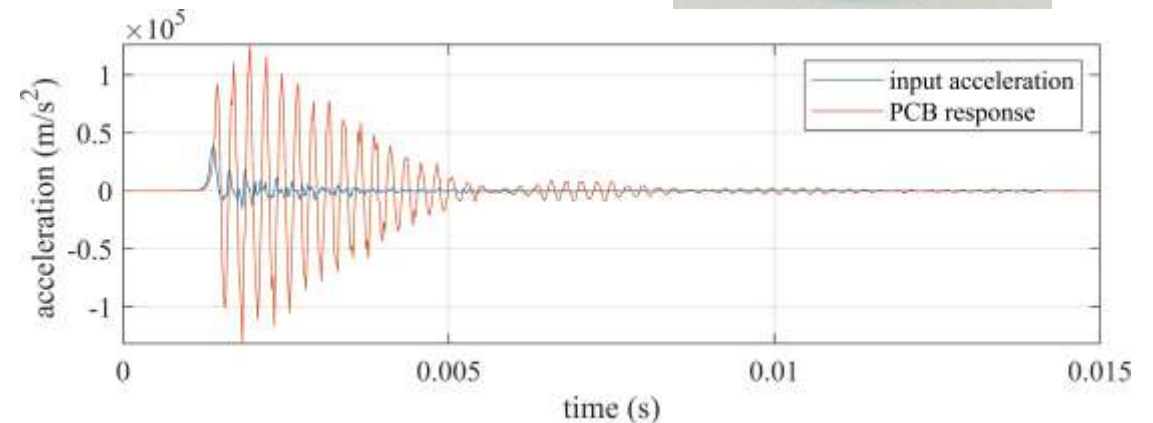
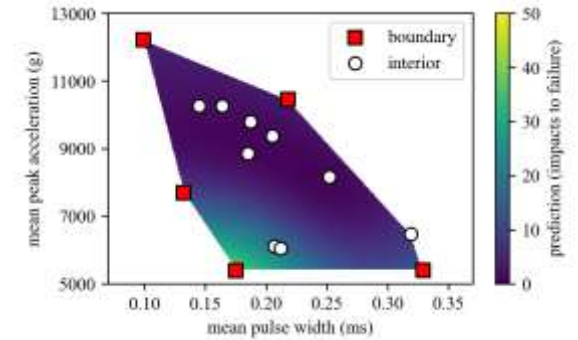
- PCBs are the structural and electrical backbone of every embedded system.
- Drop-tower shocks exceed 10,000 g with pulse widths of 0.15–0.35 ms.
- Board bending concentrates stress at solder joints and interconnects.
- Damage builds up over repeated impacts, often before any functional failure.
- Goal: estimate PCB health state from the shock response itself.



(Yount et al., 2025)

Building High-Rate Degradation Datasets

- Drop-tower testing of circular, cantilever, and fixed–fixed PCBs.
- Component-level fatigue studies on resistors and BGA dummy components.
- Impact energy and boundary conditions are varied under control.
- In-situ resistance measurements label each board's health state.
- Kriging models guide test selection and predict failure boundaries.

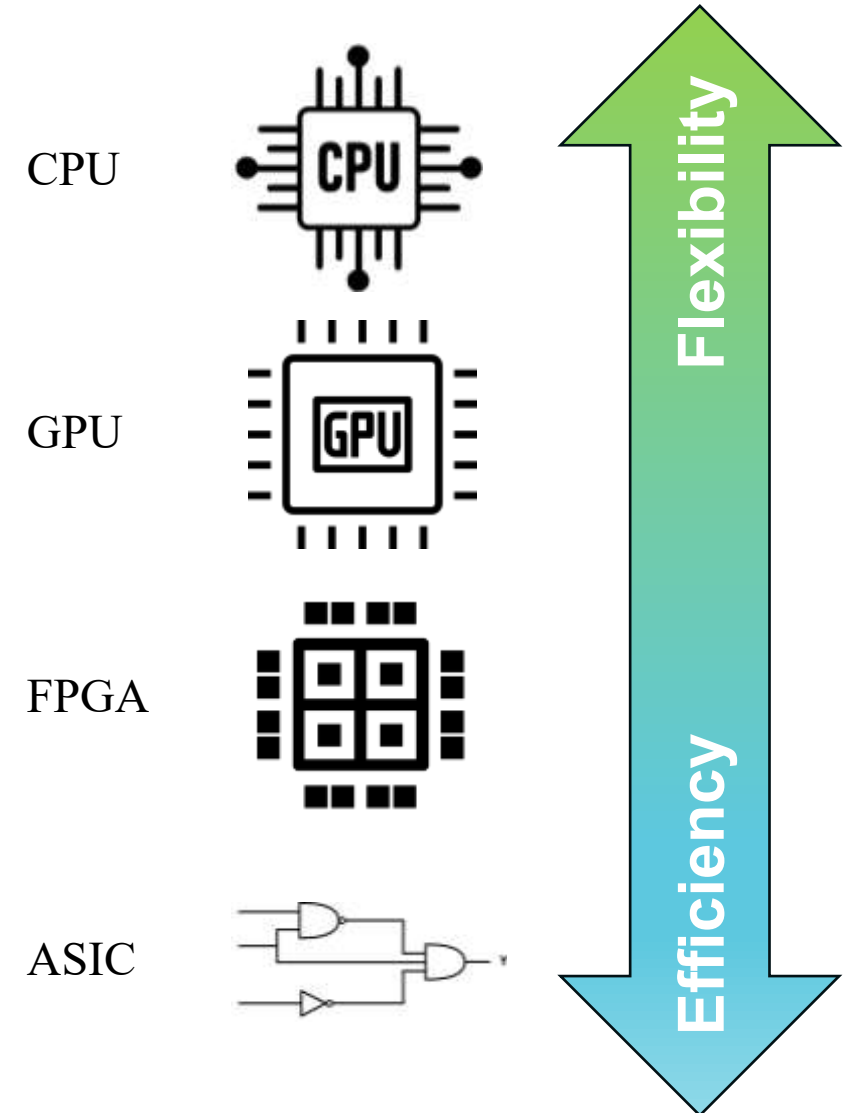


(Roberts et al., 2025a)

Why FPGAs for High-Rate Inference

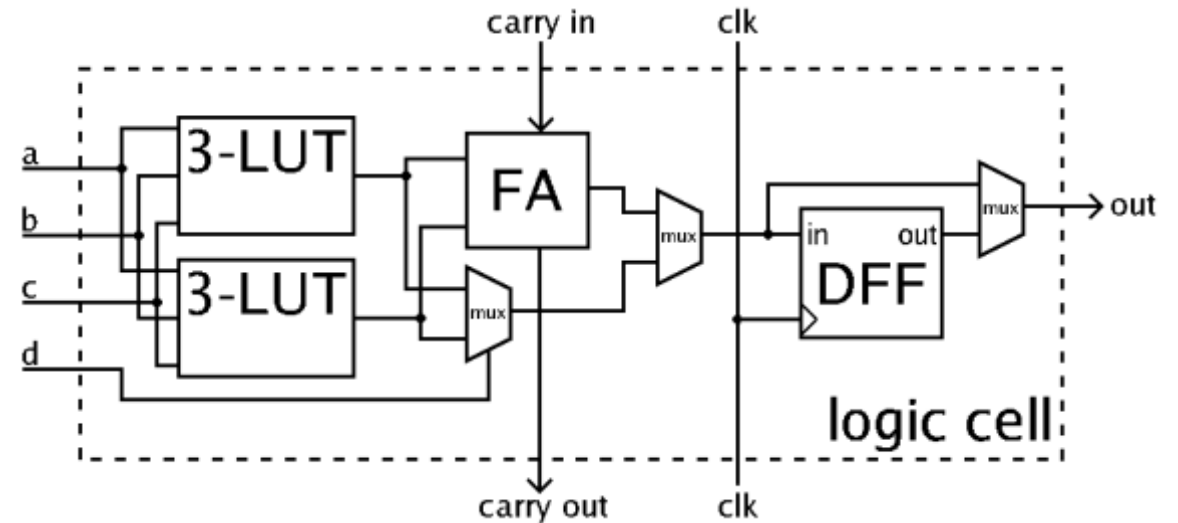
Why FPGAs for High-Rate Inference?

- CPUs are easy to program and well-suited to general-purpose processing, but are too sequential for high-rate inference.
- GPUs add substantial parallelism, but their performance and latency depend on scheduling and memory movement.
- ASICs offer excellent efficiency and deterministic timing, but are expensive and difficult to redesign.
- FPGAs provide reconfigurable, model-specific parallelism with predictable, bounded latency.



FPGAs are Programmable Hardware

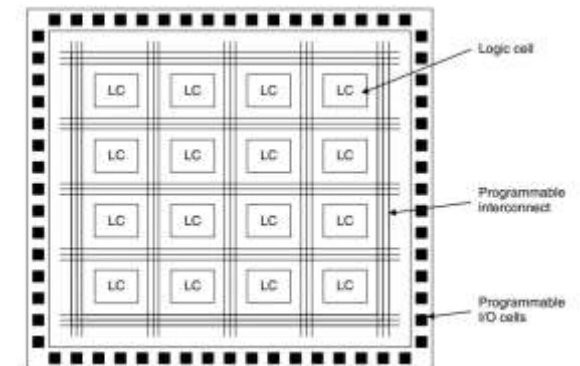
- An FPGA is a grid of logic cells joined by programmable interconnect.
- Each logic cell combines lookup tables (LUTs), a full adder, and a flip-flop register.
- Dedicated blocks add DSP slices for multiply-accumulate and BRAM for on-chip storage.
- An algorithm is built as a fixed circuit, so latency is set by the clock, not a scheduler.
- Resource budgets in slices, LUTs, registers, DSPs, and BRAM limit model size.



Petter.kallstrom, Public domain, via Wikimedia Commons



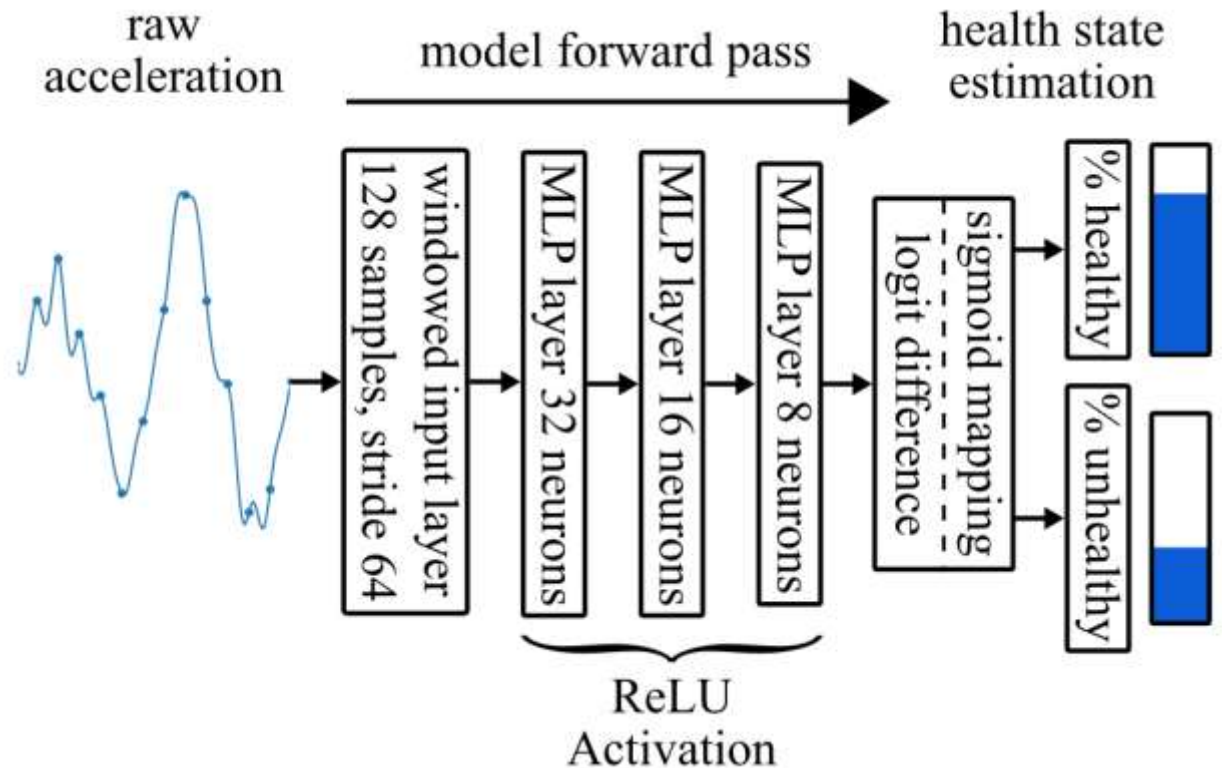
Kintex-7 K70T-2C



VLSI Master, "Introduction to FPGA & CPLD

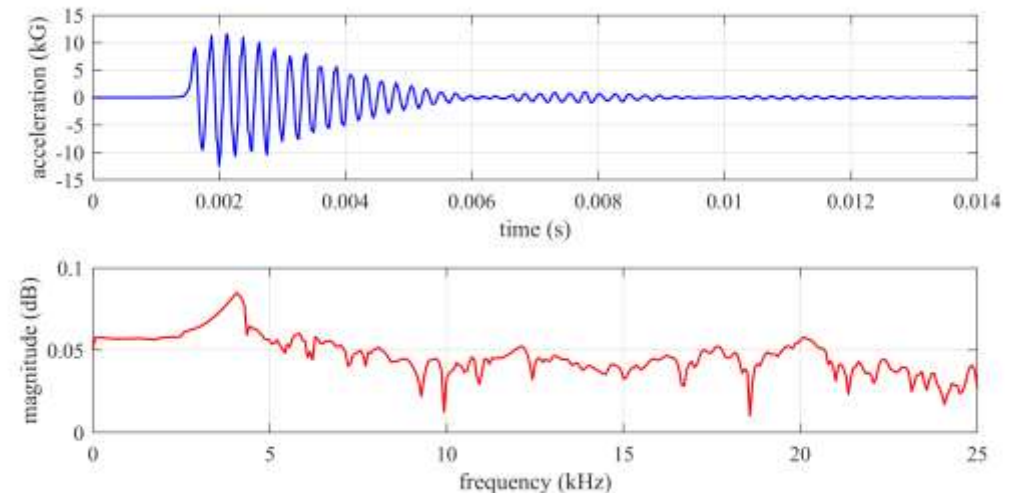
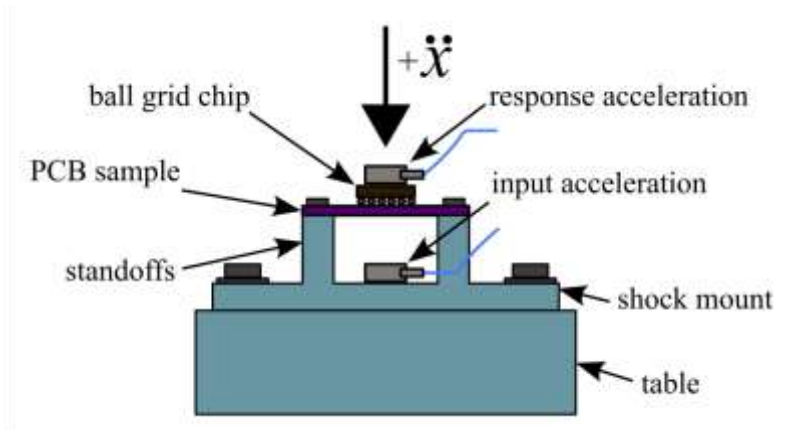
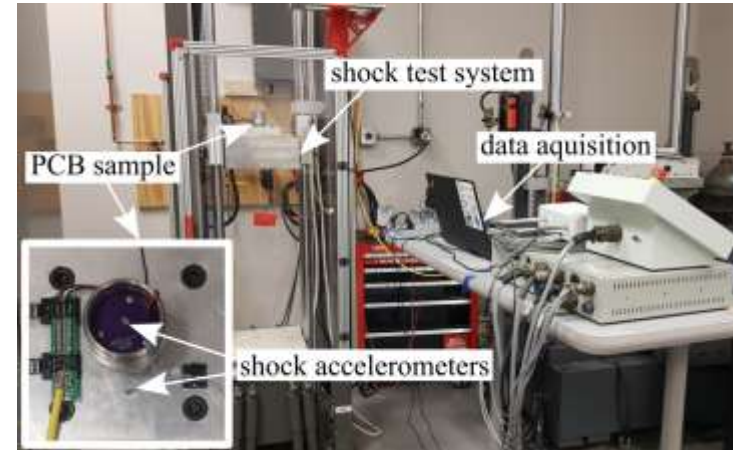
From Shock Response to Health State

- Objective: an FPGA-deployed MLP that estimates PCB health state from shock-response acceleration.
- Raw acceleration is windowed and passed through a single forward pass
- Output is a probability of healthy versus unhealthy state.
- Requirement: accurate impact-level classification.
- Requirement: deterministic, sub-millisecond inference.
- Requirement: efficient use of FPGA resources.



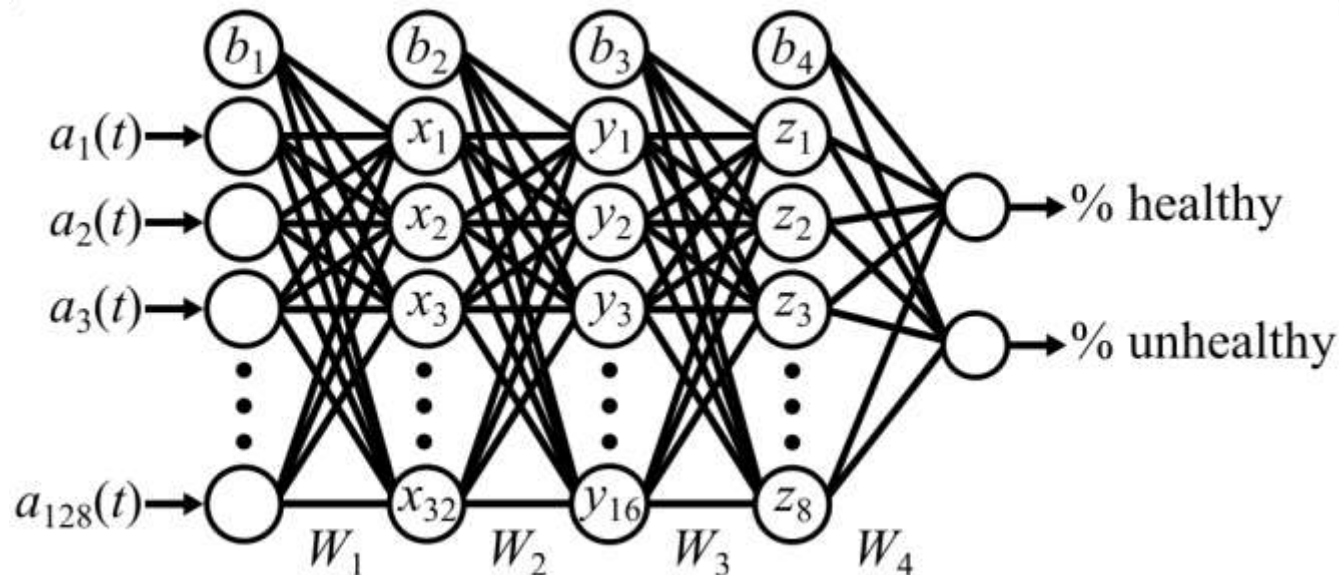
A Repeated-Shock PCB Testbed

- PCB response acceleration measured during controlled repeated-shock testing.
- 30 impacts, sampled at 51.2 kS/s.
- 10 kg_n response with 322 μs half-sine.
- Dominant PCB resonance near 4 kHz.
- Input and response accelerometers on a ball-grid-array test board.



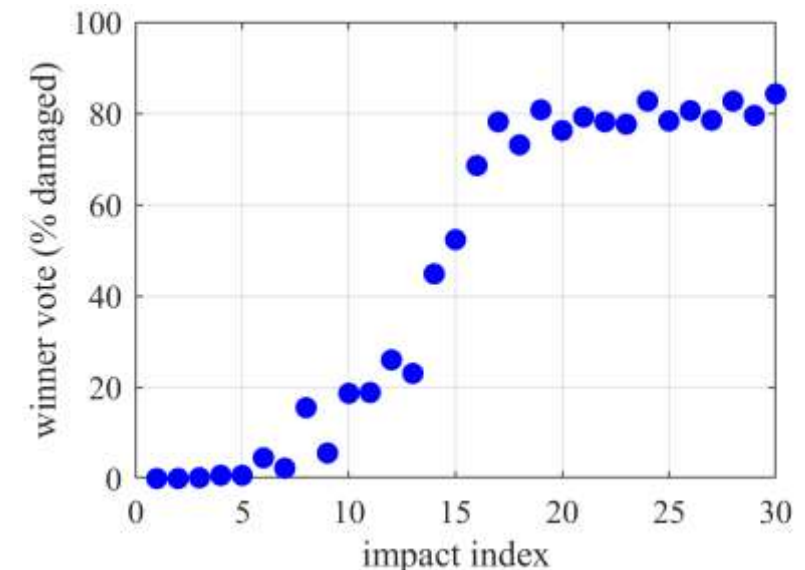
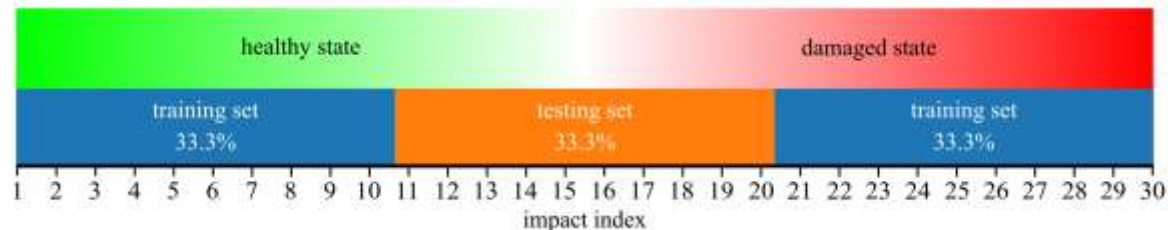
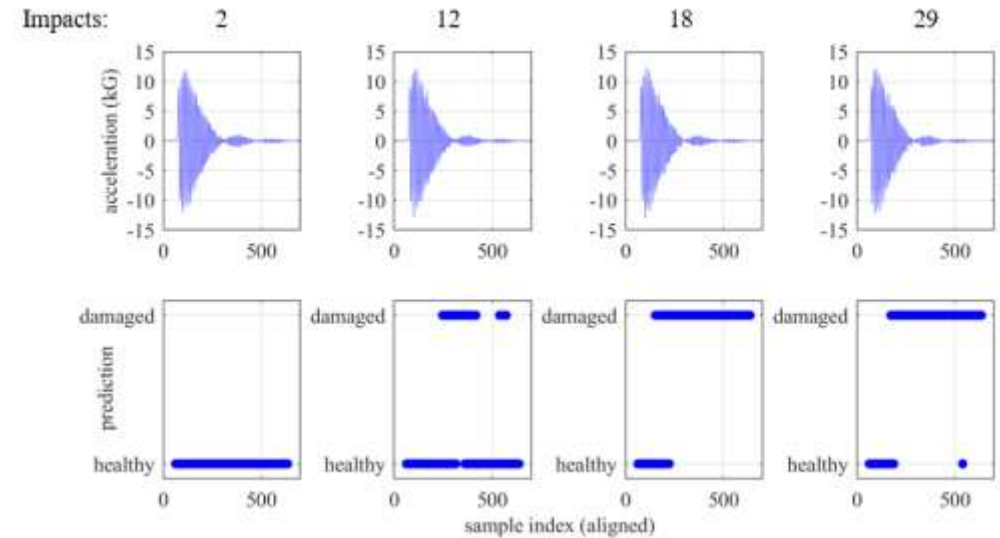
Why a Compact MLP

- A compact MLP maps each acceleration window to a health-state estimate.
- 128-sample windows with a 64-sample stride.
- Three hidden layers of 32, 16, and 8 neurons, with ReLU activation.
- The output-logit difference passes through a sigmoid to give the unhealthy-state probability.



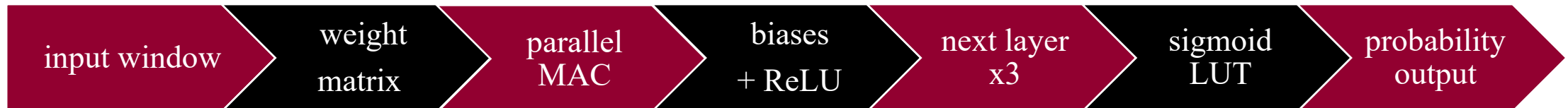
Training on Split Impacts

- Trained offline on chronologically separated impact events.
- Training set: the first 10 healthy and final 10 damaged impacts.
- Validation set: the middle 10 impacts.
- Traces aligned to the PCB-response peak and trimmed to a fixed length.
- Adam optimization with mini-batch updates and early stopping.
- Dropout in the first two hidden layers during training.



Determinism by Design

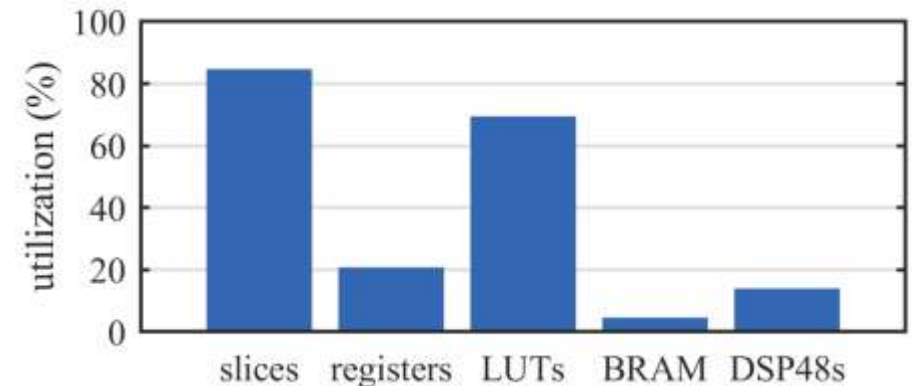
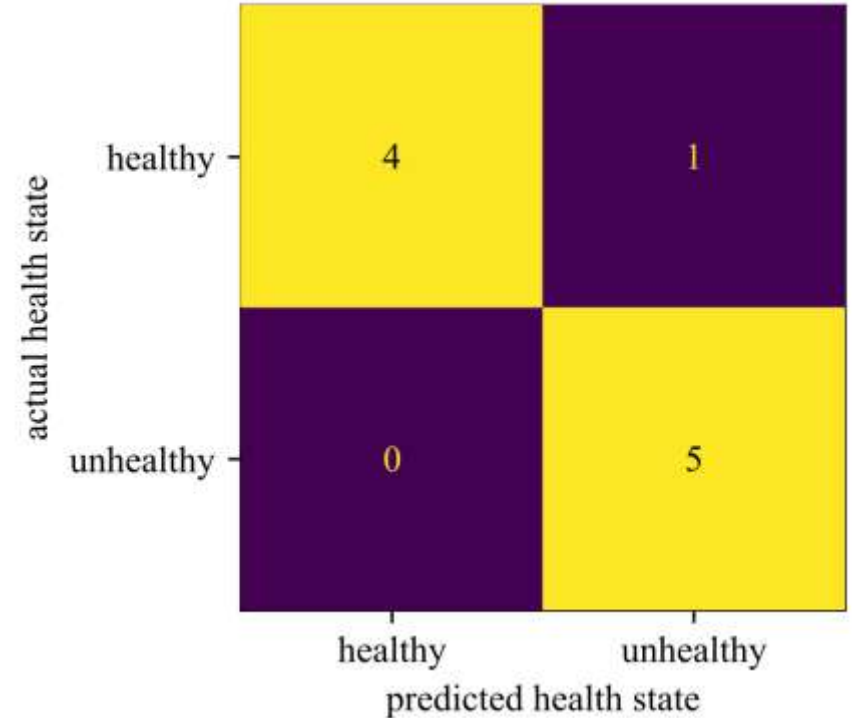
- The FPGA executes one bounded MLP forward pass per input window.
- Transposed weights and biases are stored in on-chip BRAM.
- Fully connected layers use parallel multiply-accumulate operations.
- ReLU activations are stored and passed to the next layer.
- A 512-entry sigmoid lookup table returns the unhealthy-state probability.
- Fixed arithmetic and memory-access sequences give identical execution time on every window.



Accuracy and FPGA Reso

- 90% impact-level validation accuracy (9 of 10 impacts correctly classified).
- 0.609 ms per 128-sample window at 40 MHz.
- Resource use is dominated by slices (84.7%) and LUTs (69.5%).
- Light use of registers (20.8%), DSP48s (14.0%), and BRAM (4.6%).

	slices	registers	LUTs	BRAM	DSP48s
utilized	21479	42230	70488	15	84
total	25350	202800	101400	325	600
percent	84.7%	20.8%	69.5%	4.6%	14.0%



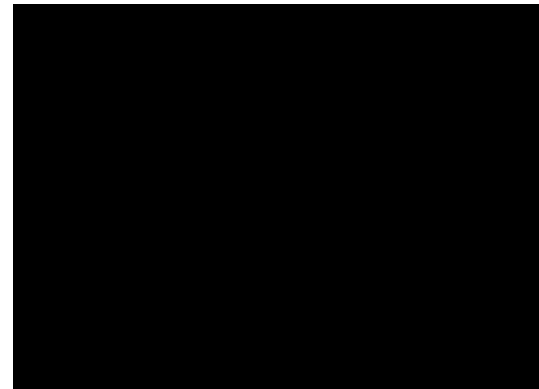
Part 4: Closing Remarks

Why the Edge, Why Now

- AI is moving onto devices: local LLMs, on-device inference, edge accelerators.
- Data sovereignty and autonomy keep data where it is collected.
- SWaP constraints turn edge computing into a rich research problem.
- Open directions: AI at the edge, hardware integration, autonomous systems, distributed intelligence.
- Civil infrastructure is an ideal proving ground: long-lived, distributed, and remote.



VTrans Rail Bridge Inspection Public Domain



Discussion

ARTS-Lab

github.com/Arts-laboratory



Name: Austin Downey
Title: Associate Professor
Email: austindowney@sc.edu



**Molinaroli College of
Engineering and Computing**
UNIVERSITY OF SOUTH CAROLINA

References

- Peter Avitabile. Twenty years of structural dynamic modification – a review. *Sound and Vibration*, 37(1):14-25, 2003.
- Jacob Dodson, Austin Downey, Simon Laflamme, Michael Todd, Adriane G. Moura, Yang Wang, Zhu Mao, Peter Avitabile, and Erik Blasch. High-Rate Structural Health Monitoring and Prognostics: An Overview. *Data Science in Engineering, Volume 9*, Proceedings of the 39th IMAC, A Conference and Exposition on Structural Dynamics 2021, Springer International Publishing, p. 213-217, Oct 2021. doi:10.1007/978-3-030-76004-5_23
- Austin Downey, Jonathan Hong, Jacob Dodson, Michael Carroll, and James Scheppepegrell. Millisecond Model Updating for Structures Experiencing Unmodeled High-rate Dynamic Events. *Mechanical Systems and Signal Processing*, vol. 138, 2020, p. 106551, doi:10.1016/j.ymssp.2019.106551.
- Austin Downey and Jason D. Bakos. High-Rate Structural Health Monitoring: Part-II Embedded System Design, pages 105-108. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-34946-1_12
- Matthew Nelson, Simon Laflamme, Chao Hu, Adriane G Moura, Jonathan Hong, Austin Downey, Peter Lander, Yang Wang, Erik Blasch, and Jacob Dodson. Generated datasets from dynamic reproduction of projectiles in ballistic environments for advanced research (DROPBEAR) testbed. *IOP SciNotes*, 3(4):044401, Nov 2022. doi:10.1088/2633-1357/aca0d2
- Emmanuel A. Ogunniyi, Claire Drnek, Seong Hyeon Hong, Austin R.J. Downey, Yi Wang, Jason D. Bakos, Peter Avitabile, and Jacob Dodson. Real-time structural model updating using local eigenvalue modification procedure for applications in high-rate dynamic events. *Mechanical Systems and Signal Processing*, 195:110318, Jul 2023. doi:10.1016/j.ymssp.2023.110318
- Trotter Roberts, Matthew Burnett, Ryan Yount, Austin R. J. Downey, Adriane Moura, and Fahad Abumohaimed. Adaptive shock testing and failure prediction of printed circuit boards using kriging-based experimental design. In H. Felix Wu, editor, *Nondestructive Characterization and Monitoring of Advanced Materials, Aerospace, Civil Infrastructure, and Transportation XX*, page 2. SPIE, April 2026. doi:10.1117/12.3090164
- Trotter Roberts, Joud N. Satme, Tiffany Yu, Austin R.J. Downey, Rasha E. Karakchi, and Jason D. Bakos. High-rate state estimation of a printed circuit board under shock using an FPGA-deployed neural network. In *Proceedings of the ASME 2026 Conference on Smart Materials, Adaptive Structures and Intelligent Systems*, Albuquerque, NM, USA, 2026. Peer-reviewed conference paper, final manuscript submitted. Paper No. SMASIS2026-191347
- Joud N. Satme, Daniel Coble, Hung-Tien Huang, Austin R. J. Downey, and Jason D. Bakos. Non-linear vibration signal compensation technique for UAV-deployable sensor packages with edge computing. In Zhongqing Su, Maria Pina Limongelli, and Branko Glisic, editors, *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2023*. SPIE, apr 2023. doi:10.1117/12.2658563
- Joud N. Satme, Ryan Yount, Nikita Goujevskii, Luke Jannazzo, and Austin R. J. Downey. Sensor package deployment and recovery cone with integrated video streaming for rapid structural health monitoring. In *ASME 2024 Conference on Smart Materials, Adaptive Structures and Intelligent Systems*, SMASIS2024. American Society of Mechanical Engineers, September 2024. doi:10.1115/smasis2024-140435
- Ryan Yount, Trotter Roberts, Jacob Dodson, Adriane Moura, and Austin R.J. Downey. Experimental Analysis to Enable Low-Latency Structural Health Monitoring for Electronics in High-Rate Dynamic Environments. *Springer Nature Switzerland*, 2025