

Adaptive Real-Time Systems Laboratory (ARTS-Lab)

An Interdisciplinary Cyber-physical Systems Lab at USC

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



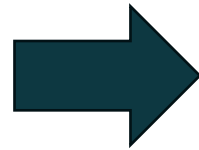
How We See Ourselves

We use

foundational
science



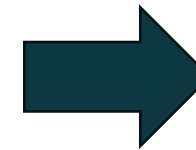
Day School



to develop
essential tools



Dan Thompson



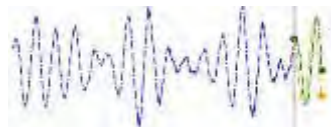
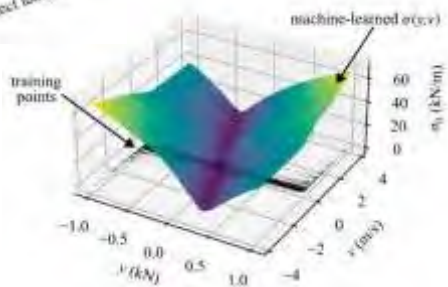
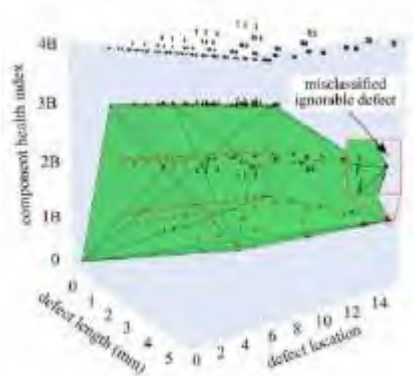
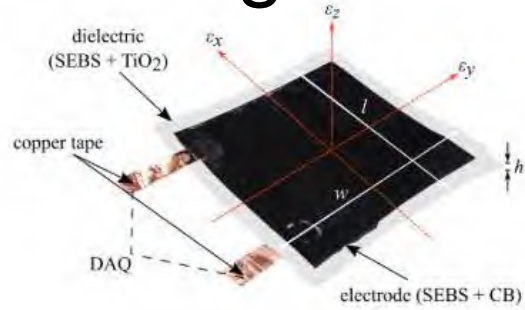
to solve real-world
problems



public domain

**We are Engineers
(mostly)**

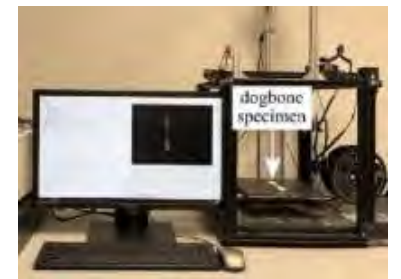
Sensing



AI/ML

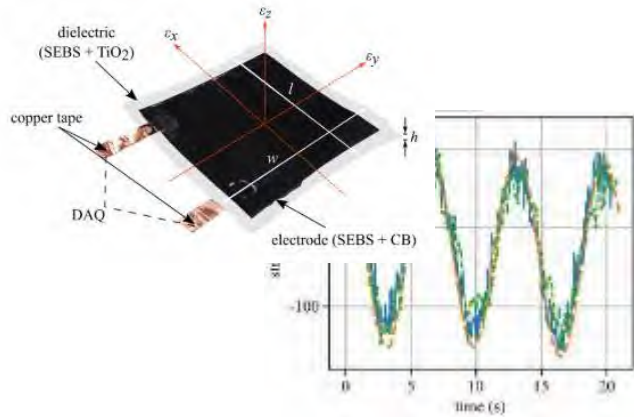


Data Assimilation

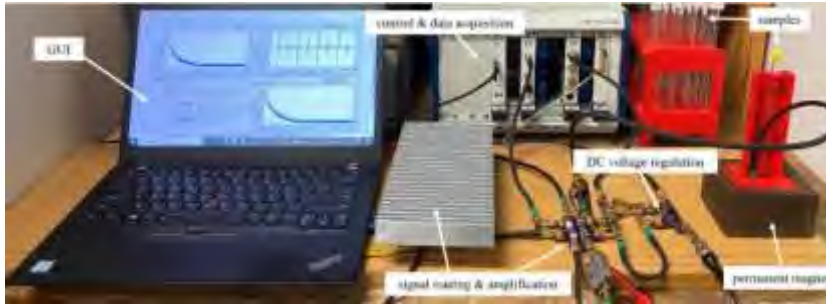
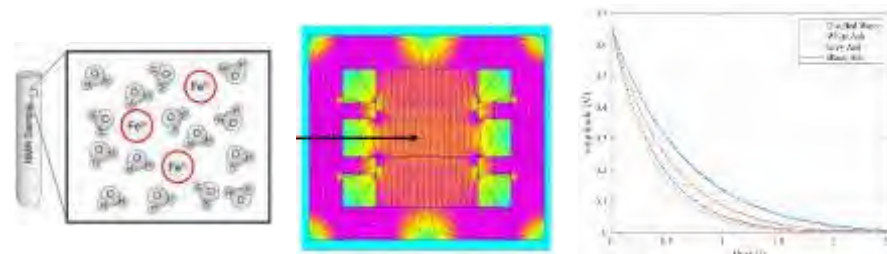


Embedded Systems

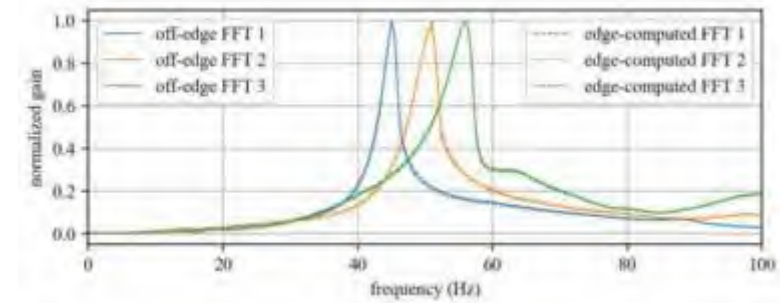
Sensing



Flexible Electronics



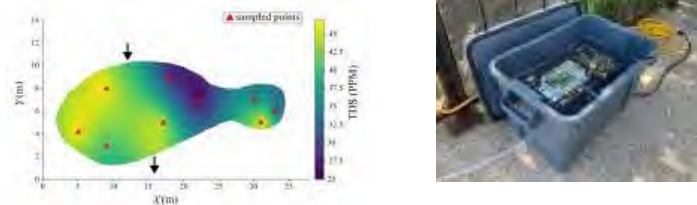
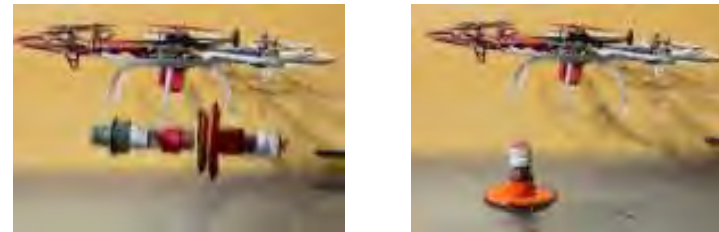
Nuclear Magnetic Resonance



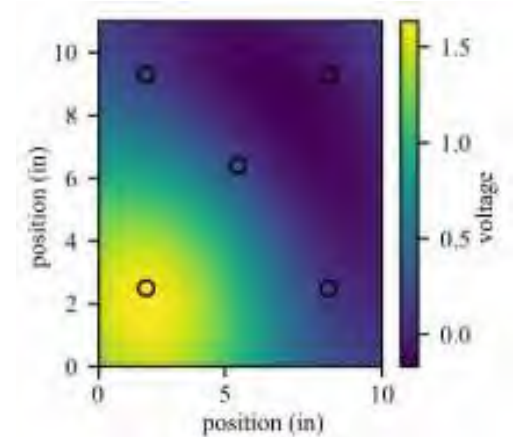
Vibration Sensors



In Situ Monitoring of AM

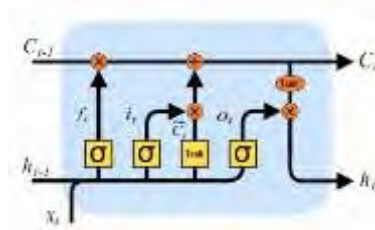
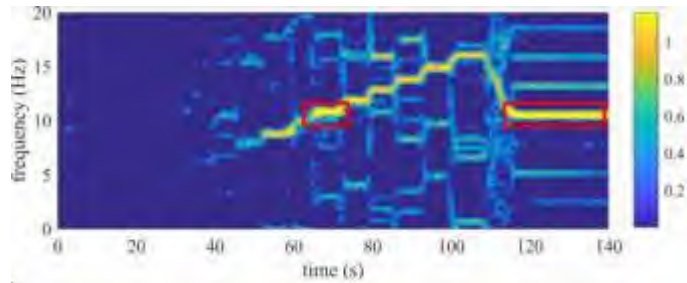


Water Quality Sensors

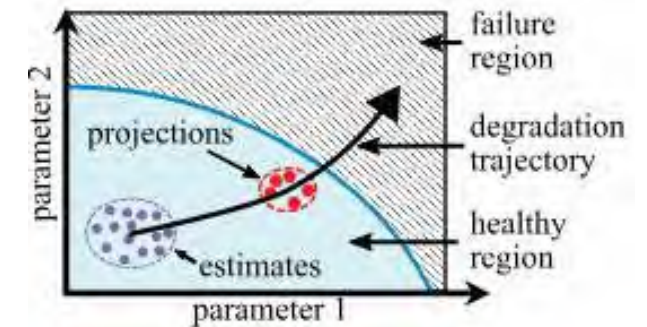
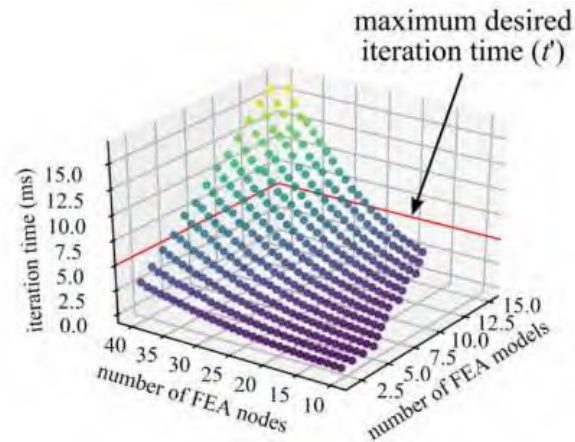
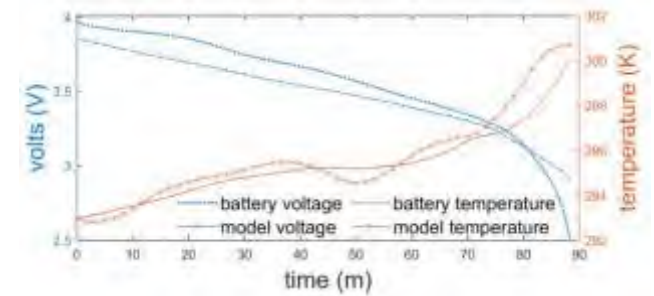


Geo Technical Sensors

Data Assimilation



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

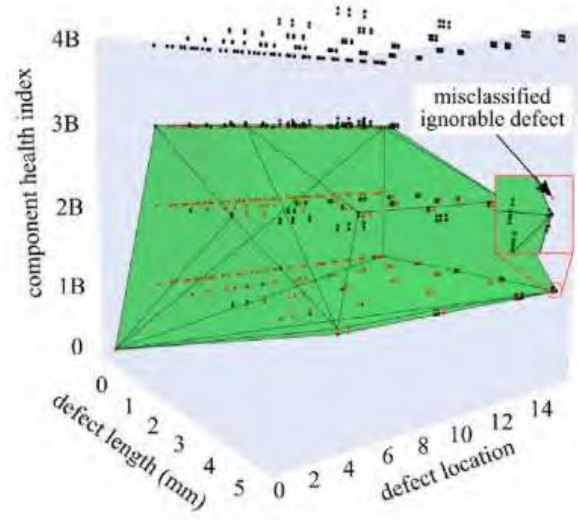
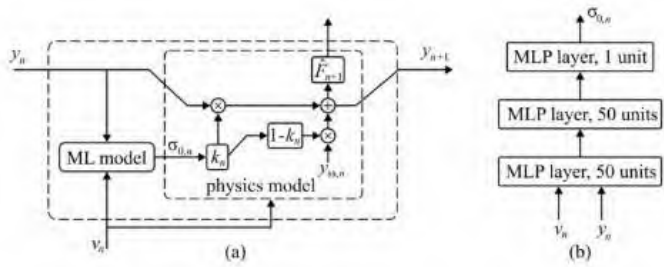


Civil Structures

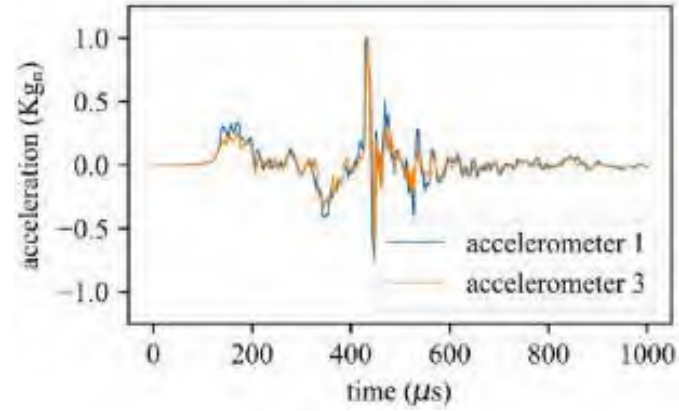
High-Rate Systems

Battery Systems

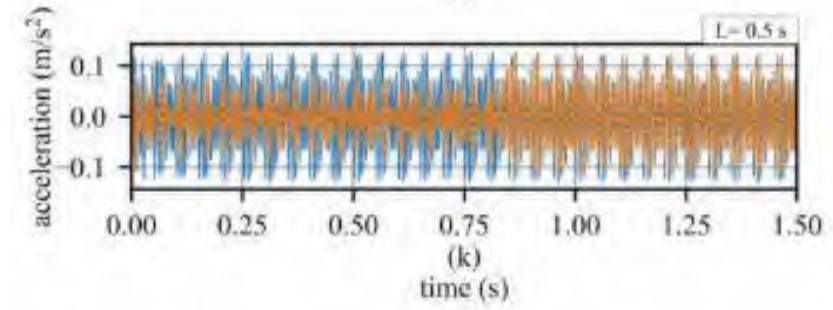
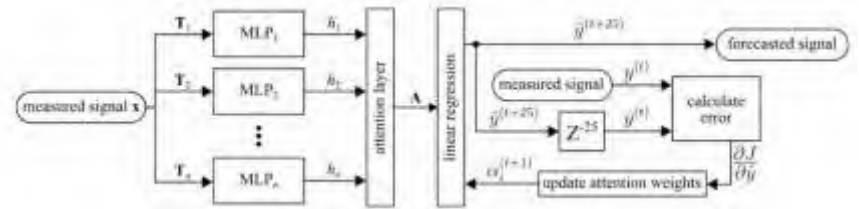
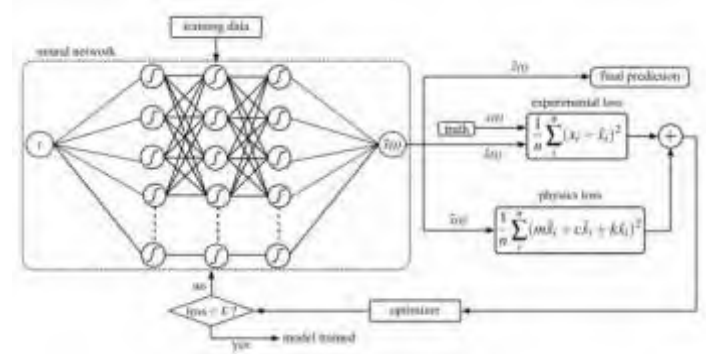
AI/ML



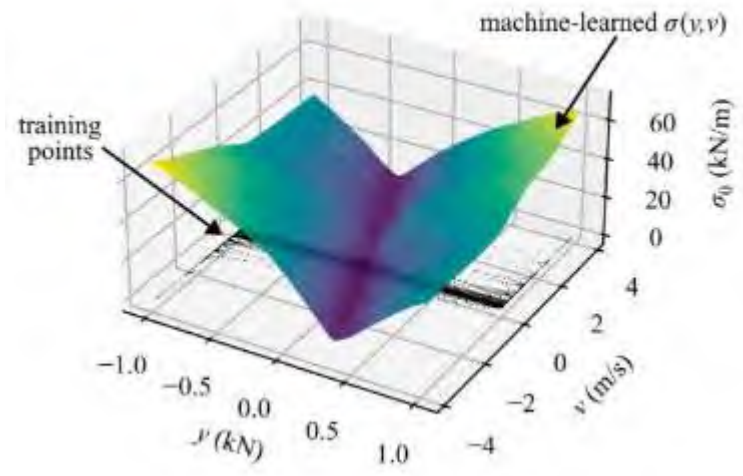
Decision-making



Generative

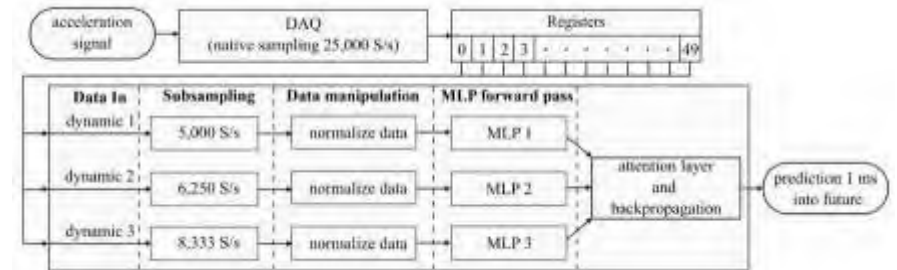
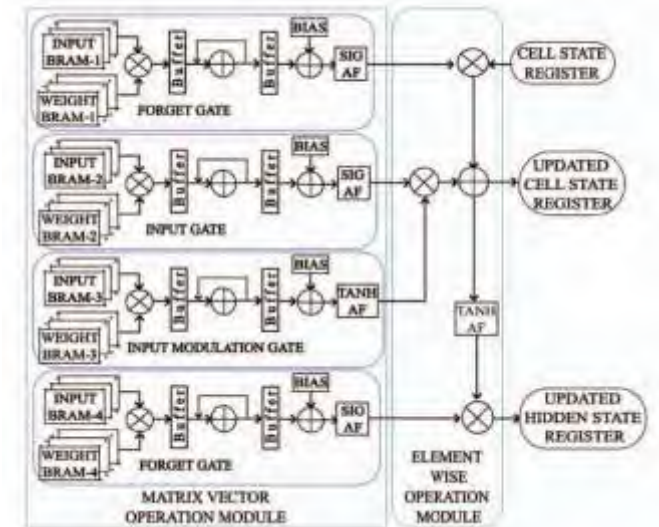
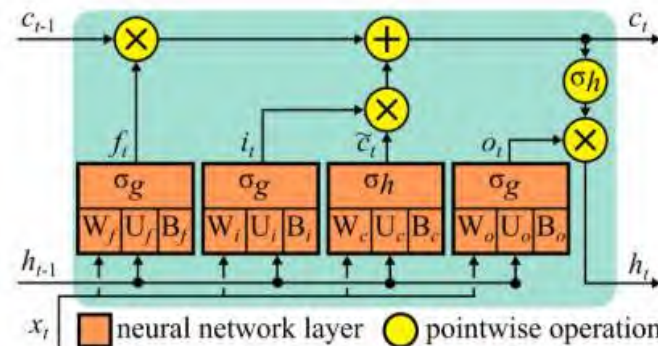
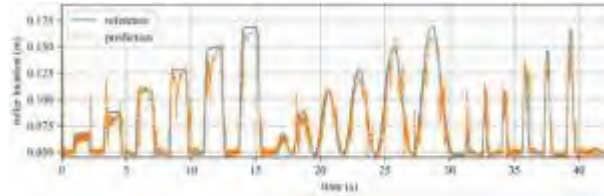
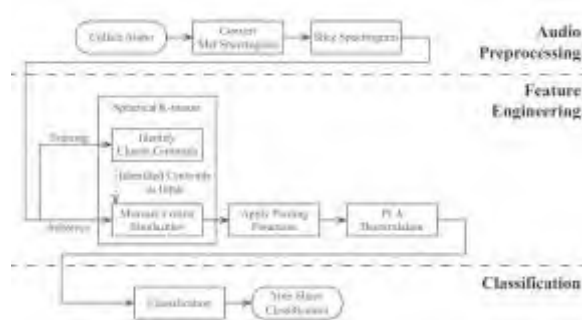


Forecasting



Explainability

Embedded Systems



7 Microcontroller/
microprocessor

Real-Time OS



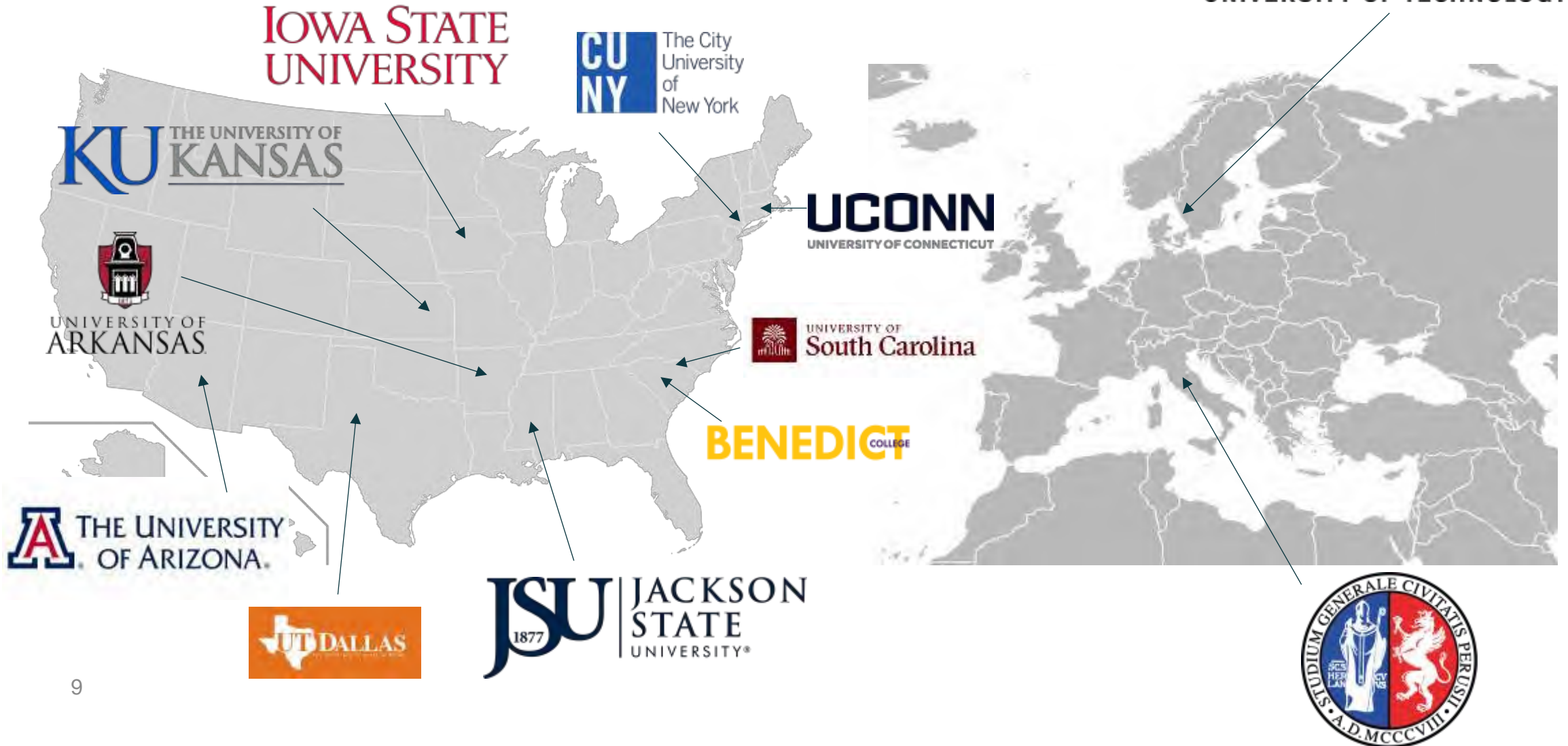
FPGA

Supporting Agencies, Companies, and Partners



Out Academic Partners

CHALMERS
UNIVERSITY OF TECHNOLOGY



The High-rate Challenge

Description of High-rate Dynamics

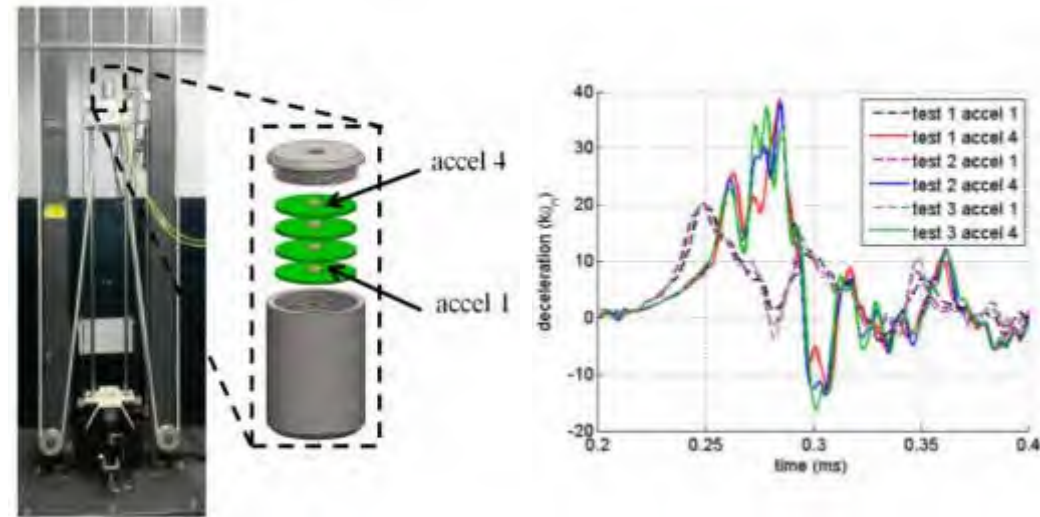
High-rate (<100ms)



High-amplitude (acceleration > 100 g)



The deceleration event in drop tower tests typically lasts for 0.5ms



- Large uncertainties in the external loads.
- High levels of nonstationarity and heavy disturbance.
- Generations of unmodeled dynamics from changes in mechanical configuration.

High-Rate Systems

Hypersonic vehicles



Ballistic packages



Debris approaching space shuttle



Lightning strikes on aircraft



Civil Structures



Fighter jets



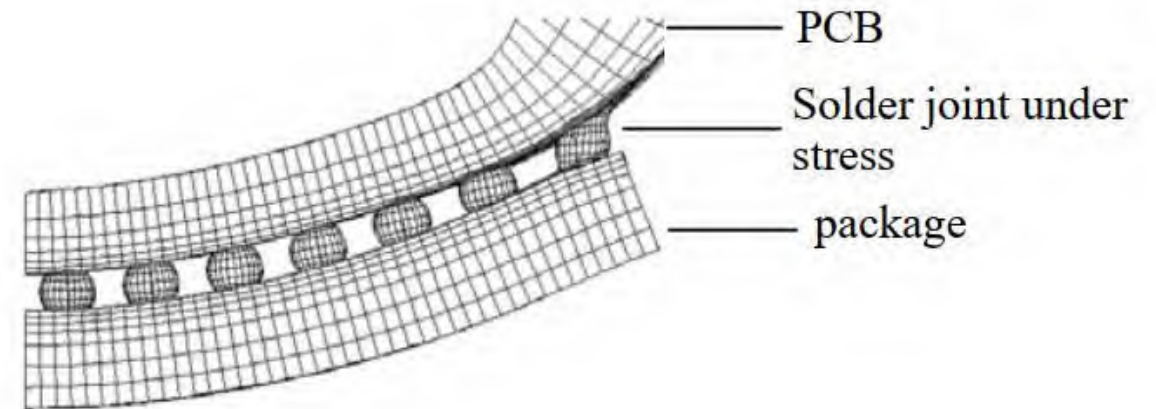
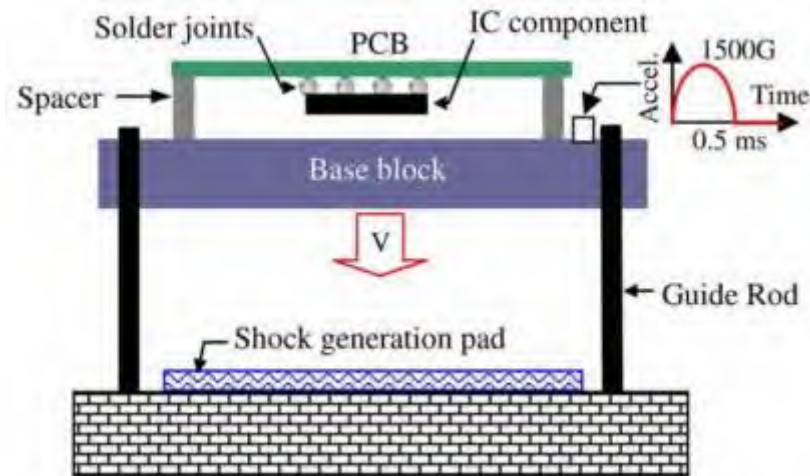
Active High-Rate Systems (Airbags)



Active High-Rate Systems (Electronics)

PCB failures under shock are caused by:

- Bending of the base PCB board, causing stresses to build up at the solder balls.
- Adhesion challenges of masses (components) accelerating away from the PCB.



Data Driven or Physics Based State Estimation



Data Driven or Physics Based State Estimation

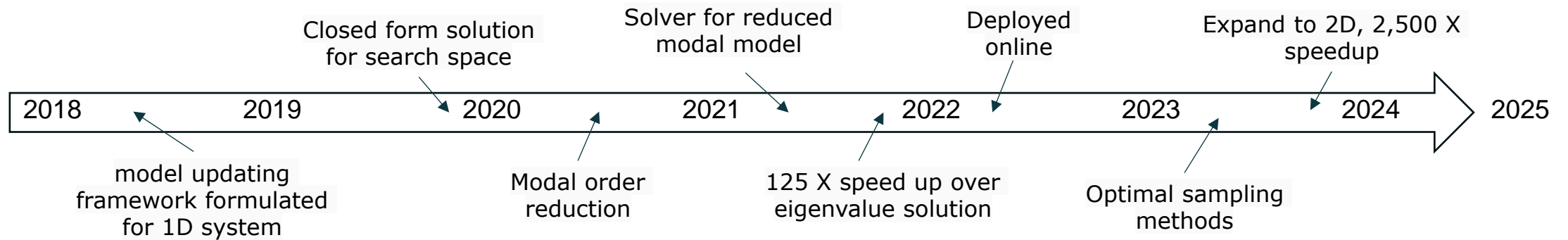
- **Data-driven:**
 - Potential to be faster
 - Easier to implement
 - Students excited to work on it
 - AI/ML is moving quickly
- **Physics-based:**
 - Potential for prognostics
 - Potential for real-time control
 - Better suited for decision-making
 - Better suited for un-foreseen dynamics



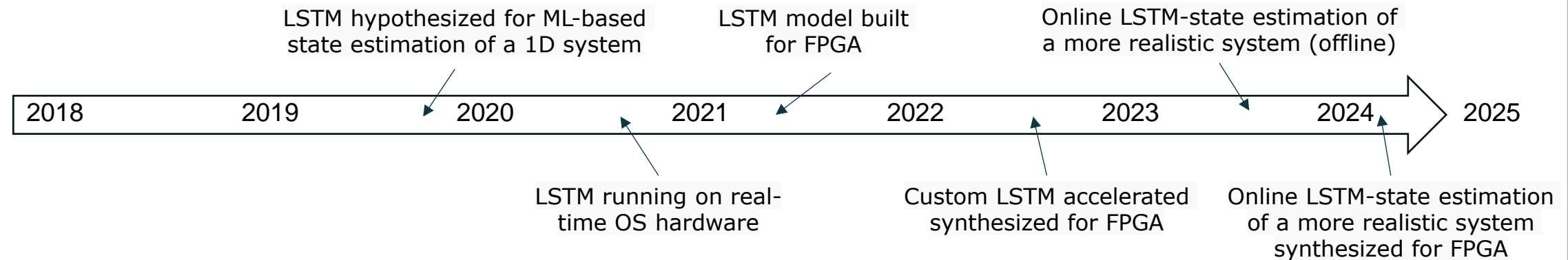
It was hard to decide,
so we did both

Timeline of Efforts on State Estimation

Physics-based



Data-driven



Data Driven Model Updating (Theory and Proof of Concept)

Data Driven Model Updating
(Theory and Proof of Concept)

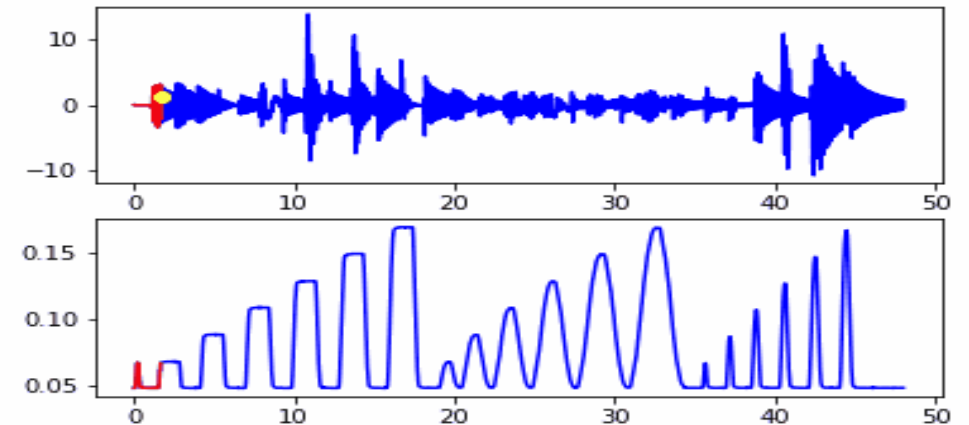
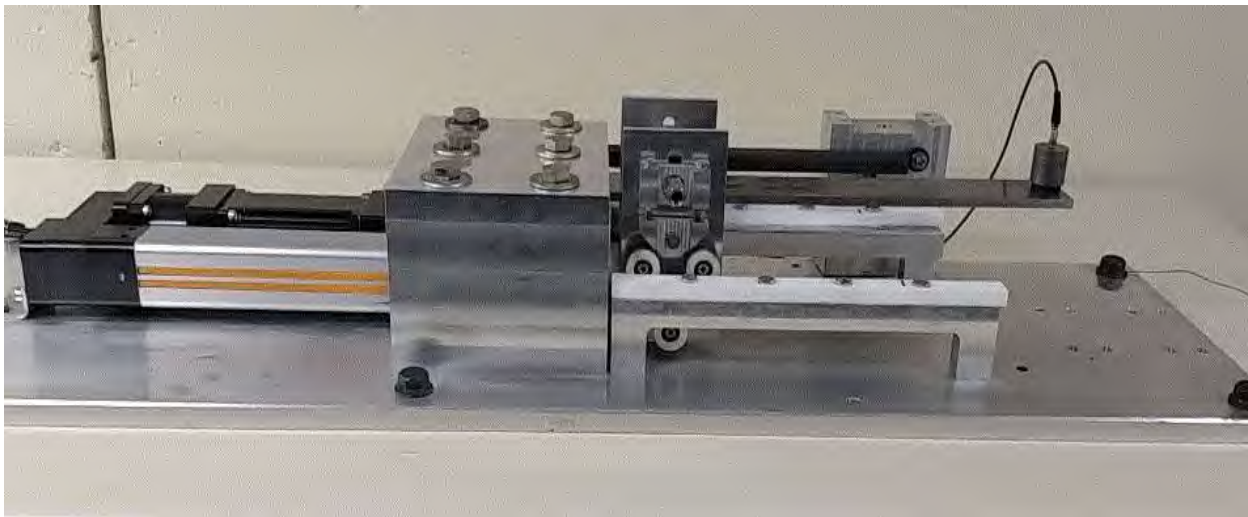
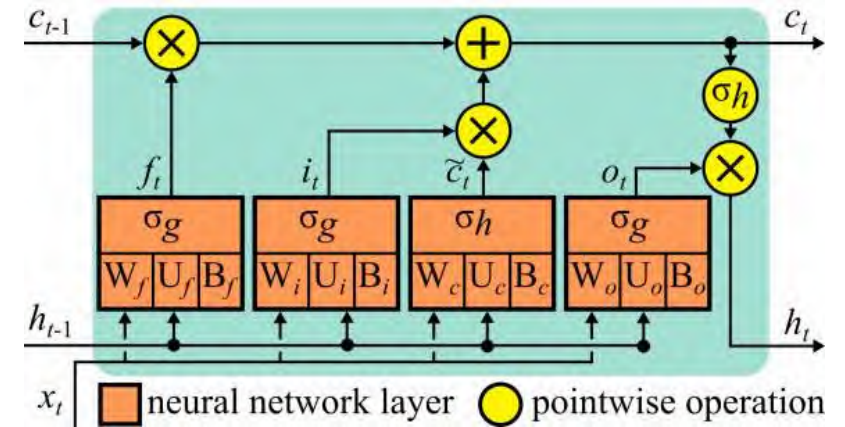
Electronic Components
Under Shock (Application)

FPGA Implementation
(Timing Consideration)

LSTM-based Real-time State Estimation

In this work:

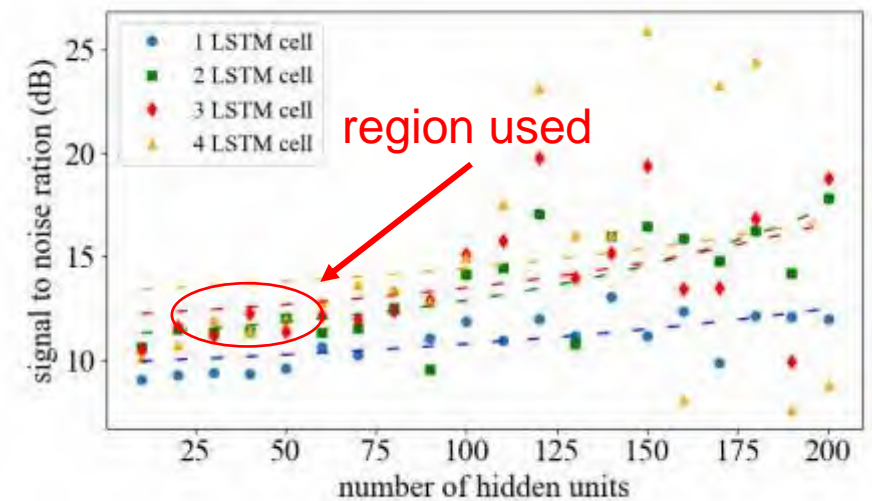
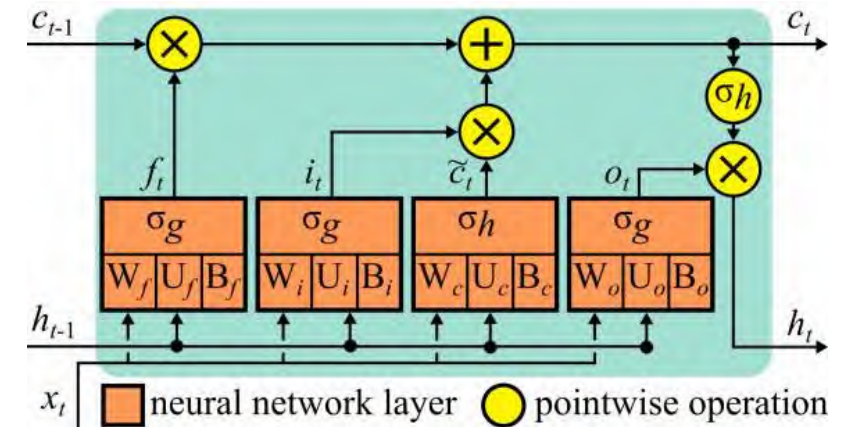
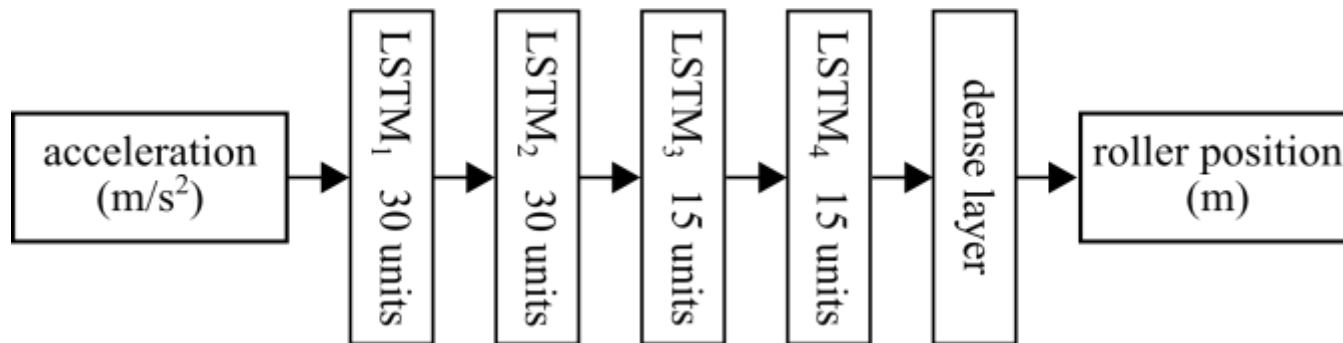
- Long short-term memory (LSTM) models are used for real-time state estimation.
- Experimentally validated on NI-Linux Real-Time.



Long Short-term Memory Model

LSTM features and development:

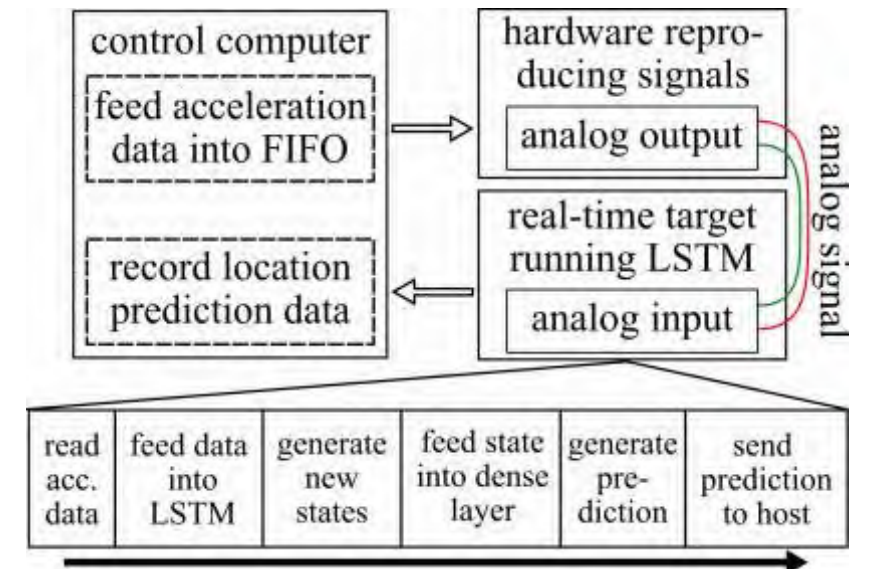
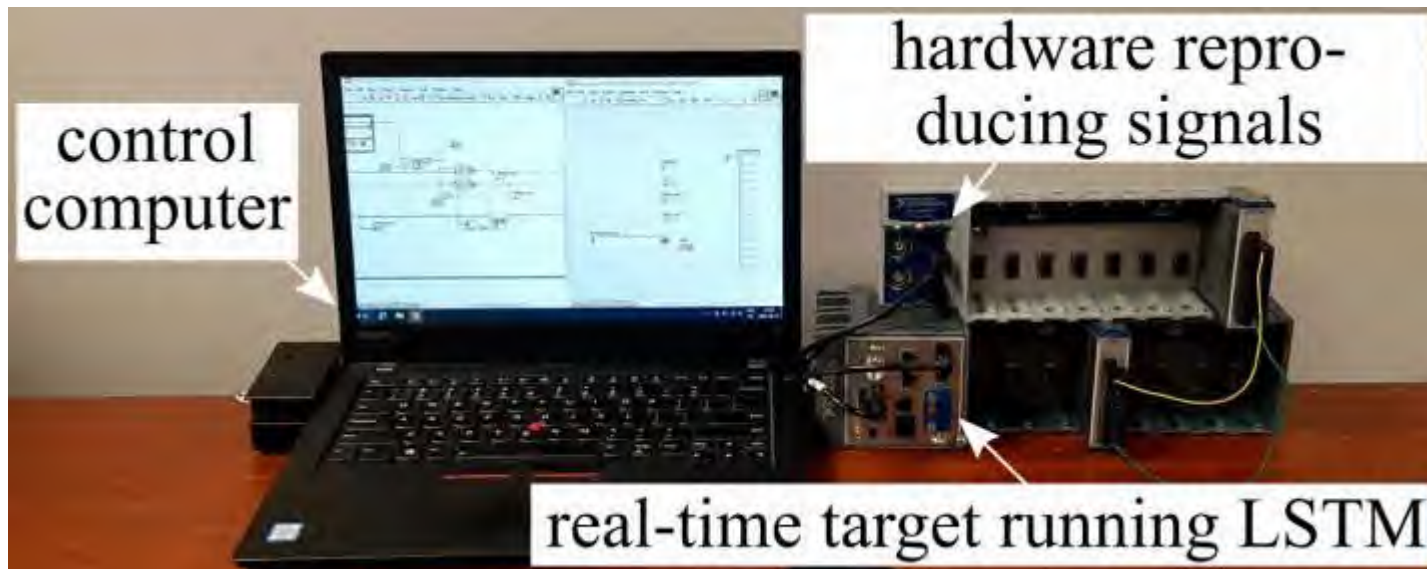
- LSTMs are a Recurrent Neural Network (RNN) that propagates through long- and short-term memory forms.
- Four stacked LSTM cells (30, 30, 15, 15 units) with a fully connected layer at the output.
- LSTM network is trained offline.



Real-time Validation on Embedded Systems

Real-time validation performed on an embedded system running:

- The experimental setup consisted of two subsystems:
 - **Hardware reproducing Signals** reproduces the DROPBEAR dataset using a digital to analog converter.
 - **Real-time Target** digitizes the analog voltage and feeds the input into the LSTM architecture (cRIO-9035).
- Data is sampled at 400 S/s, therefore, a prediction is made every 2.5 ms.
- State predictions are returned via a first-in-first-out (FIFO) buffer to the host PC).



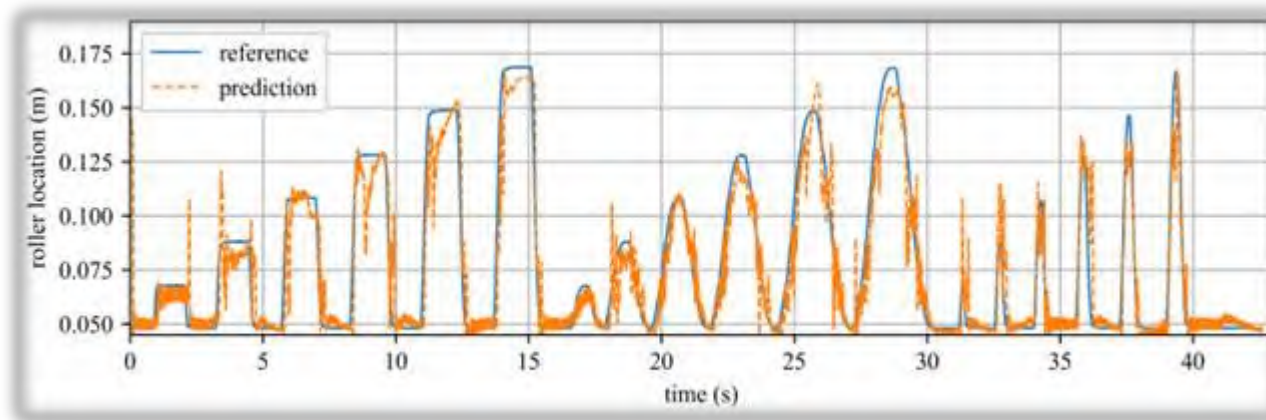
Real-time LSTM Modeling Results

LSTM model performance results:

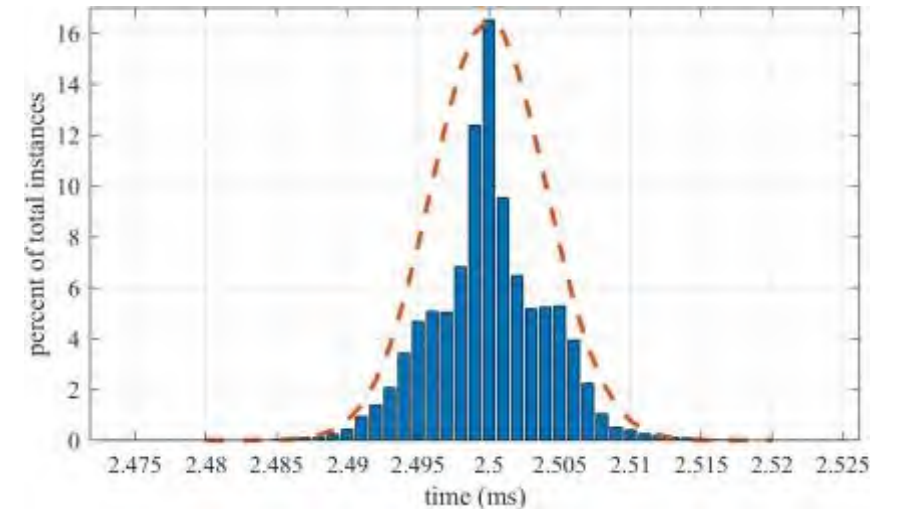
- SNR_{dB} of 43.2 dB.
- RMSE of 12.8 mm.
- LSTM traces reference pin location closely.

Timing accuracy results:

- Execution-time jitter is observed (expected).
- Timing follows a normal distribution.



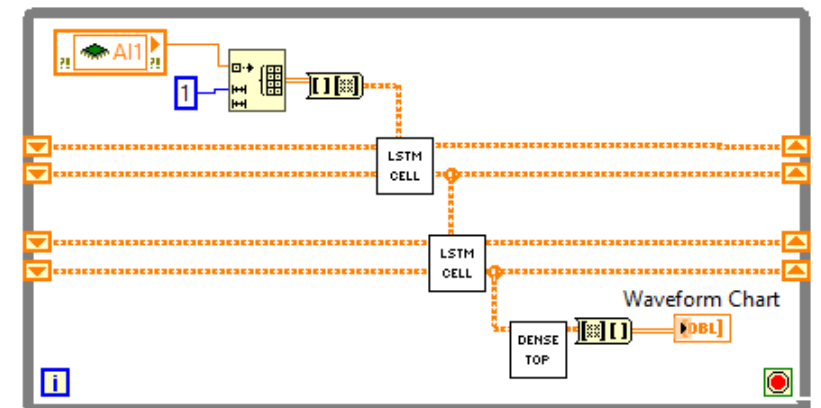
Algorithm Timing



Mean	2.5 ms
Standard deviation	0.004 ms
Max overshoot	0.019 ms

Code, Data, and Resources

- This work will be presented at ASME-IDETC under the title “Progress Towards Data-driven High-rate Structural State Estimation on Edge Computing Devices”.
- Open-Source library for Deploying LSTMs to the NI Linux Real-time Operating System at: <https://github.com/ARTS-Laboratory/LabVIEW-LSTM>
- Code for ASME-IDETC conference paper at: <https://github.com/ARTS-Laboratory/Paper-Progress-towards-data-driven-high-rate-structural-state-estimation-on-edge-computing-devices>
- Dataset available on GitHub at: <https://github.com/High-Rate-SHM-Working-Group/Dataset-2-DROPBEAR-Acceleration-vs-Roller-Displacement>



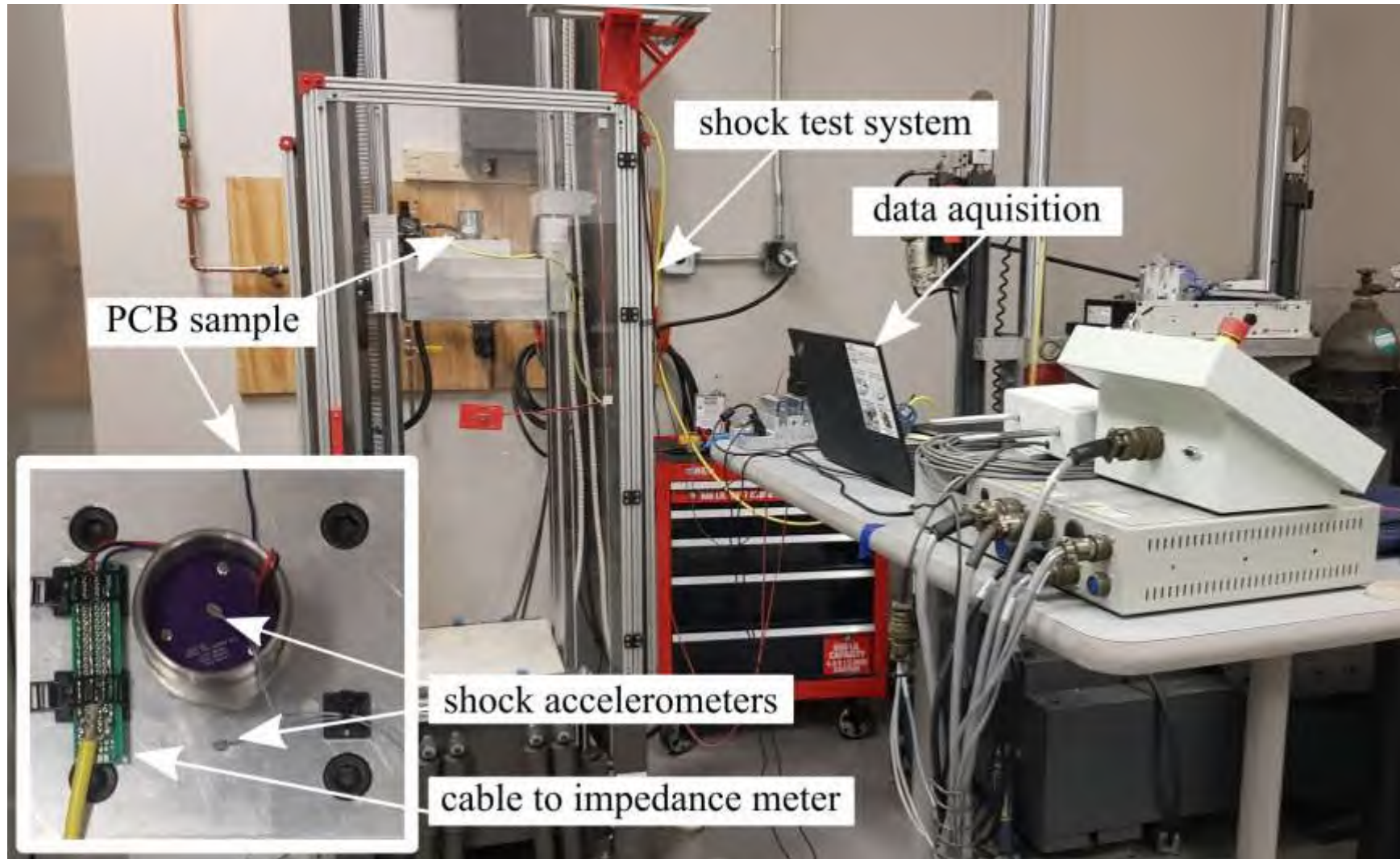
Electronic Components Under Shock (Application)

Data Driven Model Updating
(Theory and Proof of Concept)

Electronic Components
Under Shock (Application)

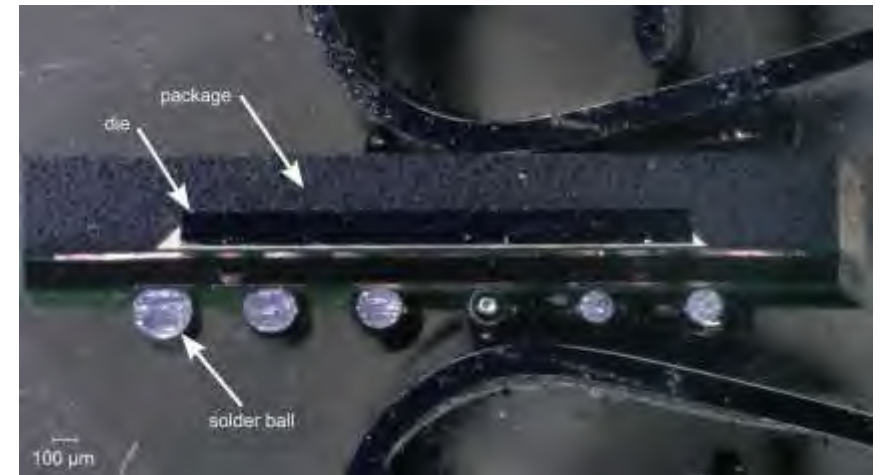
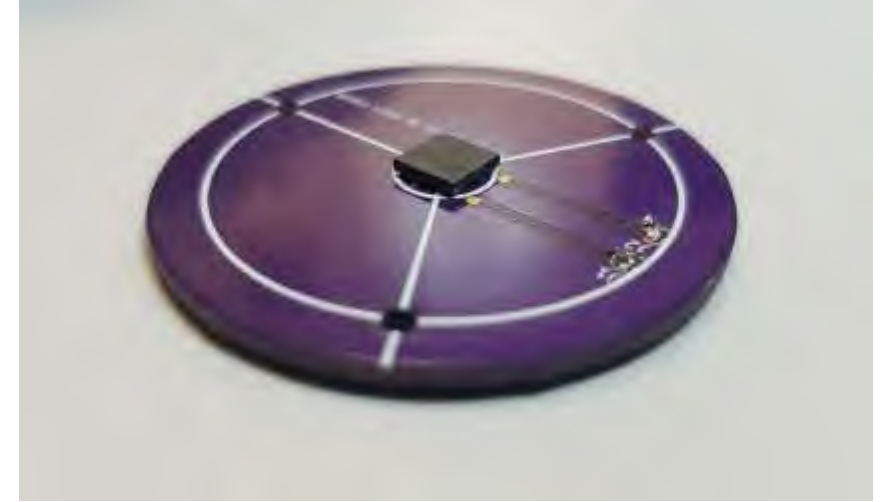
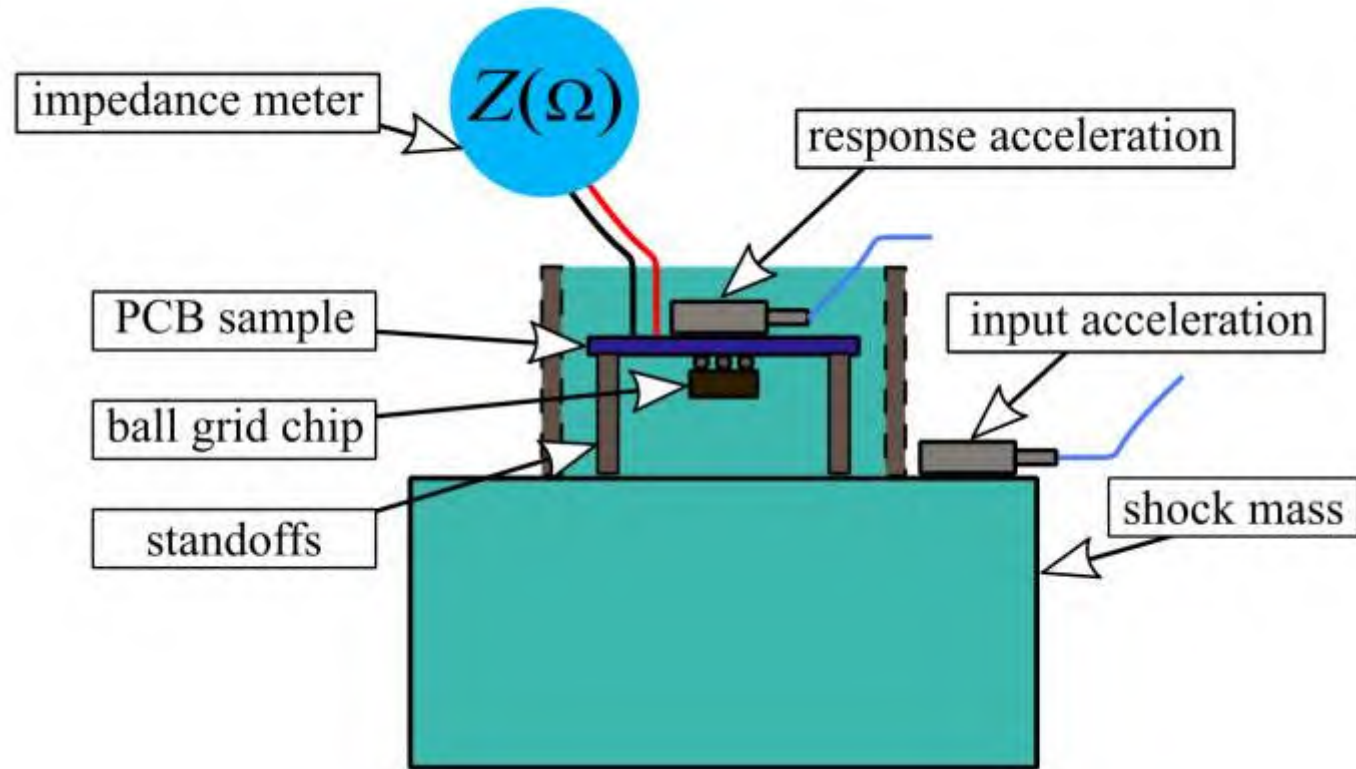
FPGA Implementation
(Timing Consideration)

Experimental System used for Validation

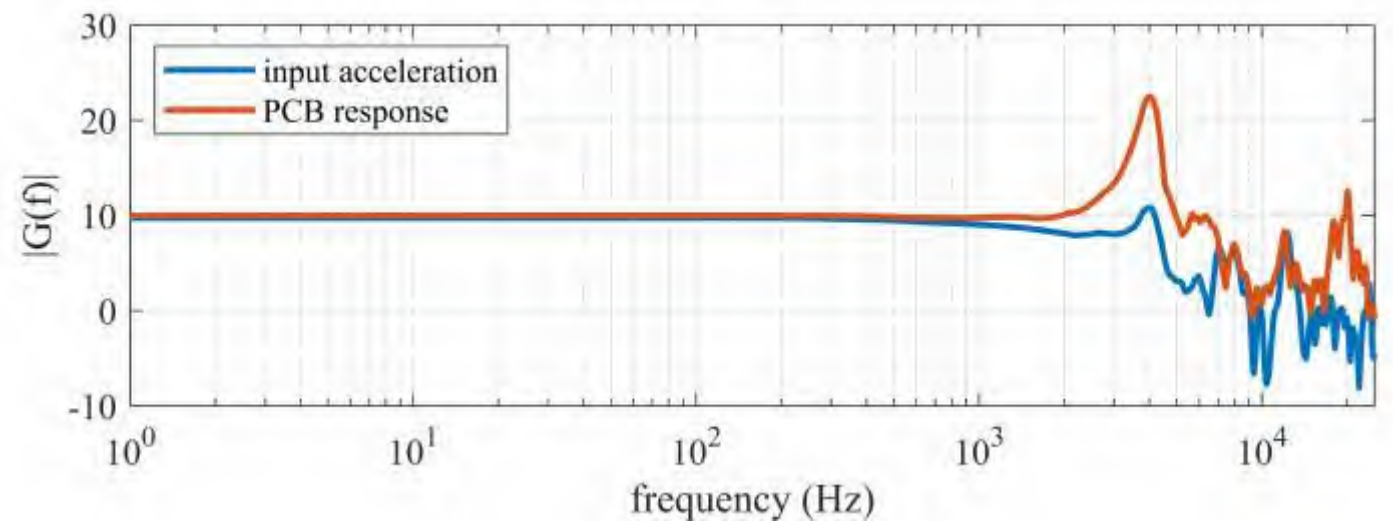
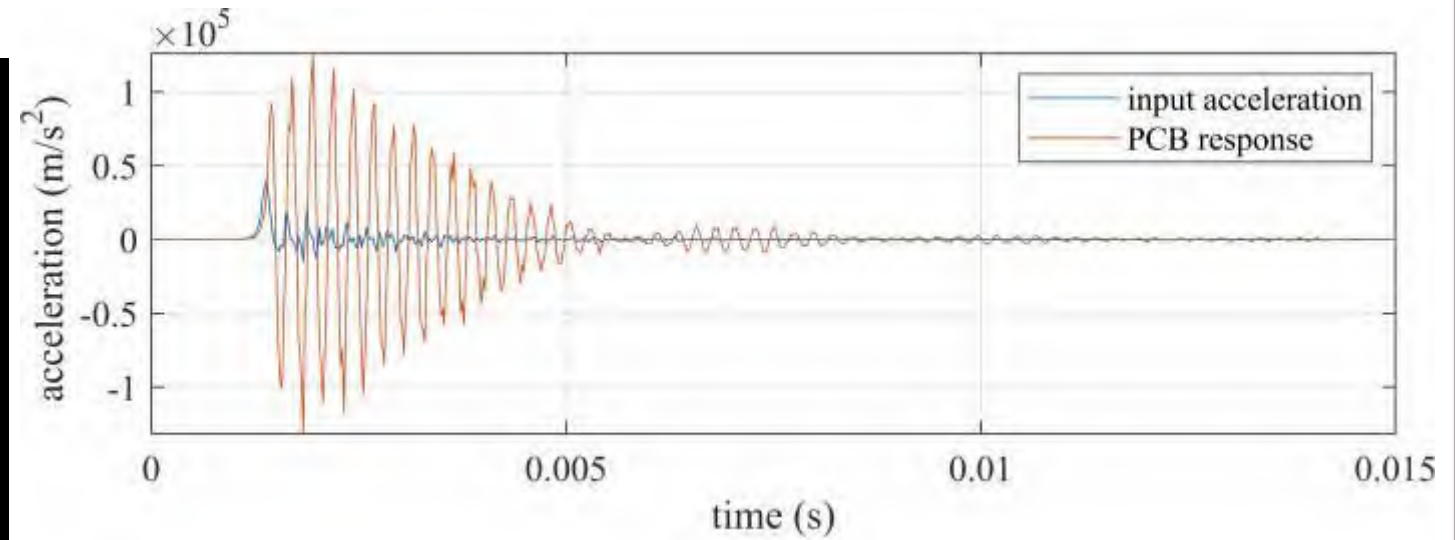
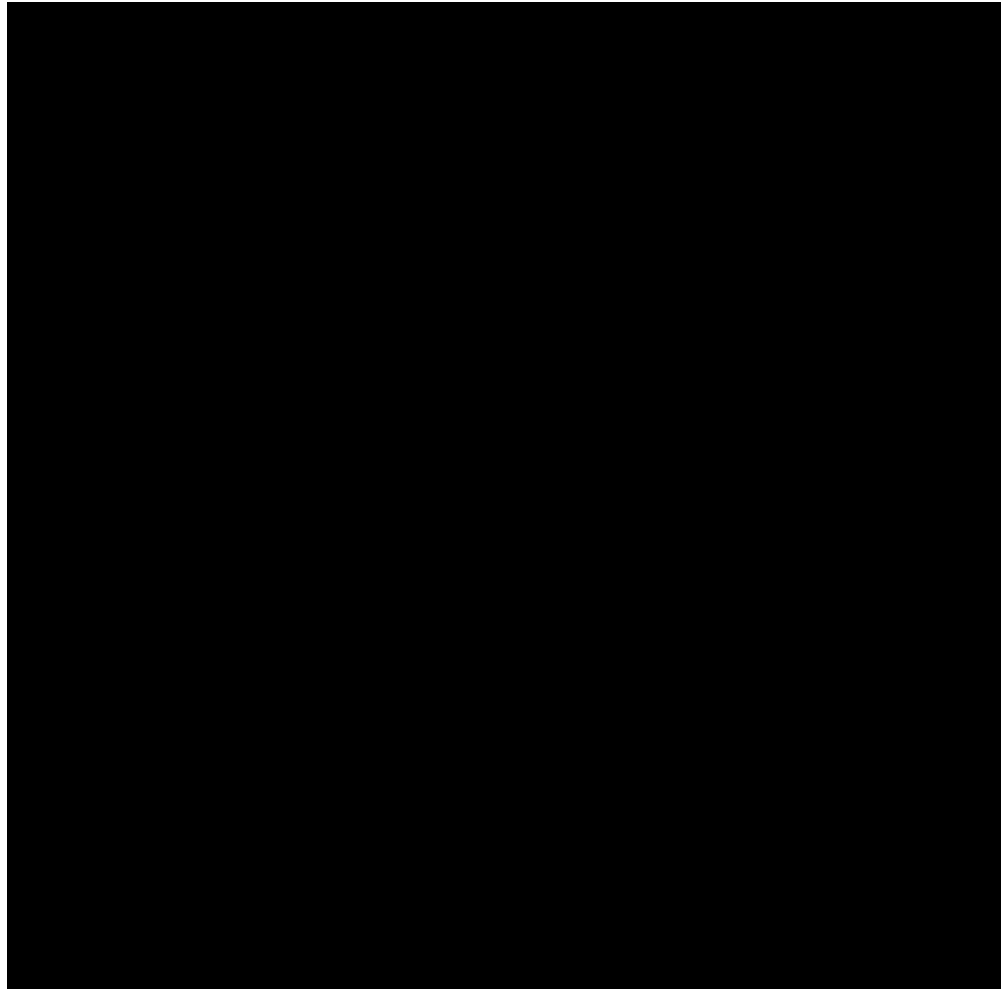


<https://github.com/High-Rate-SHM-Working-Group/Dataset-5-Extended-Impact-Testing>

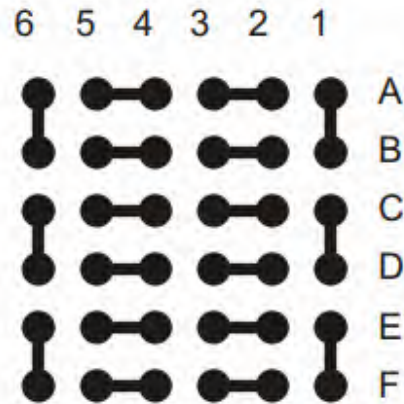
Experimental System used for Validation



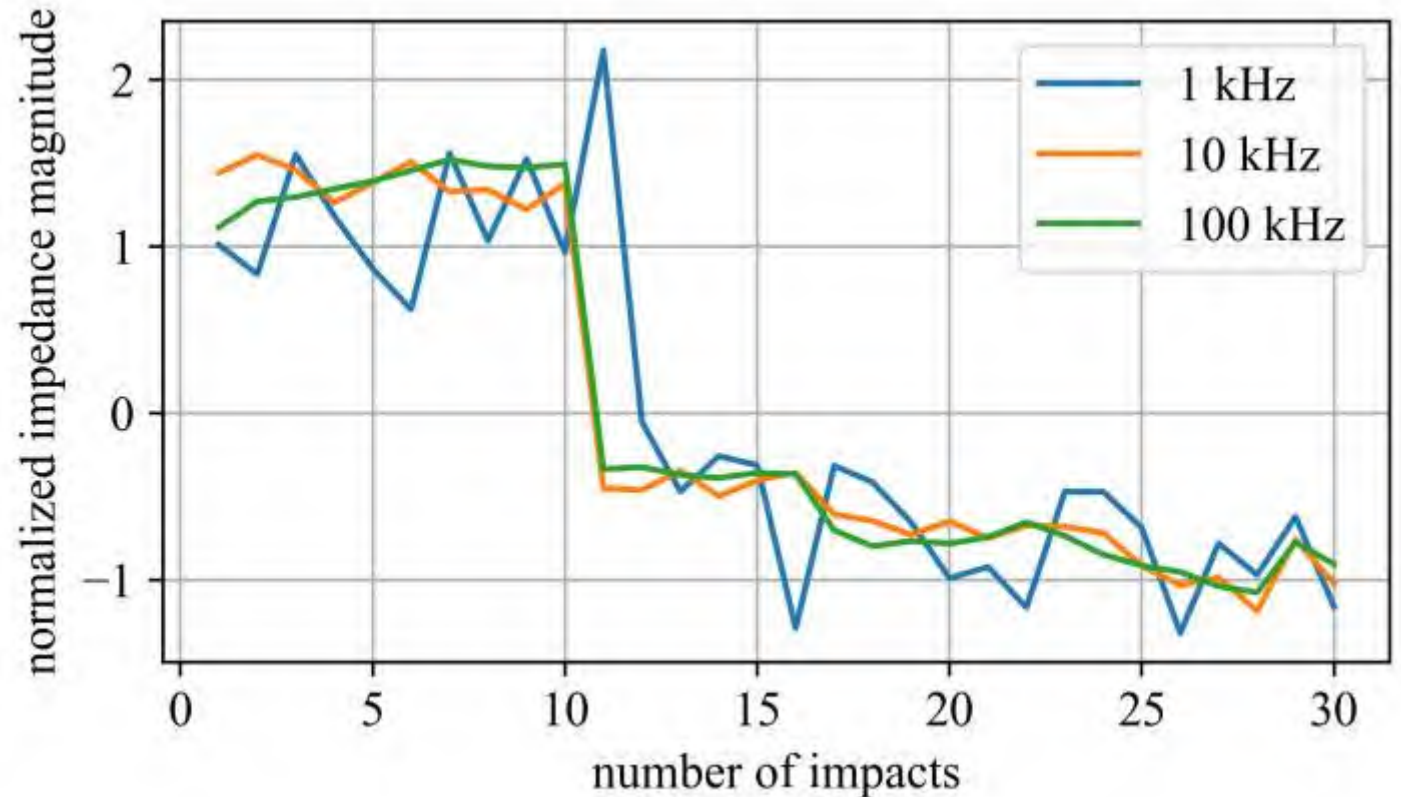
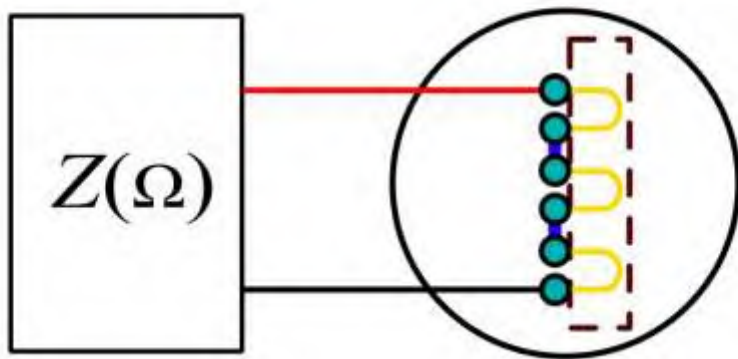
Experimental System used for Validation



Experimental System used for Validation



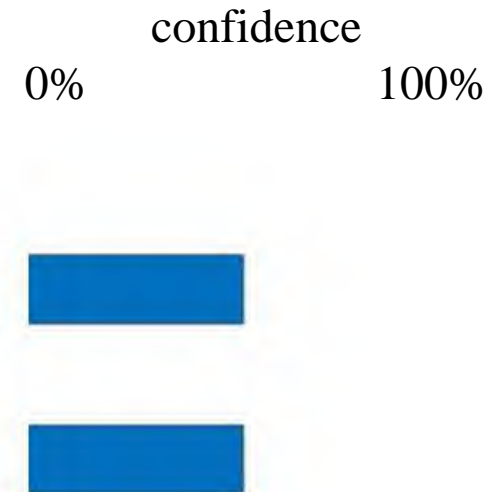
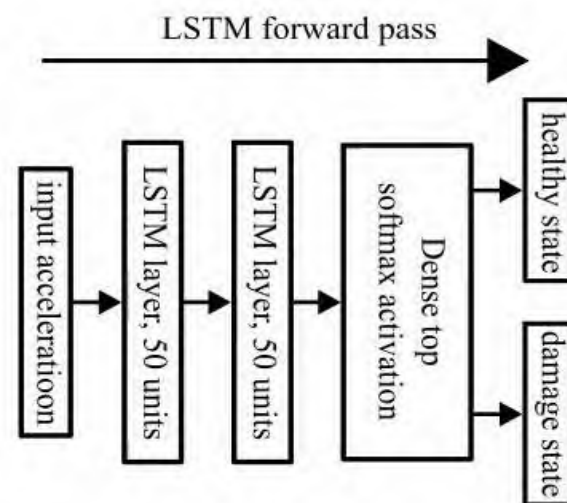
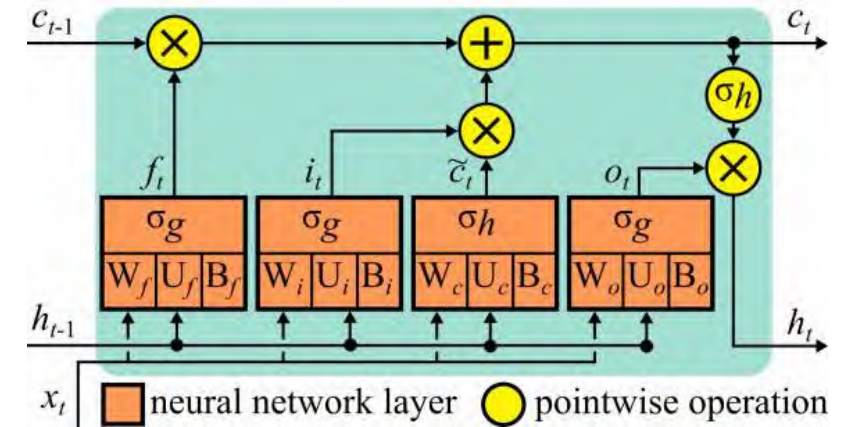
— PCB connection
— internal connections



LSTM-based Real-time State Estimation

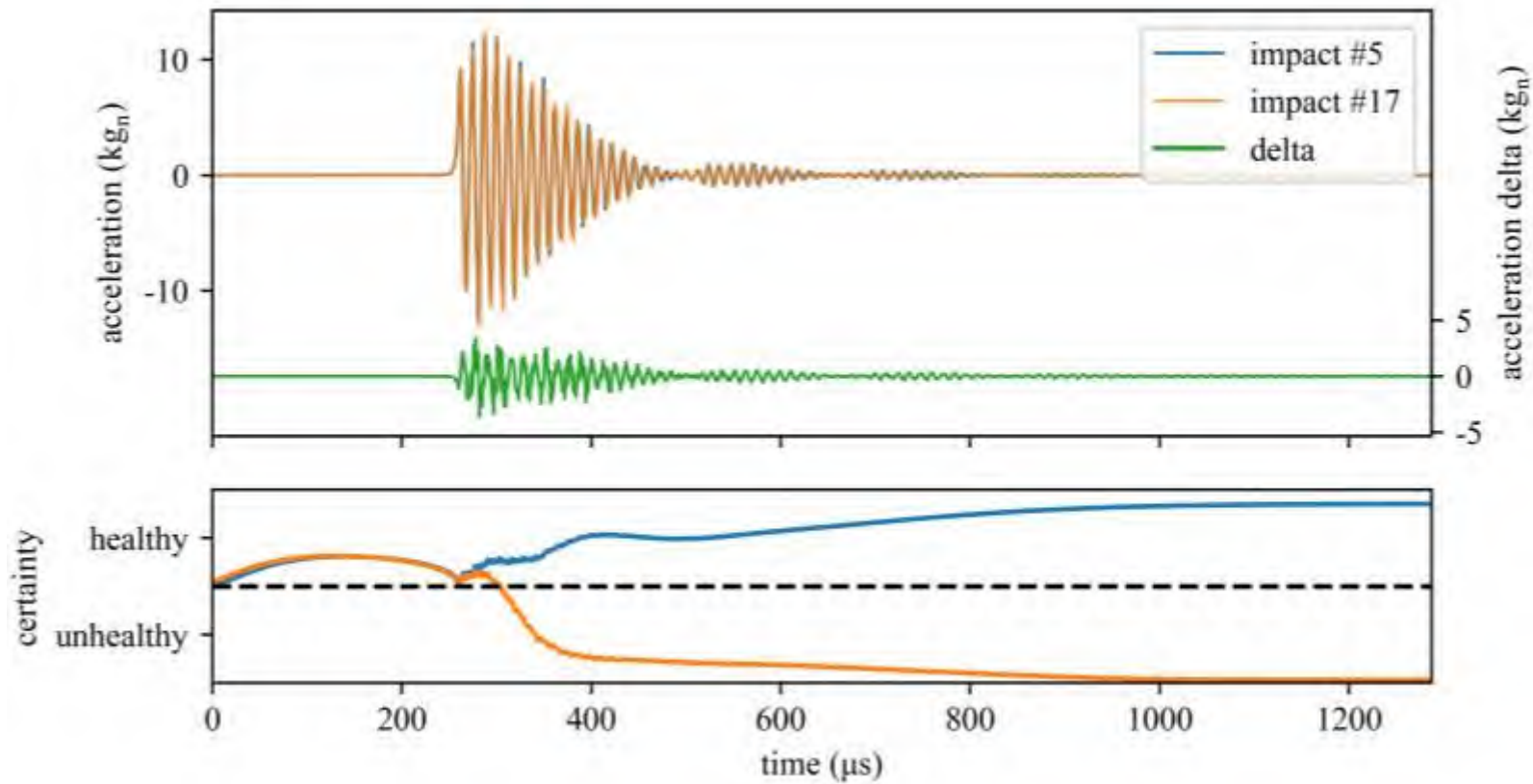
In this work:

- Long short-term memory (LSTM) models are used for real-time state estimation.
- Models are initially trained offline on pre-recorded data.
- LSTM architecture is (50, 50 units) with a dense layer at the output with SoftMax activation



Model Results

Prediction of survivability of PCB exposed to shock loads



actual health state	healthy	10	0
	unhealthy	0	10
		healthy	unhealthy
		predicted health state	

FPGA Implementation (Timing Consideration)

Data Driven Model Updating
(Theory and Proof of Concept)

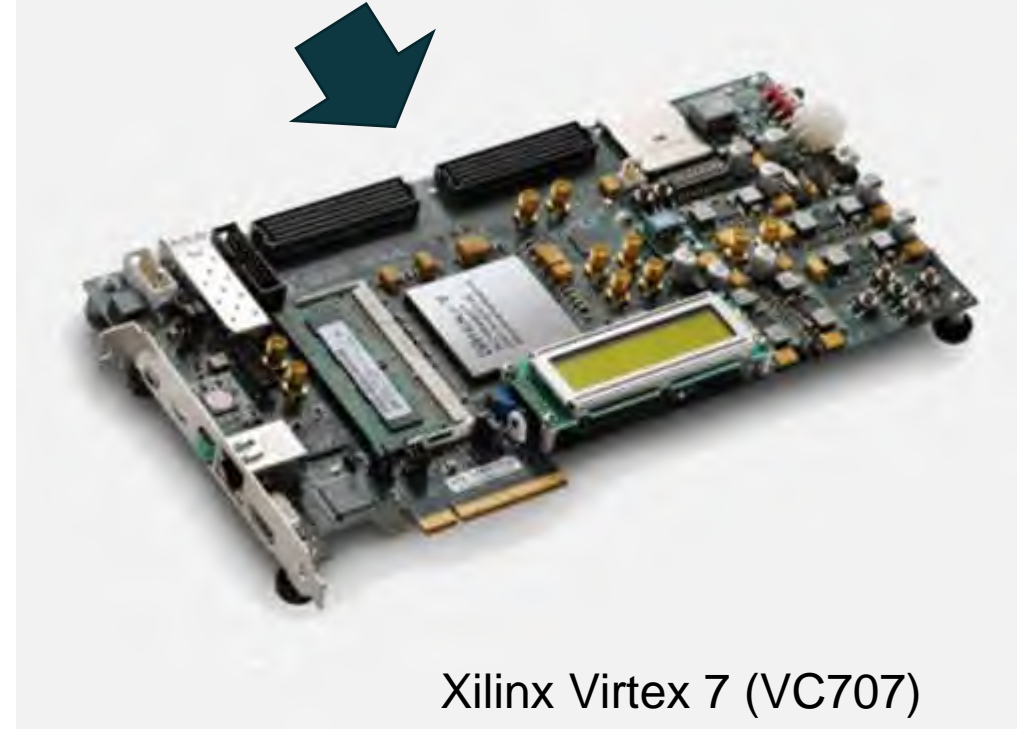
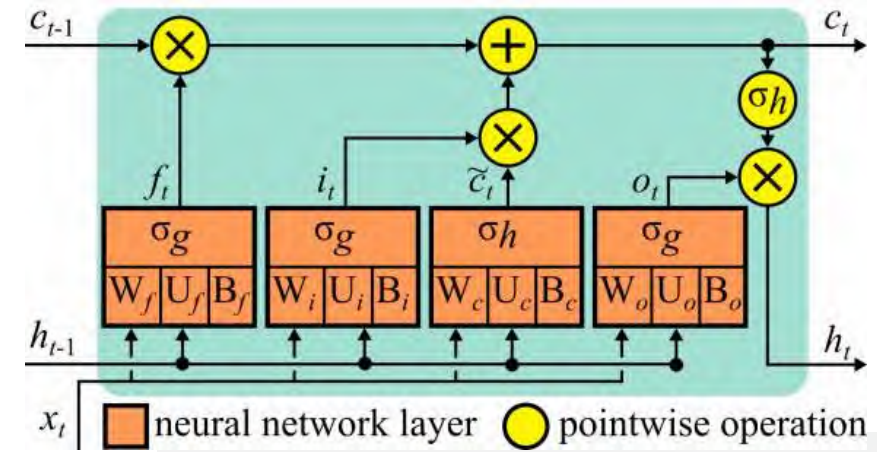
Electronic Components
Under Shock (Application)

FPGA Implementation
(Timing Consideration)

Model Deployment on

LSTM model deployed on a Xilinx Virtex 7 (VC707) FPGA:

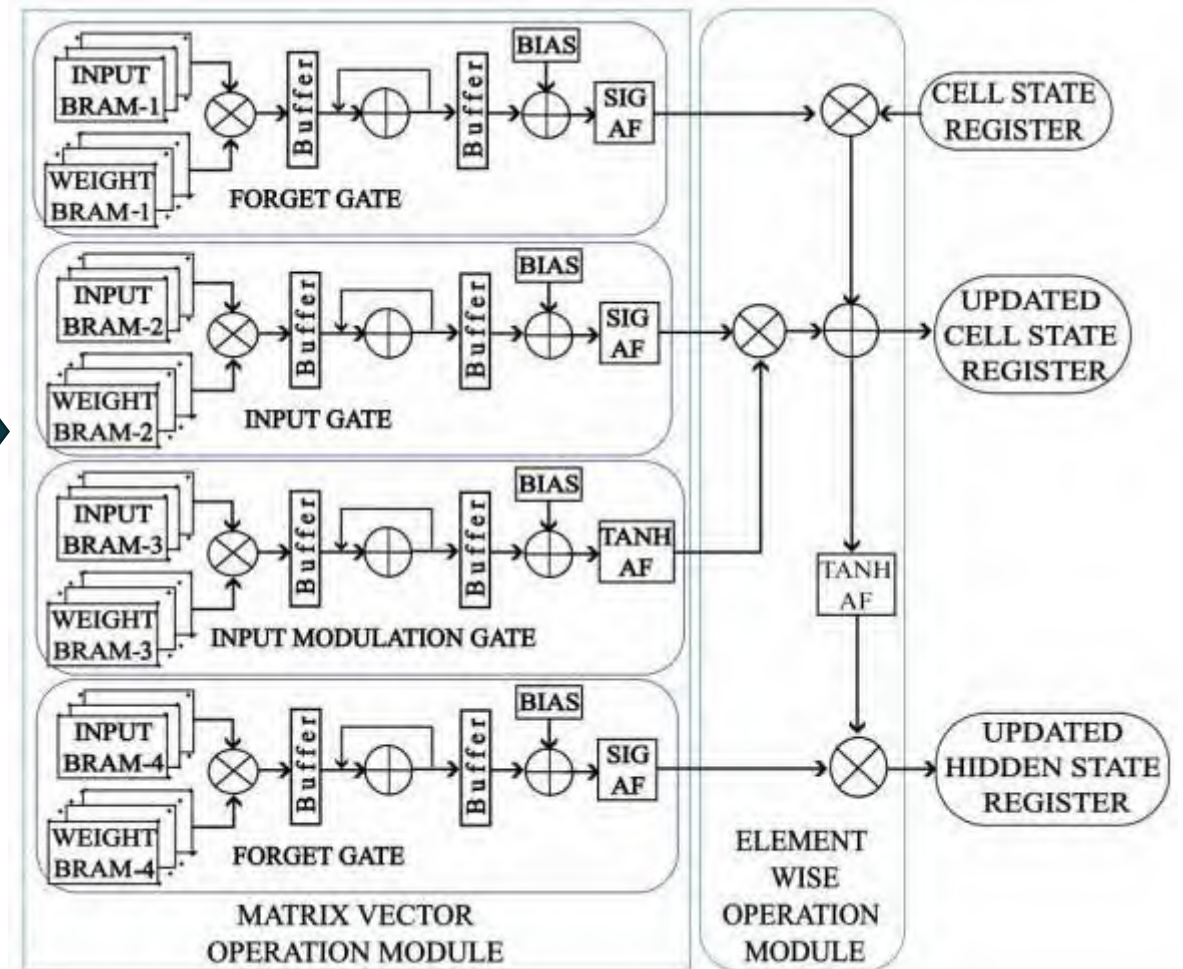
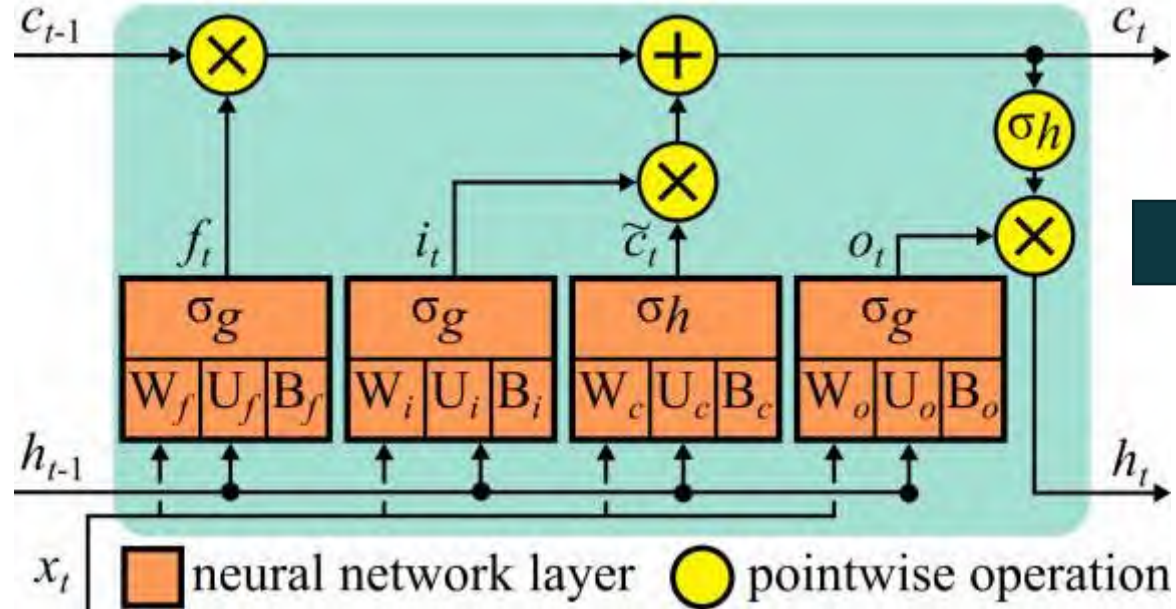
- Implemented in 8-bit, 16-bit and 32-bit fixed point .
- Developed an LSTM hardware accelerator using
- Design of an LSTM accelerator framework using high-level synthesis (HLS)
- Goal: to meets the real-time requirements set by high-rate applications.
- Findings: outermost loop pipelining generates a more efficient hardware design than outermost loop unrolling of the algorithm.



Xilinx Virtex 7 (VC707)

LSTM deployment on an FPGA

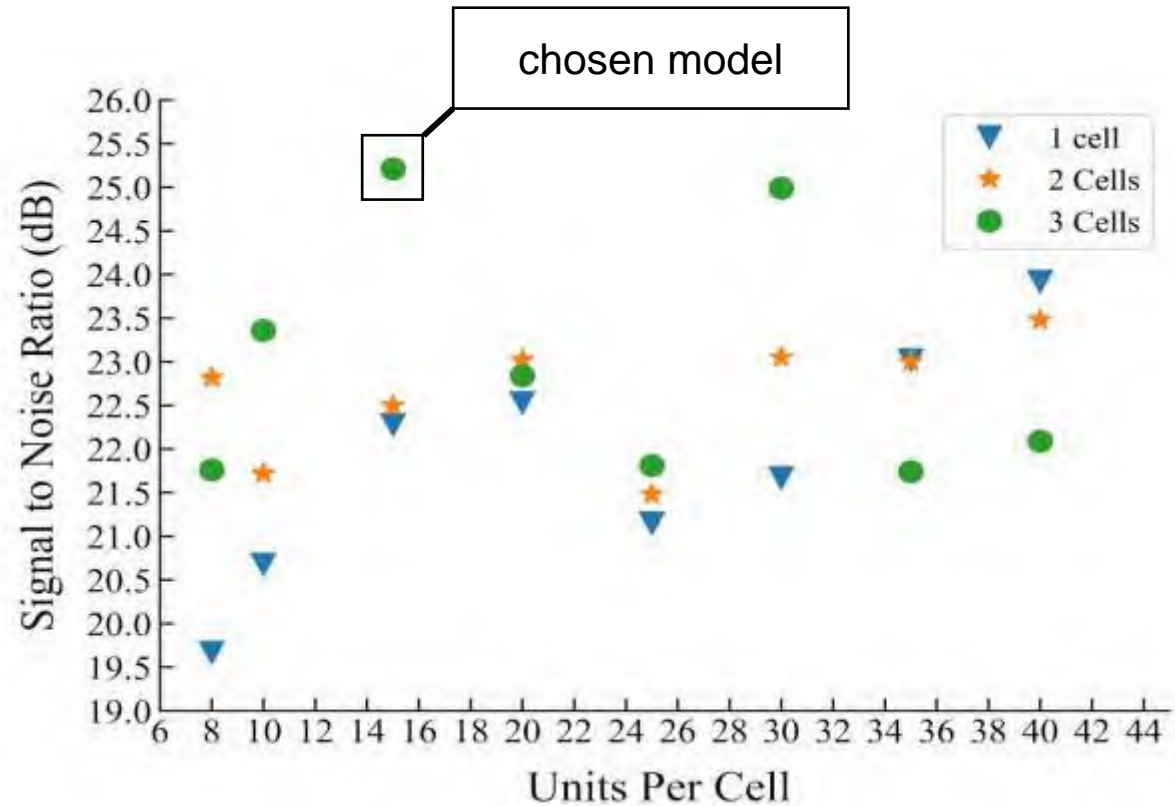
The developed hardware accelerator is split up into the LSTM's gates for deployment.



LSTM Operations for HDL Design

Model Selection

- A 3-layer configuration with 15 units per layer provided the highest signal to noise ratio (SNR).
- The model utilizes 16 input features derived from the uniformly sampled input signal at the preceding time step.
- The selected model generates an output state prediction every 500 μ s on Real-Time National Instruments testbed system.

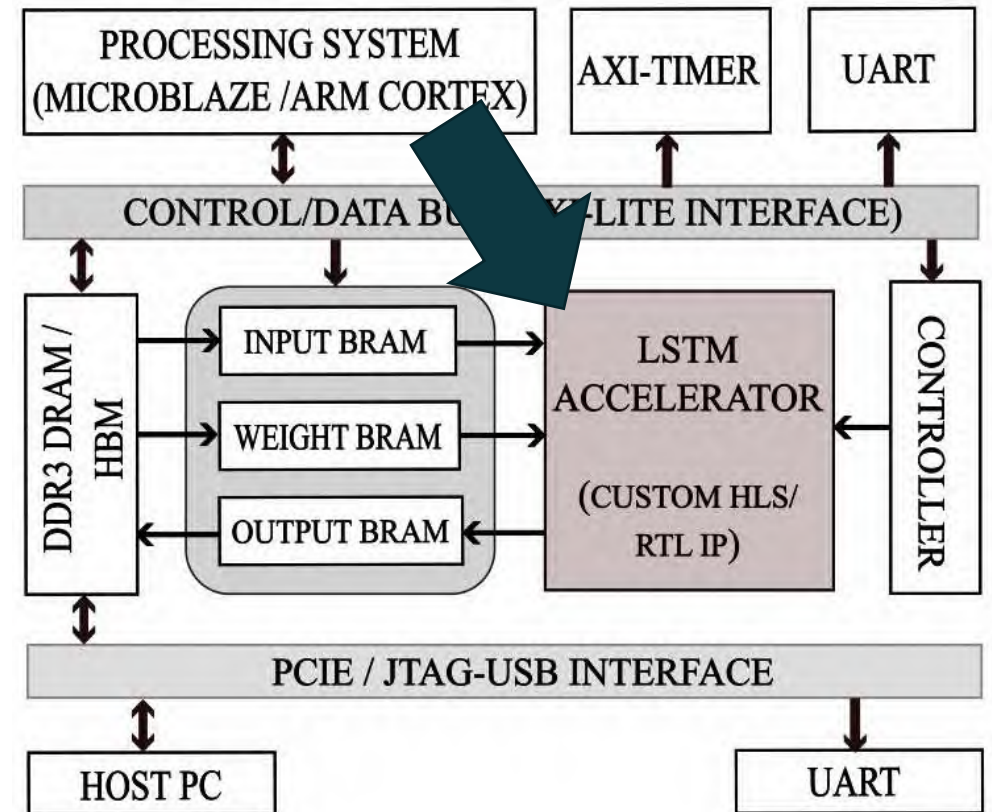


Custom LSTM Hardware Accelerator

Building a hardware accelerator that for deploying LSTMS with a focus on latency

HLS Implementation

- Designed with C++ in Vitis HLS.
- Two main units in the accelerator architecture:
 - Matrix-vector operations(MVO) unit
 - Element-wise operations(EVO) unit.
- Depending on the size of the LSTM network and the compiling capacity of the synthesis tool, arrays were partially or entirely partitioned to generate multiple BRAMs.
- Pipeline pragmas were used for the outer loops of the functions associate with LSTM gates. This unrolled the internal loops facilitating parallel multiplication.
- However, due to limited port of Block RAMs, full parallelization of operations was not achievable with HLS.

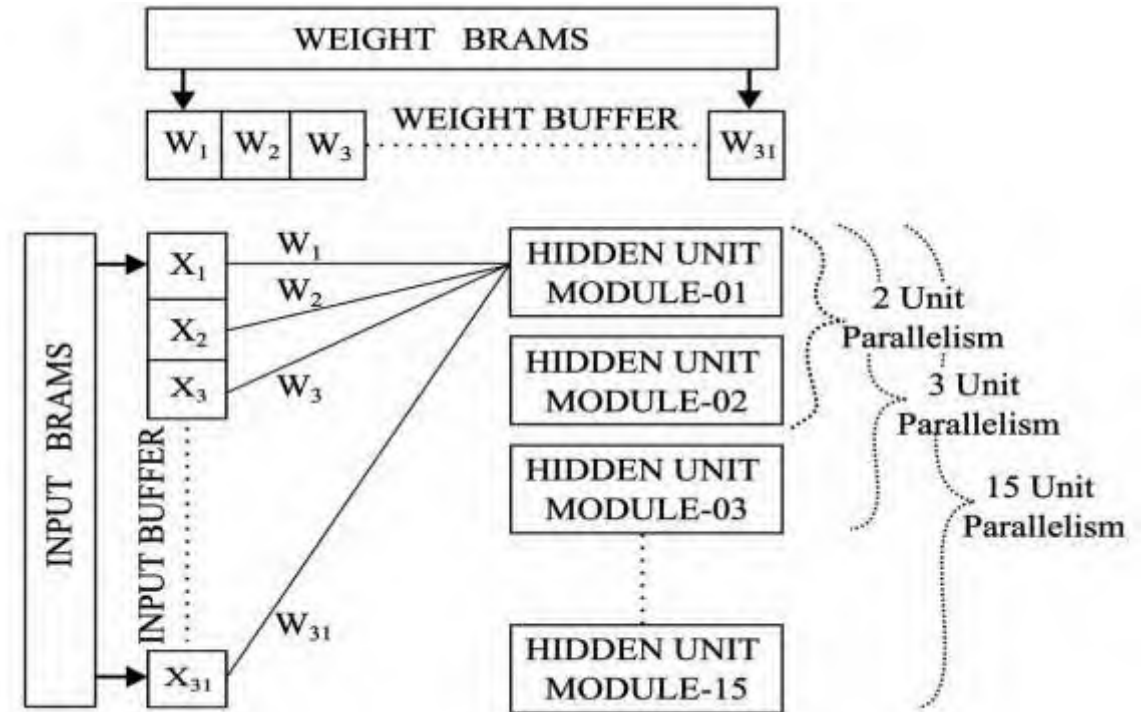


Custom LSTM Hardware Accelerator

Building a hardware accelerator that for deploying LSTMS with a focus on latency

HDL Implementation

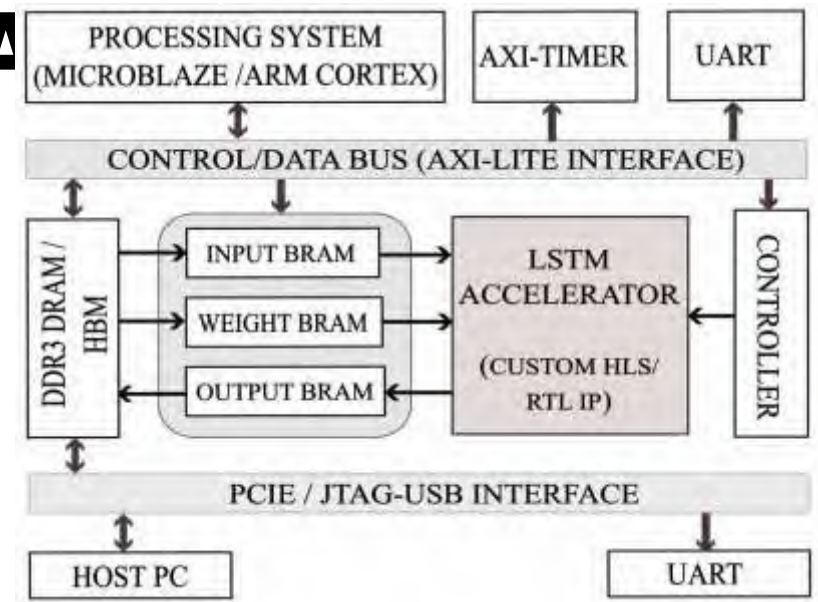
- Designed with Verilog to obtain more flexibility than HLS design.
- RTL module named 'hidden unit' does the matrix-vector operations. RTL design adds reconfigurable feature to this module which can be utilized to boost DSP consumption, thereby enhancing parallelism - a feat not attainable with HLS.
- Performance improved dramatically over HLS design, but the design becomes congested with the increase of DSP, limiting high-frequency operation.



Modules block diagram of the RTL design

Model Deployment on FPGA

- Model deployment on targeted datacenter platforms
 - ZCU104
 - VC707
 - U55C



ZCU104

(Zynq UltraScale+XCZU7EV-2FFVC1156 MPSoC)



VC707

(Virtex-7 XC7VX485TFFG1761-2)



U55C

(UltraScale+XCU55C-FSVH2892-2L-E)

Real-time LSTM Modeling Results (HLS)

Results for High-Level Synthesis (HLS) Design

Platform	Bit Precision	LUT	FF	BRAM 36k	DSP	Fmax (MHz)	Latency (μ S)	Throughput (GOPS)	GOPS/LUT	GOPS/DSP
Virtex 7	FP-32	70380 (23%)	86579 (14%)	41.5 (4%)	712 (25%)	210	8.75	1.28	18.19	1.80
	FP-16	30532 (10%)	36186 (6%)	22 (2%)	224 (8%)	213	7.4	1.51	49.46	6.74
	FP-8	26889 (9%)	20683 (3%)	0 (0%)	30 (1%)	235	6.36	1.76	65.45	58.67
ZCU104	FP-32	78850 (34%)	94936 (21%)	17.5 (16%)	712 (41%)	305	3.74	2.99	37.92	4.20
	FP-16	36458 (16%)	39326 (9%)	10 (3%)	224 (13%)	350	2.92	3.83	105.05	17.10
	FP-8	23575 (10%)	21590 (5%)	0 (0%)	15 (1%)	400	2.83	3.95	167.55	263.33
U55C	FP-32	64930 (5%)	80191 (3%)	29.5 (1%)	711 (8%)	362	6.86	1.63	25.10	2.29
	FP-16	25346 (2%)	31136 (1%)	16 (1%)	224 (2%)	375	4.72	2.36	93.42	10.57
	FP-8	23899 (2%)	17422 (1%)	0 (0%)	15 (0.2%)	380	4.65	2.4	100	160.00

Real-time LSTM Modeling Results (HDL)

Results for Hardware Design Language (HDL) Synthesis Design

Platform	Bit Precision	LUT (%)	FF (%)	BRAM 36k (%)	DSP (%)	Fmax (MHz)	Latency (μ S)	Throughput (GOPS)	GOPS/LUT	GOPS/DSP
Virtex 7	FP-32	17	16	1	43	150	11.48	0.97	19.34	0.81
	FP-16	22	23	5	41	166	3.71	3.01	45.19	2.64
	FP-8	13	12	5	35	200	3.10	3.61	95.06	3.64
ZCU104	FP-32	22	21	4	69	230	7.11	1.57	31.62	1.31
	FP-16	30	29	15	66	250	2.14	5.21	76.69	4.56
	FP-8	16	16	15	57	300	1.72	6.50	171.61	6.55
U55C	FP-32	4	4	1	13	250	6.826	1.64	6.83	1.37
	FP-16	5	5	2	13	256	2.492	4.48	2.49	3.92
	FP-8	3	3	2	11	300	2.108	5.30	2.11	5.34

Parallelism study

Effect of Parallelism on HDL Design

- LSTM hardware accelerator replacement created in both Hardware Description Language (HDL) and High Level Synthesis (HLS). HDL exposed more parallelism.
- Software baseline system developed on National Instruments testbed. State prediction output every 500 μ s.

Platform	Bit Precision	LUT (%)	DSP (%)	Highest Level of Parallelism	Fmax (MHz)	Latency (μ S)
Virtex 7	FP-32	28	69	4 Units	142	5.78
	FP-16	39	72	15 Units	166	2.06
U55C	FP-32	11	38	8 Units	150	2.38
	FP-16	9	22	15 Units	250	1.42

Takeaway

It is possible to use online data-driven models for micro-second tracking of structures during impact.



Acknowledgement

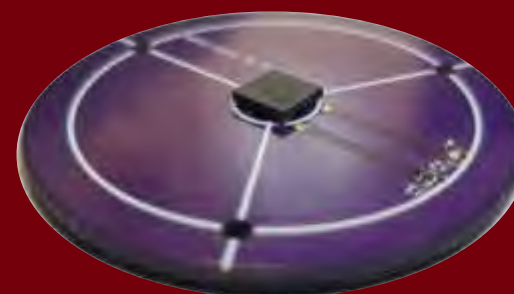
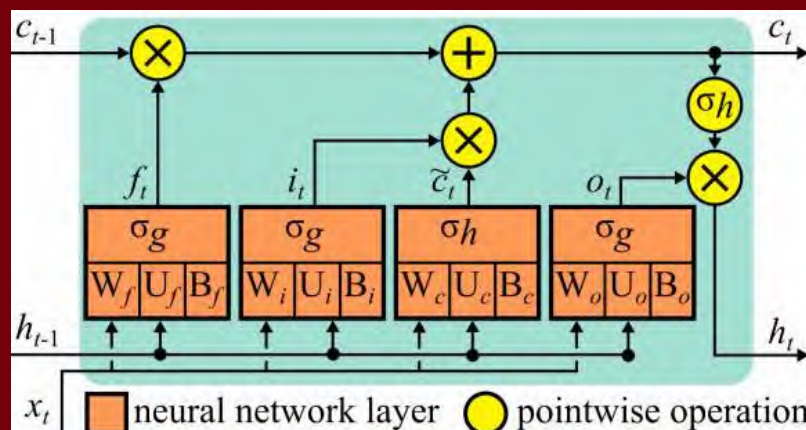


This material is based upon work supported by the Air Force Office of Scientific Research (AFOSR) through award no. FA9550-21-1-0083. This work is also partly supported by the National Science Foundation Grant numbers 1850012 and 1956071. The support of these agencies is gratefully acknowledged. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors, and they do not necessarily reflect the views of the National Science Foundation or the United States Air Force.

DISCUSSION



NEXT-STEPS



Contact Information: Austin Downey
Email: austindowney@sc.edu
Github: <https://github.com/austindowney>
Github-Lab: <https://github.com/Arts-laboratory/>

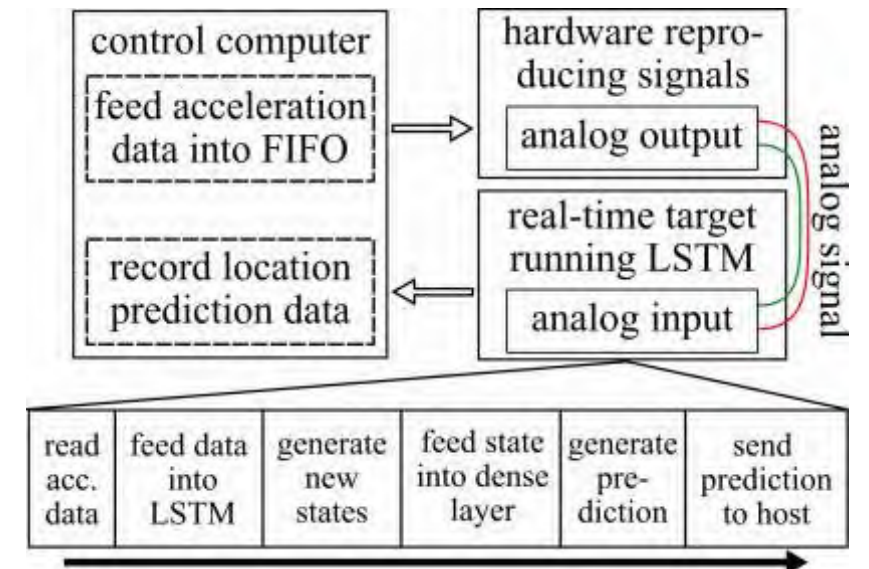
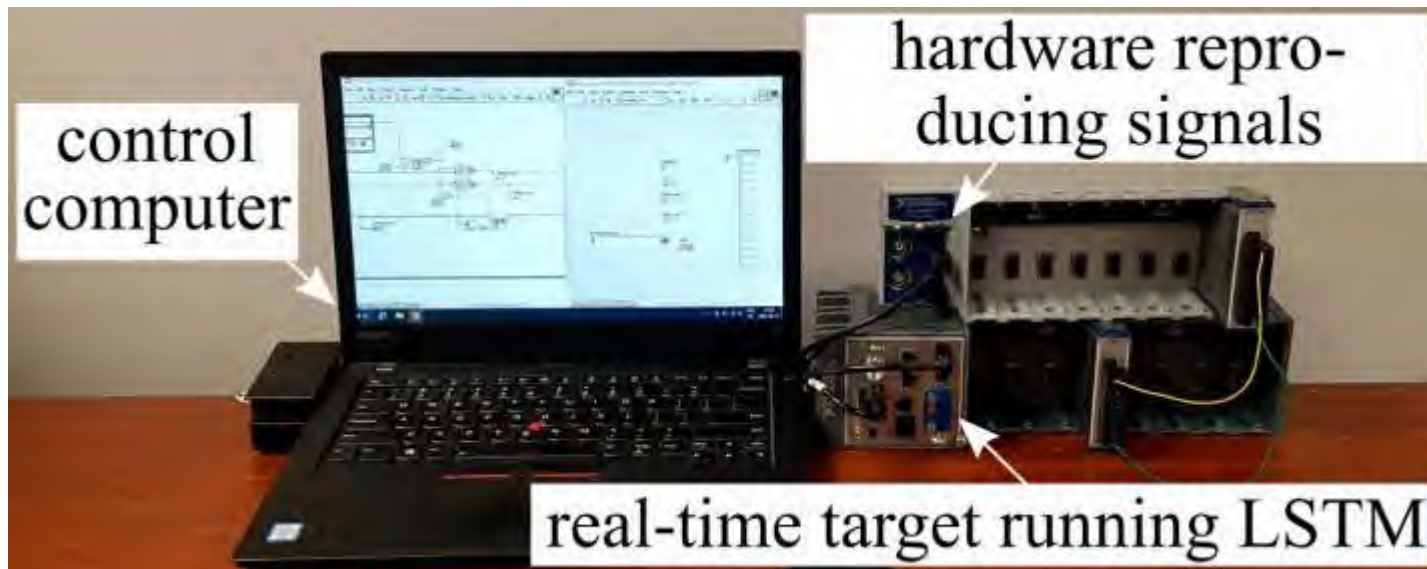


**Molinaroli College of
Engineering and Computing**
UNIVERSITY OF SOUTH CAROLINA

Backup Slides

Model Deployment on RTOS

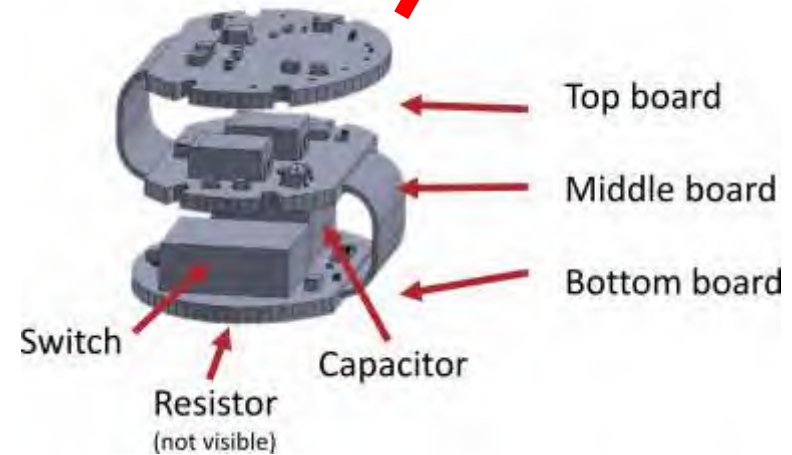
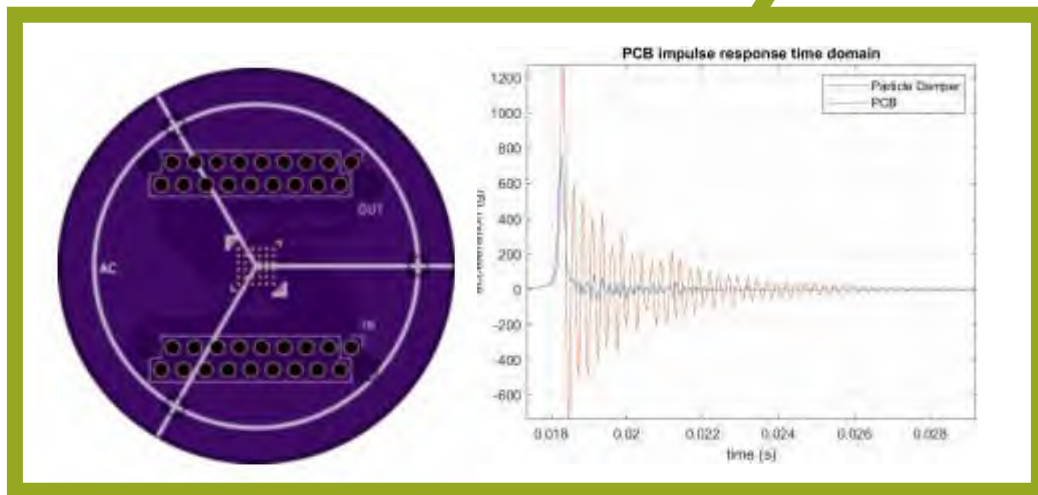
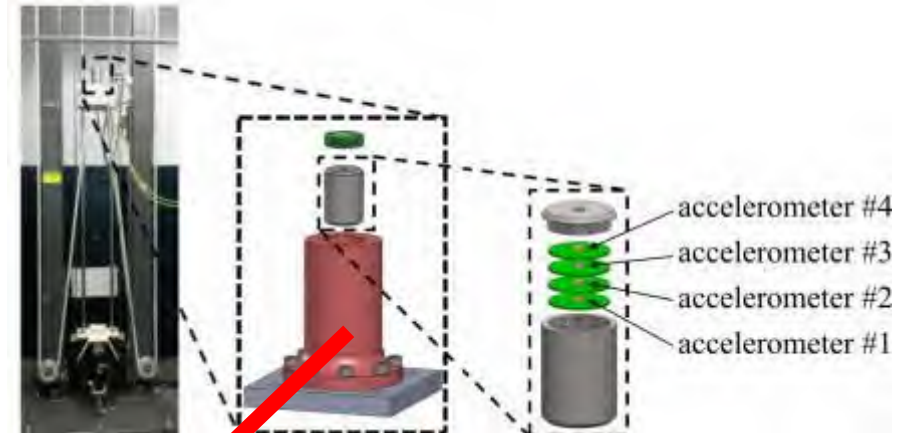
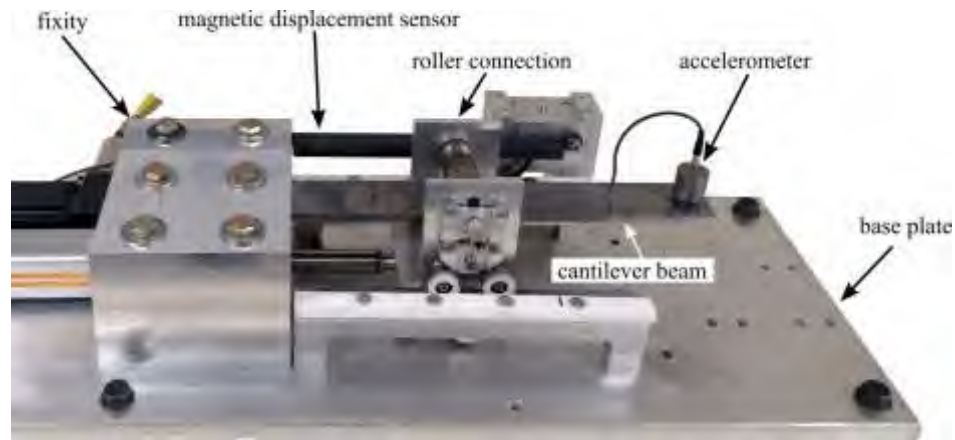
- A real-time system demonstrates an end-to-end prediction system.
- Signal reproduction is isolated from real-time system performing signal acquisition, state updating, and health estimation.
- Health estimates are communicated back to the host computer.



Experimental System used for Validation

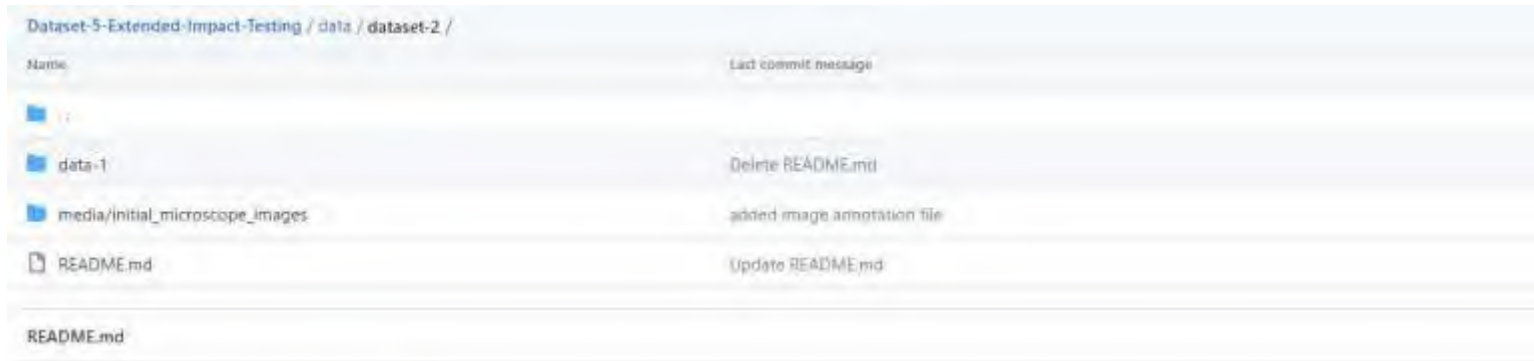
Datasets of Varying Complexity

Air Force Systems



Dataset Layout

<https://github.com/High-Rate-SHM-Working-Group/Dataset-5-Extended-Impact-Testing/tree/main/data/dataset-2>



The screenshot shows a GitHub repository directory for 'dataset-2'. It lists several folders and files with their respective commit messages:

Name	Last commit message
..	Last commit message
data-1	Delete README.md
media/initial_microscope_images	added image annotation file
README.md	Update README.md

Dataset 2

Dataset 2 consists of 32 tests performed May 5 2023. Tests were performed consecutively on the same PCB. Following each impact test, impedance was measured at five LCR excitation frequencies. The folder also contains a python file with a demonstration for extracting data from the .lvm files and plotting, and figures plotting the acceleration and measured impedance.

