Event Prediction in Processors using Deep Temporal Models

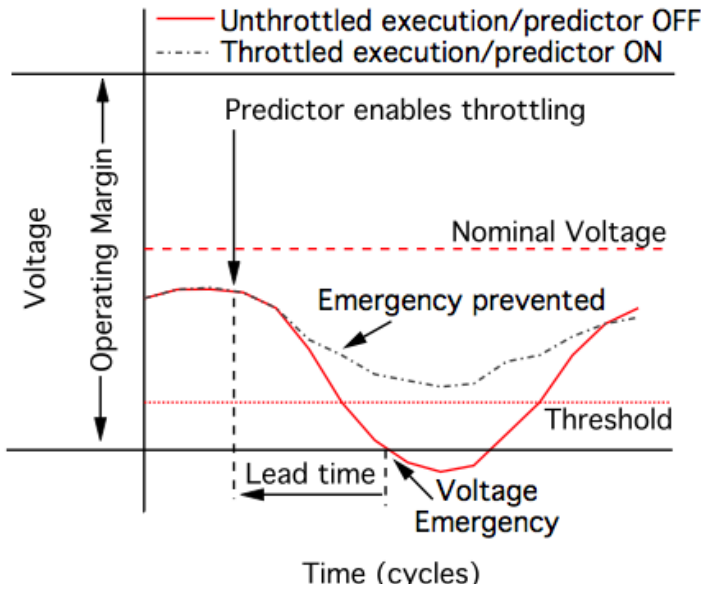
Tharindu Mathew, Aswin Raghavan, Sek Chai
SRI International

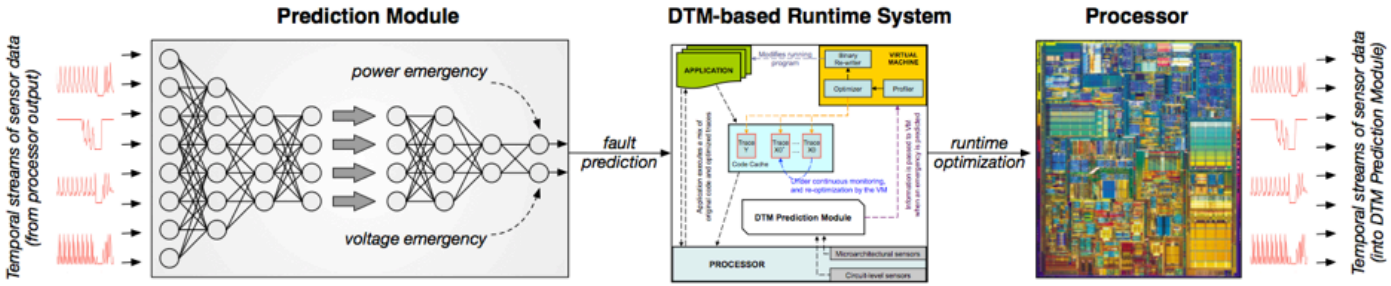
© 2018 SRI International. All Rights Reserved. Confidential

NSF Resilient Computing

The overall objective of our work is to lay the foundation for a proactive computing platform that manages its own fault resiliency and reliability.

The ultimate goal is to eliminate the penalties to system design and operation that arise in the use of incremental circuit techniques and micro-architectural changes that increase system complexity, affect power consumption and cost.

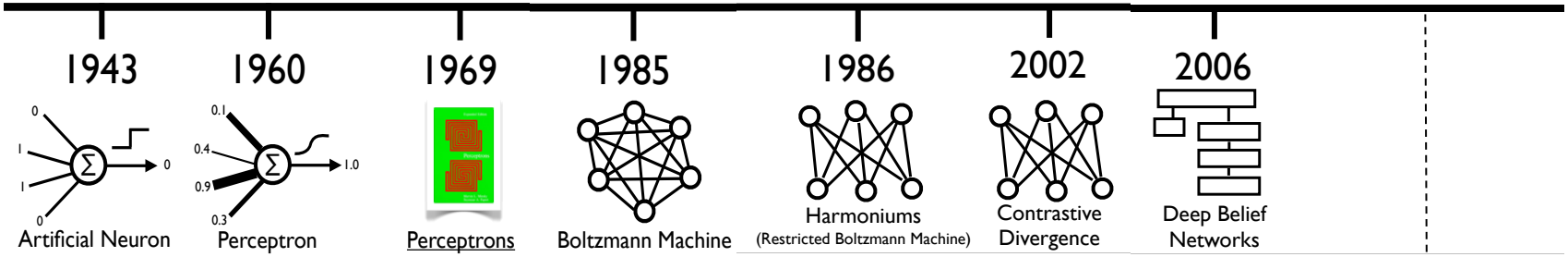




© 2018 SRI International. All Rights Reserved. Confidential

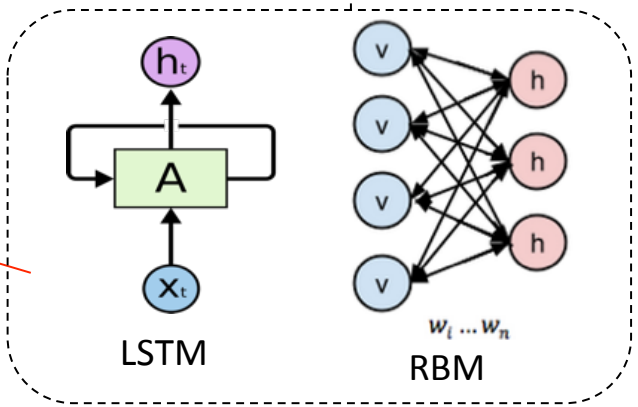
SRI International

Brief Taxonomy



Deep Temporal Models (DTM)

- Are hierarchical and multi-layered
- Address spatio-temporal data
- Typically, recurrent connections.
- Examples: LSTM, RBM



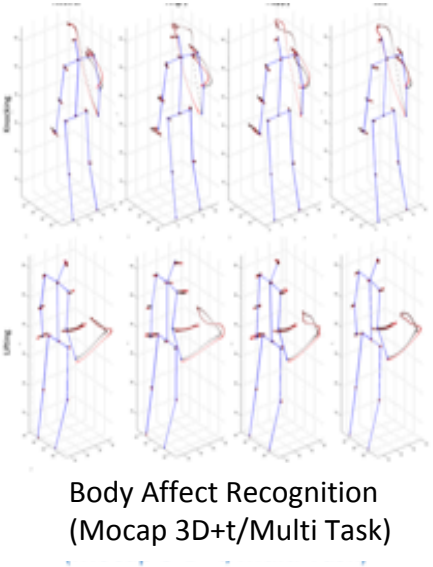
Deep Temporal Models

Timeline info loosely based on Yoshua Bengio's talk:
https://homes.cs.washington.edu/~rcg/talks/BengioDeepArchitectures_GensCommentary.pdf

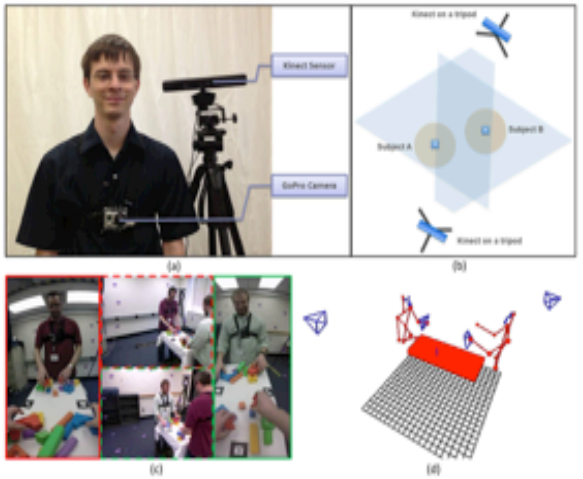
© 2018 SRI International. All Rights Reserved. Confidential

Applications using DTM (Benchmarks 1D – 4D dimensional data)

Facial Attributes Classification
(Images 2D/Multi Task)



Body Affect Recognition
(Mocap 3D+t/Multi Task)

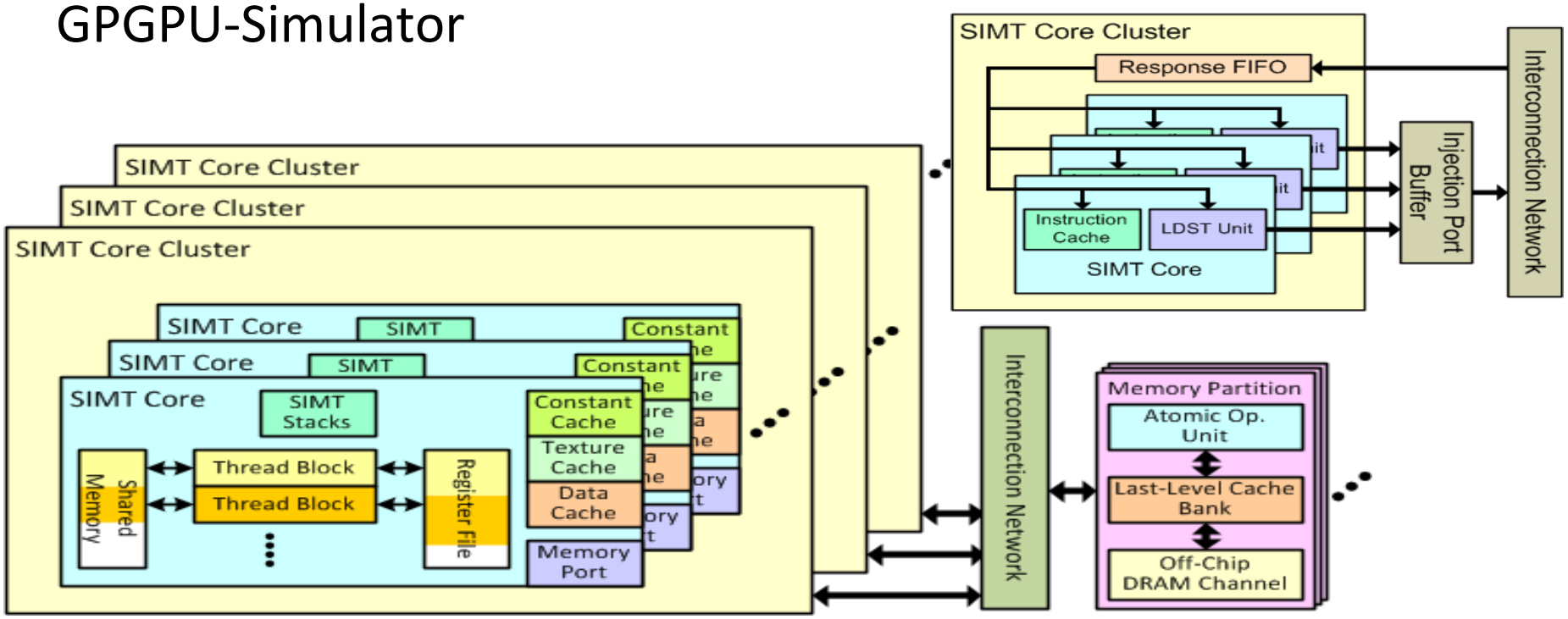


Tower Game for Engagement Classification
(Mocap, Audio, Depth, RGB: N-D+t/Multi Task)

Amer, M. R., Shields, T., Siddiquie, B., Tamrakar, A., Divakaran, A., & Chai, S. (2018). Deep multimodal fusion: A hybrid approach. International Journal of Computer Vision, 126(2-4), 440-456.

© 2018 SRI International. All Rights Reserved. Confidential

GPGPU-Simulator



http://www.gpgpu-sim.org/

- Kernels is a block of code such as a C function.
- Each core contains many thread blocks that execute SIMT.

© 2018 SRI International. All Rights Reserved. Confidential

GPGPU-Sim Kernels

	Application	Dwarf	Domain	Problem Sizes
1	**Back Propagation (BP)**	Unstructured Grid	Pattern Recognition	65536 input nodes
2	**Breadth First Search (BF)**	Graph Traversal	Graph Algorithms	1000000 nodes
3	**CFD Solver (C)**	Unstructured Grid	Fluid Dynamics	97k elements
4	**Heart Wall Tracking (HW)**	Structured Grid	Medical Imaging	609 x 590 pixels/frame
5	**HotSpot (HP)**	Structured Grid	Physics Simulation	500 x 500 data points
6	**Kmeans (K)**	Dense Linear Algebra	Data Mining	204800 data points, 34 features
7	**Leukocyte Tracking (LK)**	Structured Grid	Medical Imaging	219 x 640 pixels/frame
8	**LU Decomposition (LU)**	Dense Linear Algebra	Linear Algebra	256 x 256 data points
9	**MUMmer (M)**	Graph Traversal	Bioinformatics	50000 25-character queries
10	**Needleman-Wensch (N)**	Dynamic Programming	Bioinformatics	2048 x 2048 data points
11	**SRAD (SR)**	Structured Grid	Image Processing	512 x 512 data points
12	**Stream Cluster (SC)**	Dense Linear Algebra	Data Mining	65536 points, 256 dimensions

Twelve (12) programs of the Rodinia Benchmark is run through the GPGPU simulator simulating a NVidia GTX 480. Executed instructions across cores and data cache read miss (DC_RM) data are extracted from all of the programs.

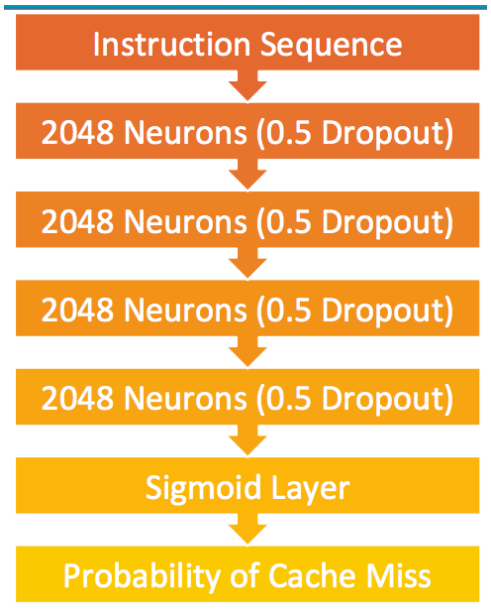
© 2018 SRI International. All Rights Reserved. Confidential

Time-series analysis using Deep Temporal Models

DTM is used as a sequence predictor.

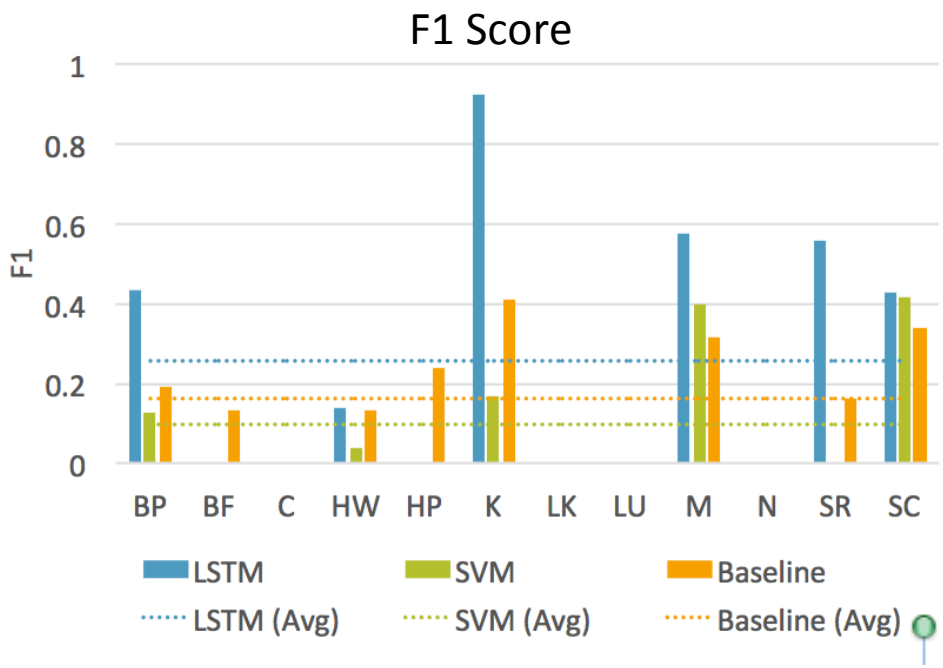
The GPU instruction sequences (e.g. add, mov, call) are *m* dimensional inputs with the occurrence of a cache miss as the binary output at each execution cycle. The deep temporal model is then trained to look at sequence of instructions, and predict the occurrence of a cache miss, *n* cycles ahead.

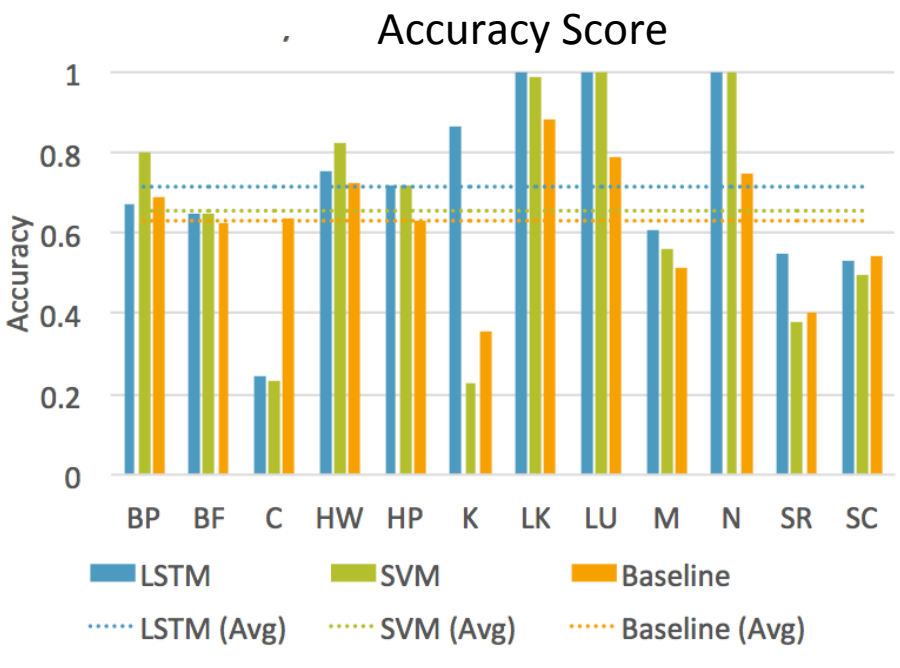
- For our Deep Temporal model, we use 2048-neuron LSTM cell in 4 successive layers.
- To improve regularization, a drop out ratio of 0.5 is applied after each layer.
- A sigmoid layer outputs the probability of a cache miss.
- A threshold of 0.5 is applied to create a binary output that represents the cache miss.



© 2018 SRI International. All Rights Reserved. Confidential

Predicting Cache Events

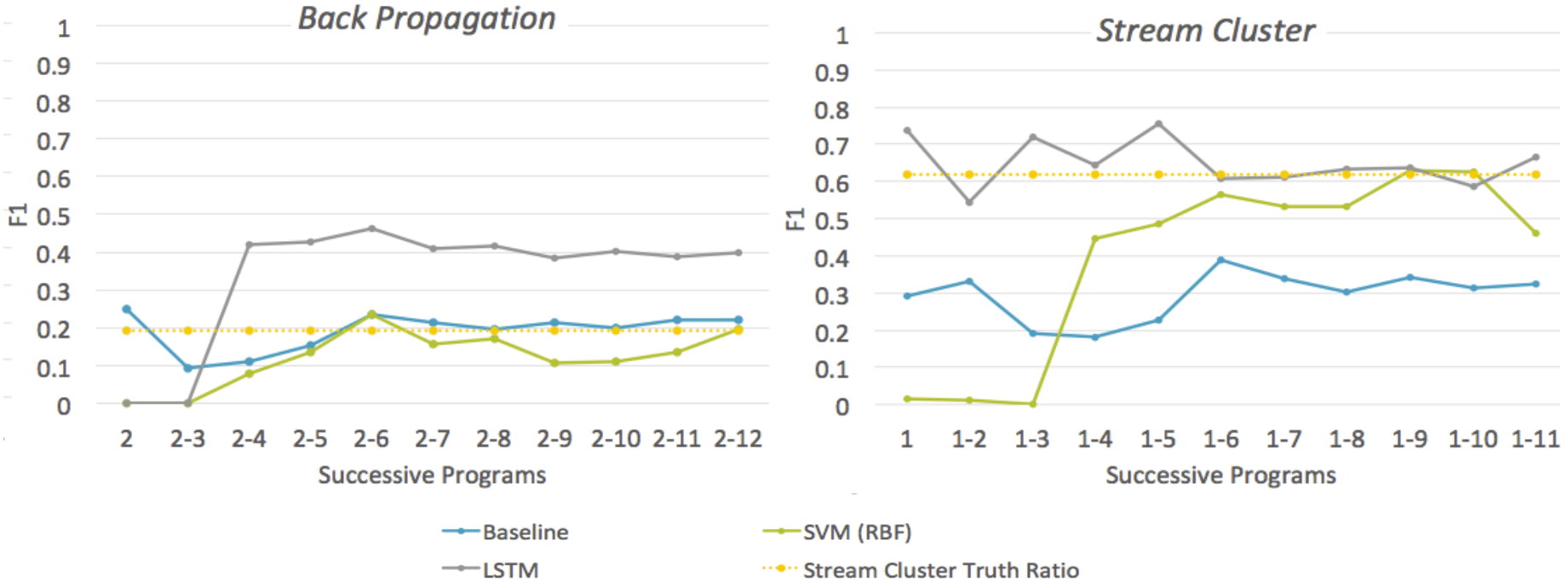




Five (5) randomly picked training programs are used as training programs to predict misses of an unseen test program

© 2018 SRI International. All Rights Reserved. Confidential

Successive cross-dataset prediction



Multiple program instructions are stacked successively for training. e.g. 1-6: instructions from BP+BF+C+HW+HP+K are stacked together

© 2018 SRI International. All Rights Reserved. Confidential