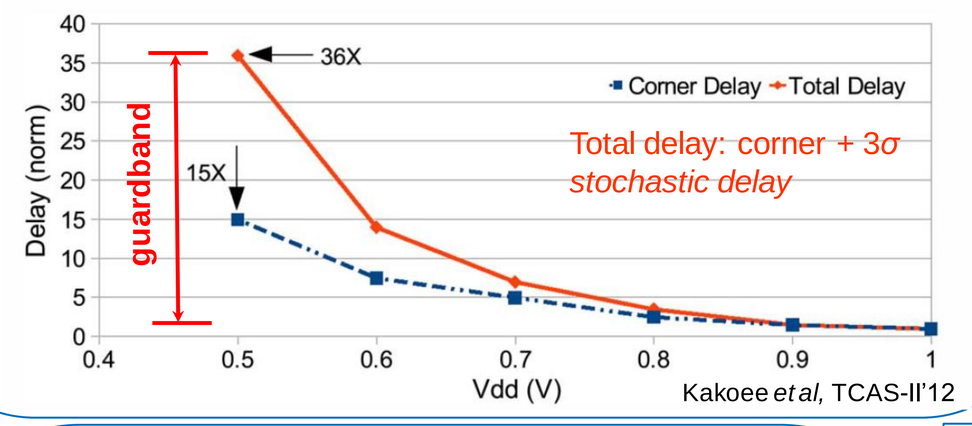
Energy-Efficient GPGPUs via Collaborative Compilation and Memristive Memory-Based Computing

Abbas Rahimi[†], Amirali Ghofrani[‡], Miguel A. Lastras-Montano[‡], Kwang-Ting Cheng[‡], Luca Benini[*] and Rajesh K. Gupta[†]

[†]UCSD, [‡]UCSB, [*]UNIBO, [*]ETHZ

Energy-Efficient GPGPUs

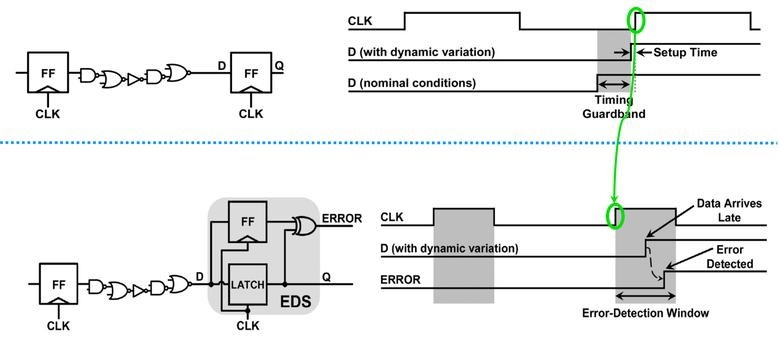
- NT and VOS achieve energy efficiency but
- Performance loss: SIMD ☺
- Susceptibility to variations: guardband ☹



Kakoee *et al*, TCAS-II'12

Variability is about Cost and Scale

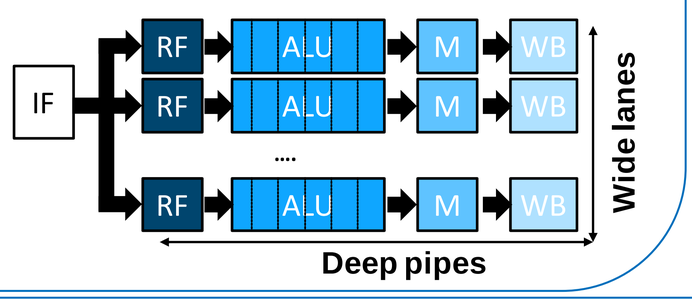
Eliminating guardband ☺



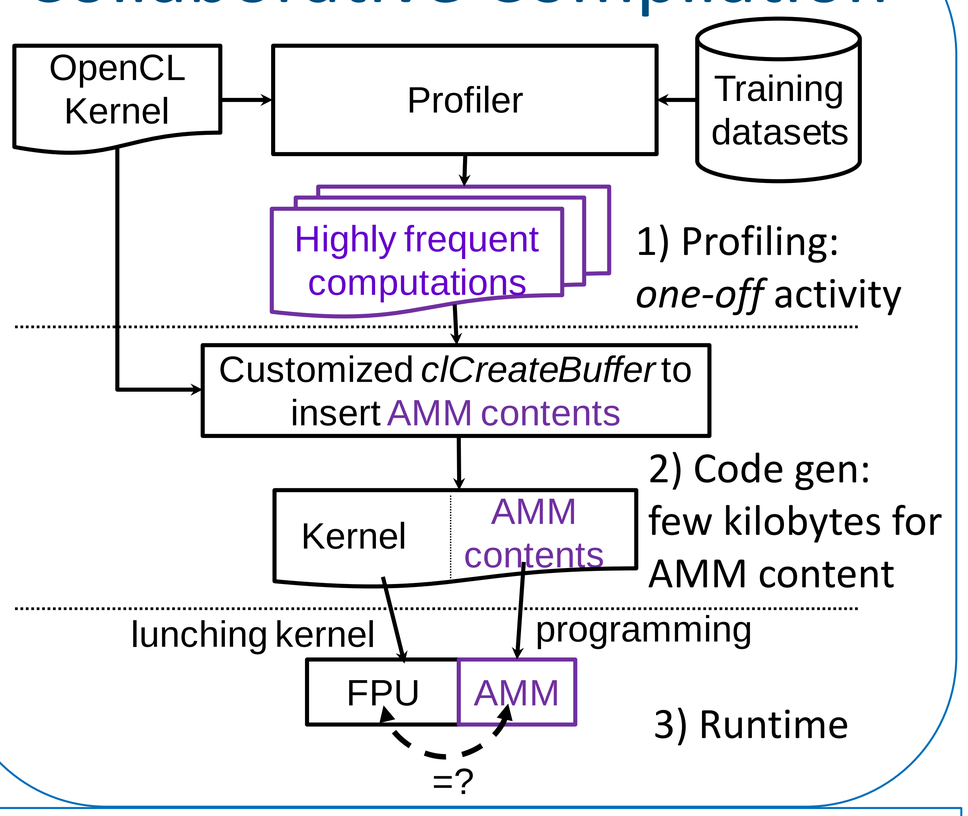
Timing error 😐

Costly error recovery for SIMD ☹

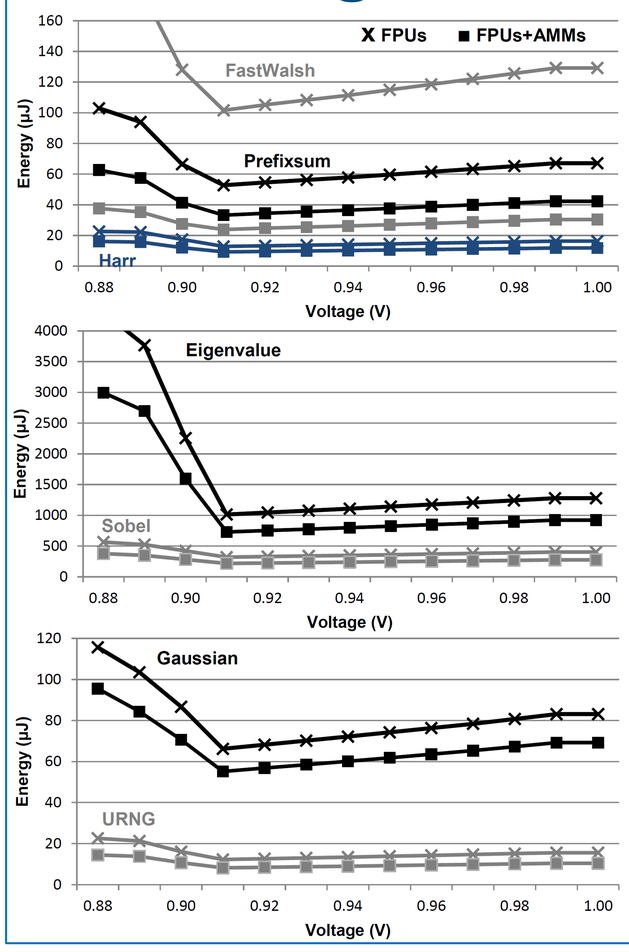
*** Recovery cycles increases linearly with pipeline length**
*** Error rate multiplies with wider pipeline width**
*** Quadratically expensive error correction**



Collaborative Compilation



Voltage Overscaling



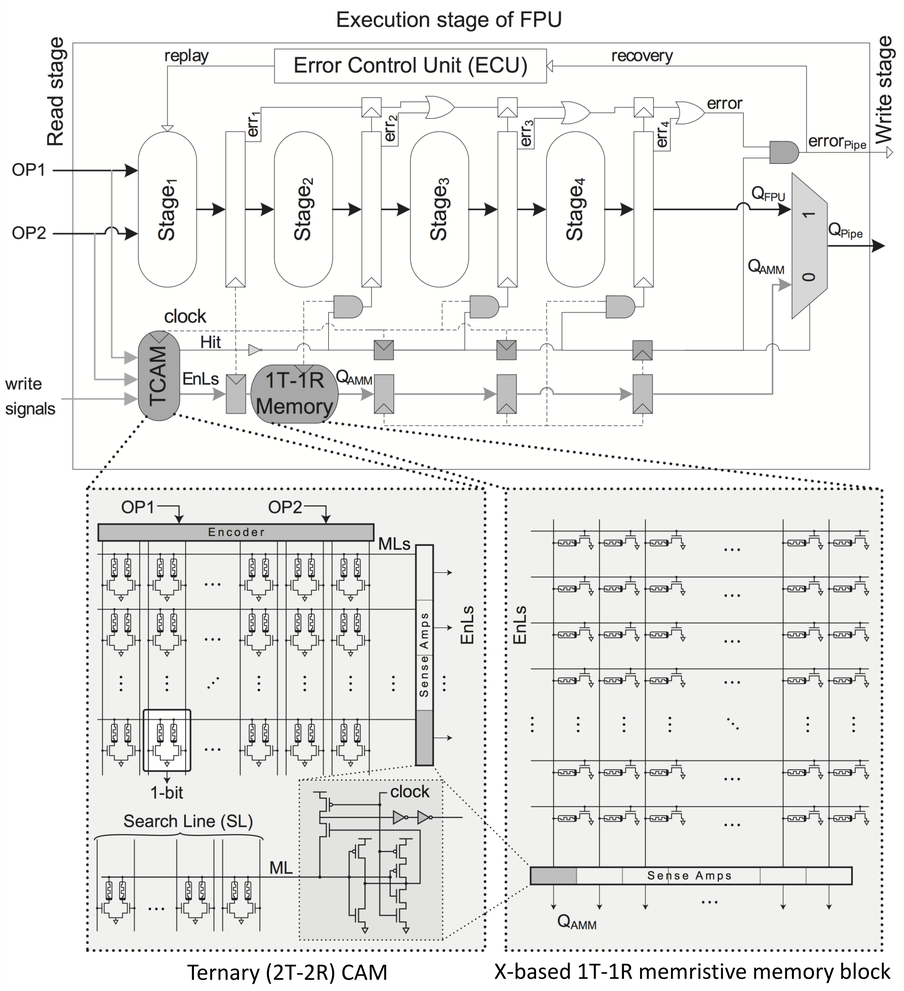
*** FPUs in the range of 1.0V to 0.88V, while AMMs at 1.0V**

*** At 1.0V, without timing error, up to 76% energy saving (for FastWalsh) and 36% average energy saving**

*** In the range of 0.92V-0.88V, the kernels face 10%-38% timing error (in the baseline FPUs) which is further reduced to 3%-24% in the proposed architecture (FPUs+AMMs).**

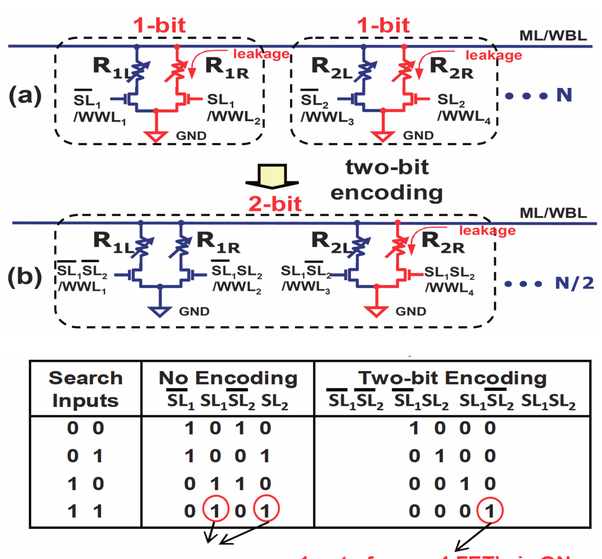
*** At 0.88V, an average energy saving of 39%.**

Associative Memristive-Based Computing



Ternary (2T-2R) CAM — X-based 1T-1R memristive memory block

*** Associative memristive memory (AMM):**
*** Software programmable**
*** Mimics partial functionality of FPU**
*** Reliable operation: 15% positive slack at (SS/0.81V/125°C) 45nm**



Search Inputs	No Encoding $\overline{SL_1}$ SL_1 $\overline{SL_2}$ SL_2	Two-bit Encoding $\overline{SL_1}\overline{SL_2}$ $\overline{SL_1}SL_2$ $SL_1\overline{SL_2}$ SL_1SL_2
0 0	1 0 1 0	1 0 0 0
0 1	1 0 0 1	0 1 0 0
1 0	0 1 1 0	0 0 1 0
1 1	0 1 0 1	0 0 0 1

2 out of every 4 FET's are ON: more leakage paths

1 out of every 4 FET's is ON: reduced leakage paths

*** Two-bit encoding improves sense margins**
*** To track PVT variations, a self referenced timing signal is generated by a reference subarray in the same 2T-2R configuration**

Overall AMM Hit Rates

*** For Sobel:**
*** All AMM modules display a hit rate of greater than 25% with a tiny TCAM of 32 lines.**
*** The AMM modules for MUL exhibit a hit rate of up to 49%**

*** For Haar:**
*** 39%-41% hit rate on MUL.**

*** Profiling: 20 random images from dataset$_1$**
*** The best hit rate: 84% (dataset$_2$) for Sobel**

*** Average hit rate:**
*** 38% (dataset$_1$) and 36% (dataset$_2$) for URNG**
*** 32% (dataset$_1$) and 24% (dataset$_2$) for Gaussian**
*** 34% (dataset$_1$) and 34% (dataset$_2$) for Sobel**

NSF Expedition in Computing, Variability-Aware Software for Efficient Computing with Nanoscale Devices http://variability.org

UC San Diego · UCLA · Michigan · Stanford University · UC Irvine · University of Illinois at Urbana-Champaign