



28 January 2008

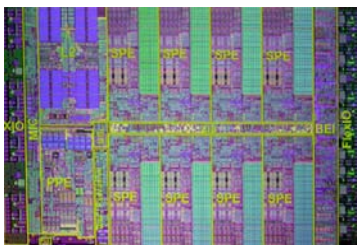
Architecture Optimization for Multidimensional Signal Processing

Prof. Dejan Markovic
dejan@ee.ucla.edu

Parallel Data Processing

- ◆ **Power limited technology scaling**
 - Increased impact of process variations
 - More leakage power, multiple threshold devices
- ◆ **Single dimensional → Multidimensional data**

Multi-core Processors



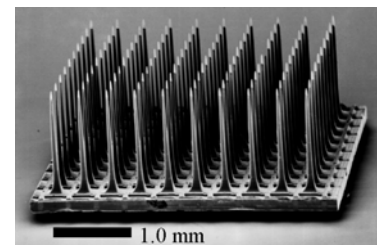
IBM / Sony / Toshiba

MIMO Communications



Belkin

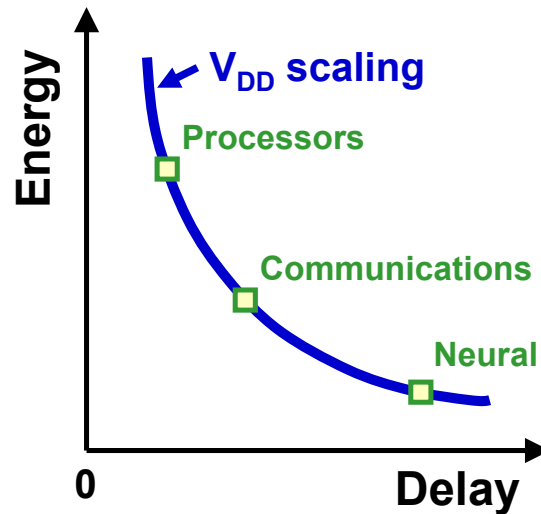
Neuroscience



www.sci.utah.edu

Different Energy-Delay Requirements

- ◆ **Processors**
 - Maximize performance
 - Highest V_{DD} required
- ◆ **Communications**
 - Minimize energy & area
 - Typically, sensitivity ~ 1
- ◆ **Neuroscience**
 - Power density: 0.8mWmm^2
 - Aggressive V_{DD} scaling



Same principle, different optimization goals

3

Circuit Optimization Framework

- ◆ **Sensitivity based optimization**
 - Balance sensitivity to all variables
 - Variables: gate size, V_{DD} , V_{TH}

minimize $Energy(V_{dd}, V_{th}, W)$
 subject to $Delay(V_{dd}, V_{th}, W) \leq D_{con}$

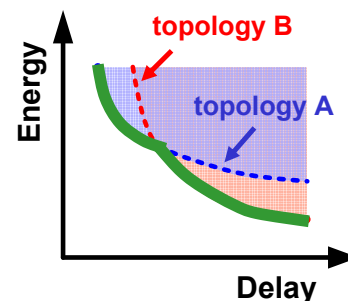
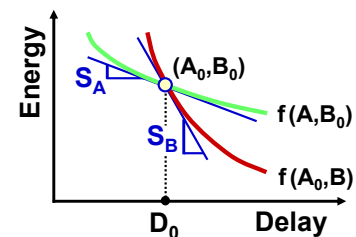
Constraints

$$V_{dd}^{min} < V_{dd} < V_{dd}^{max}$$

$$V_{th}^{min} < V_{th} < V_{th}^{max}$$

$$W^{min} < W$$

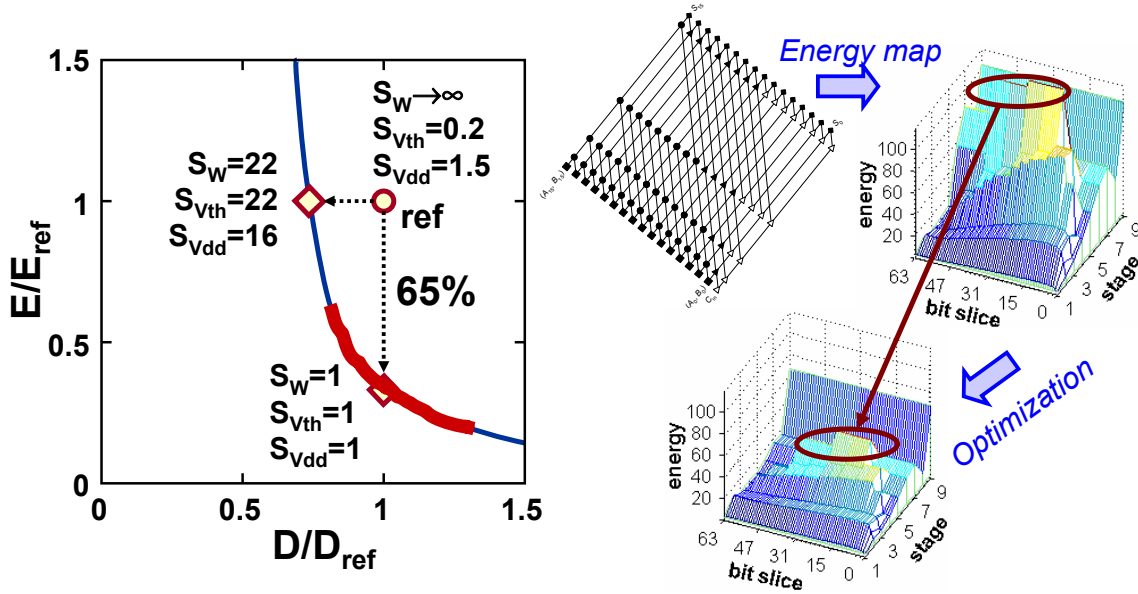
- ◆ **Reference design**
 - D_{min} sizing @ V_{dd}^{max} , V_{th}^{ref}



Goal: find optimal E-D tradeoff for a datapath

4

Circuit-Level Results: Tree Adder



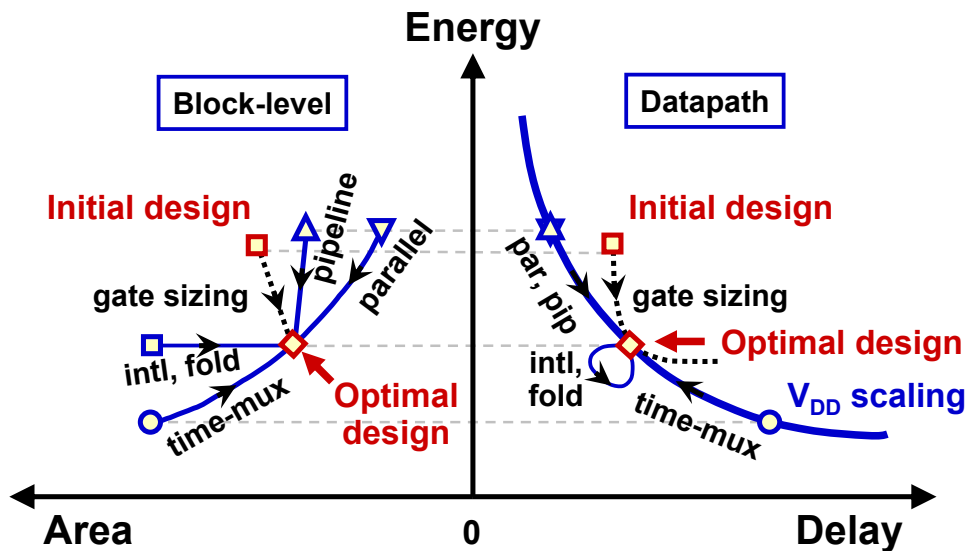
[D. Markovic et al, "Methods for True Energy-Performance Optimization" JSSC, Aug'04]

E-D space is the key for architecture optimization

5

Methodology for Architecture Selection

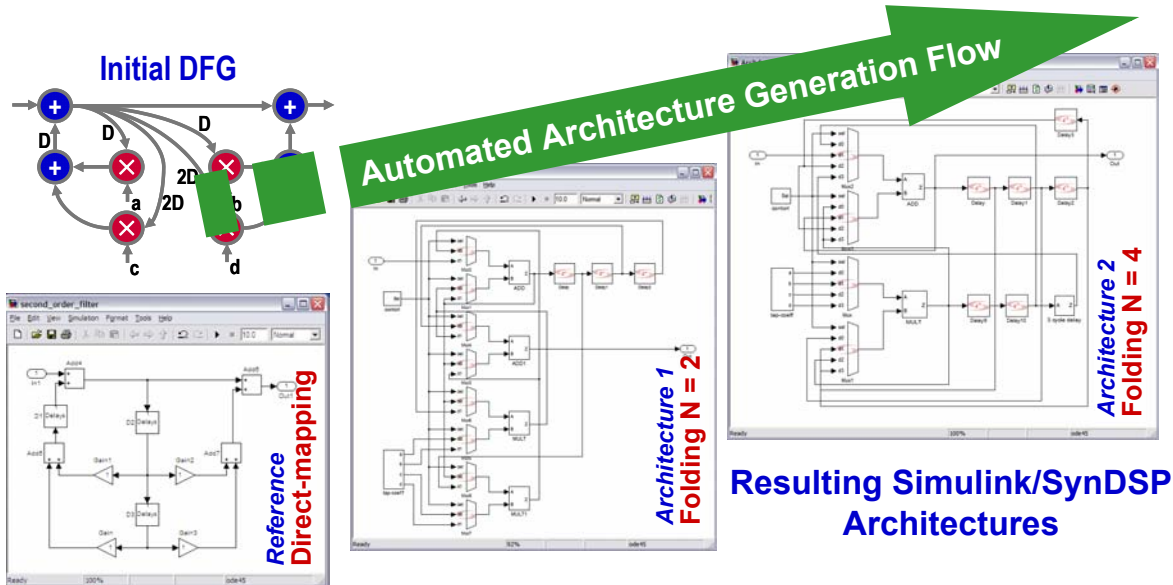
- ◆ **Energy-Area-Delay space for architecture comparison**
 - Time-mux, parallelism, pipelining, retiming, V_{DD} scaling, sizing



6

From Simulink to Optimized Hardware

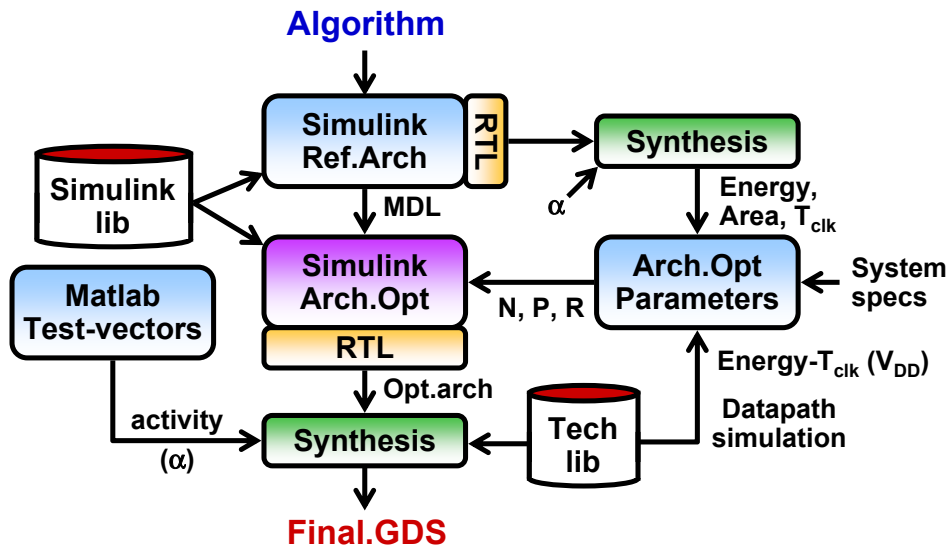
Direct mapped DFG → Scheduler → Architecture Solutions → Hardware
 (Simulink) (C++ / MOSEK) (Simulink/SynDSP) (FPGA/ASIC)



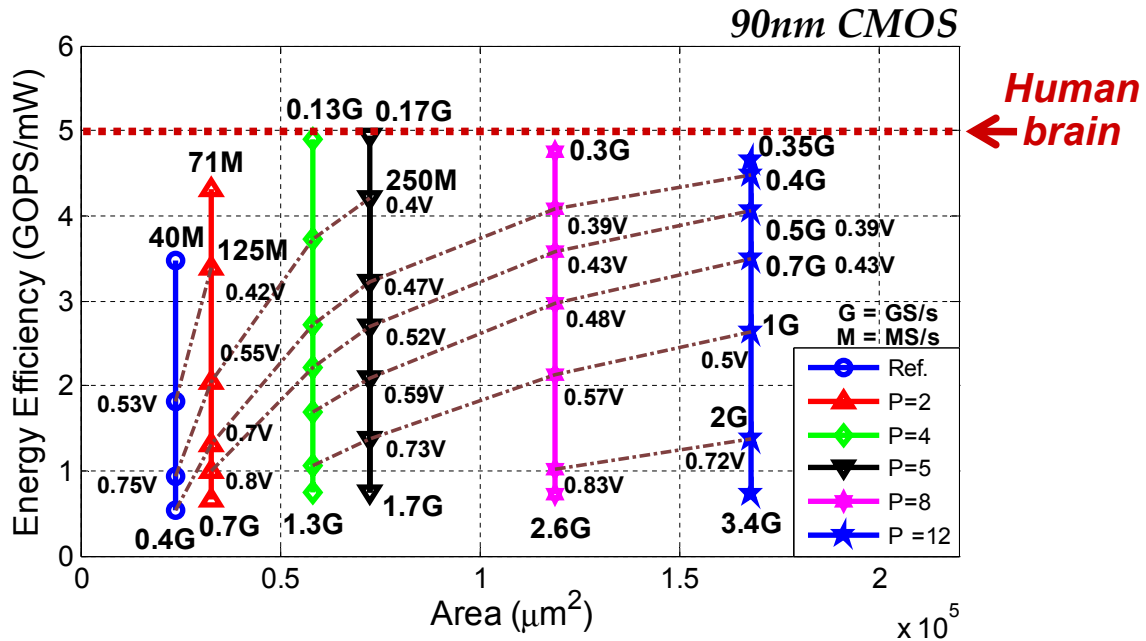
ILP Scheduling & Bellman-Ford Retiming: optimal + reduced CPU time

Optimization Flow

- ◆ Based on Reference E-D curve and system specs, fix degree of Retiming (R), Time-multiplexing (N) or Parallelism (P)
 - Generate optimized architectures/RTL in Simulink and synthesize in any given technology



Warm-up Example: 16-tap FIR

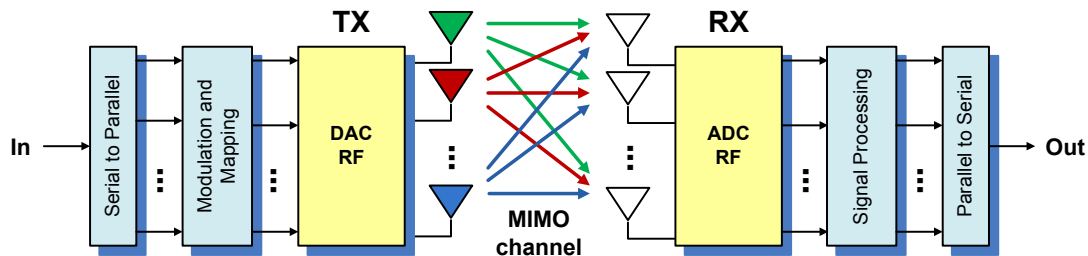


- ◆ Parallelism improves throughput or energy-efficiency
- ◆ 90nm CMOS can do neuromorphic computing

9

MIMO Communication

- ◆ MIMO used for range and rate increase



- ◆ Complex signal processing
 - Diversity algorithms (increased range)
 - Repetition, Alamouti scheme
 - Space-time coding
 - Spatial multiplexing algorithms (increased rate)
 - Bell Labs Layered Space Time (BLAST) algorithm
 - Singular Value Decomposition (SVD)
 - QR decomposition

10

Example: 4x4 MIMO SVD Chip

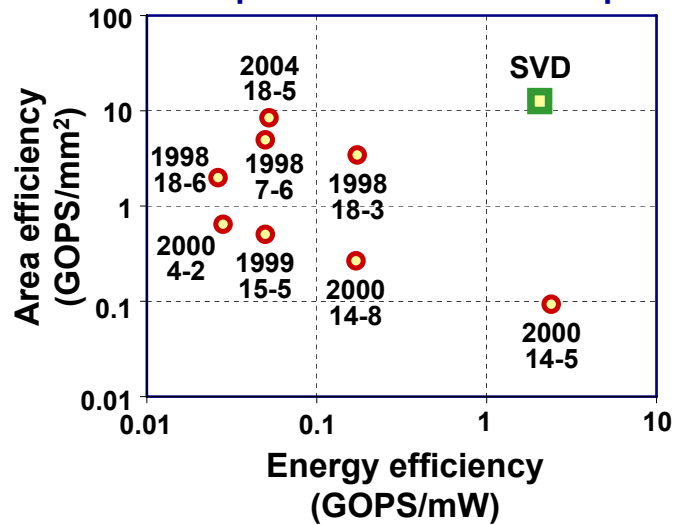
◆ Result of Energy-Area-Performance Optimization



(90nm ST Micro)

- ◆ **2.1 GOPS/mW**
 - 70 GOPS @ 100MHz
 - Power = 34mW
- ◆ **20 GOPS/mm²**
 - 70 GOPS in 3.5mm²

Comparison with ISSCC chips



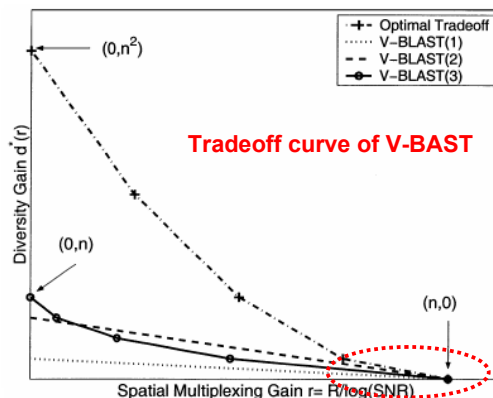
[D. Markovic, JSSC, Apr'07]

Next step: include flexibility for multi-mode operation

11

Diversity-Multiplexing Tradeoff

◆ Tradeoff curve in diversity-multiplexing space



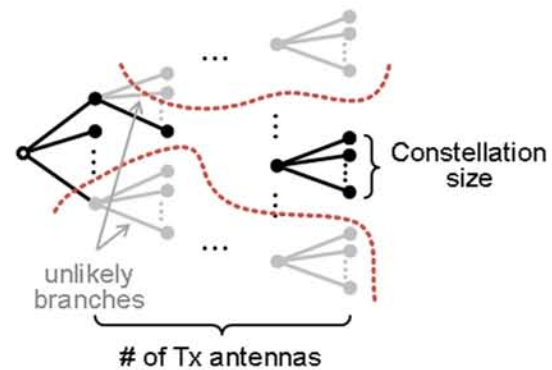
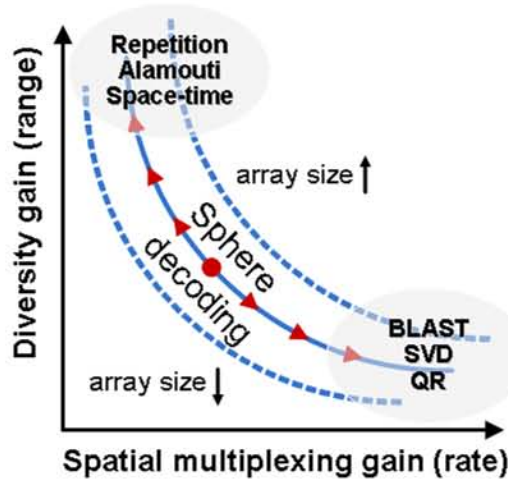
L. Zheng and D. Tse, "Diversity and Multiplexing: A Fundamental Tradeoff in Multiple-Antenna Channels," *IEEE Tran. Inf. Theory*, vol. 49, no. 5, pp. 1073-1096, May 2003.

Can we span the entire curve (unify point-wise solutions)?

12

MIMO Sphere Decoder

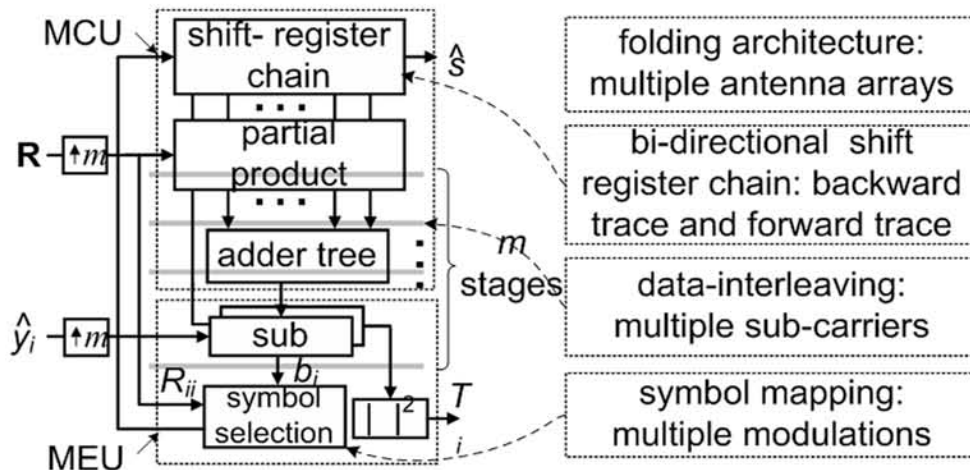
- ◆ Optimal trade-off in diversity-multiplexing space
- ◆ ML detection with polynomial complexity $O(N^3)$
- ◆ Flexible architecture required for multi-standard systems



Discarding a symbol removes all the nodes of its branch in the tree search space

Poster: Chia-Hsiang Yang

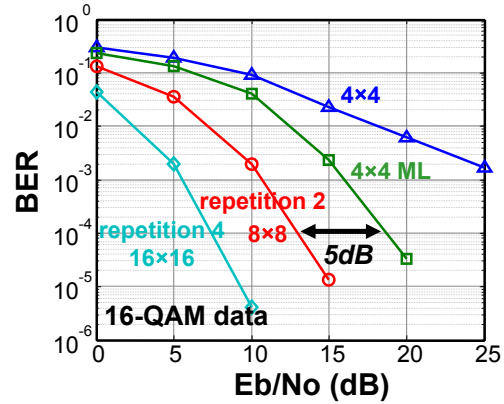
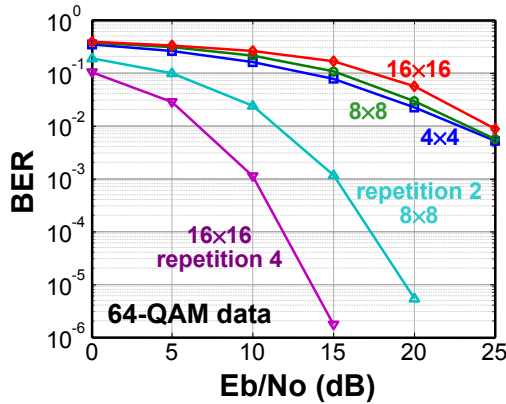
Scaleable PE Architecture



Parameter	Configuration Modes
Antenna array	Any square matrix # b/w 2×2 and 16×16
Modulation	BPSK, QPSK, 16-QAM, 64-QAM
# sub-carriers	16, 32, 64, 128
Detection	Depth-first, K-best

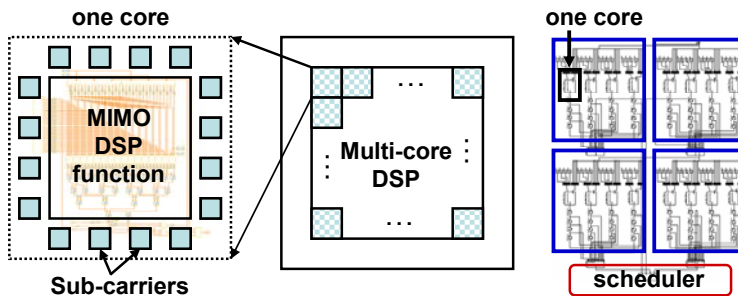
Hardware Emulation Results

- ◆ Comparable BER performance of 4×4, 8×8, and 16×16, with different throughput given a fixed bandwidth
- ◆ Repetition coding by a factor 2 reduces the throughput by 2×, but improves BER performance
- ◆ An 8×8 system with repetition coding by 2 outperforms the ML 4×4 system performance by 5dB



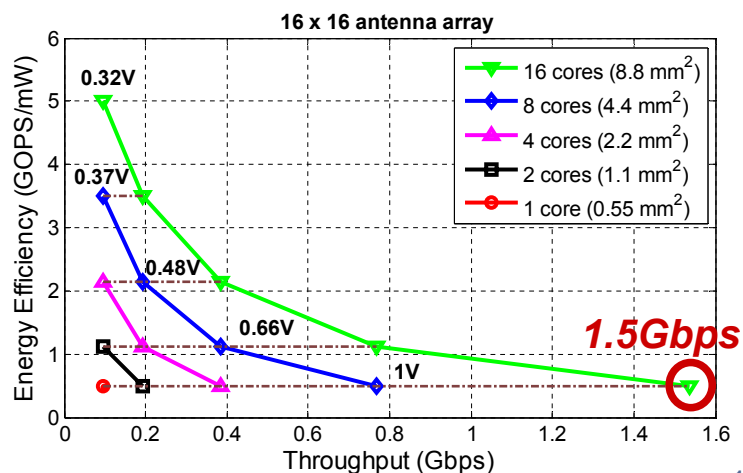
15

Multi-PE for Higher Throughput



- ◆ Hierarchical scheduling

- ◆ 58 pJ/bit
- ◆ 10x higher energy-effcyy than 1-PE

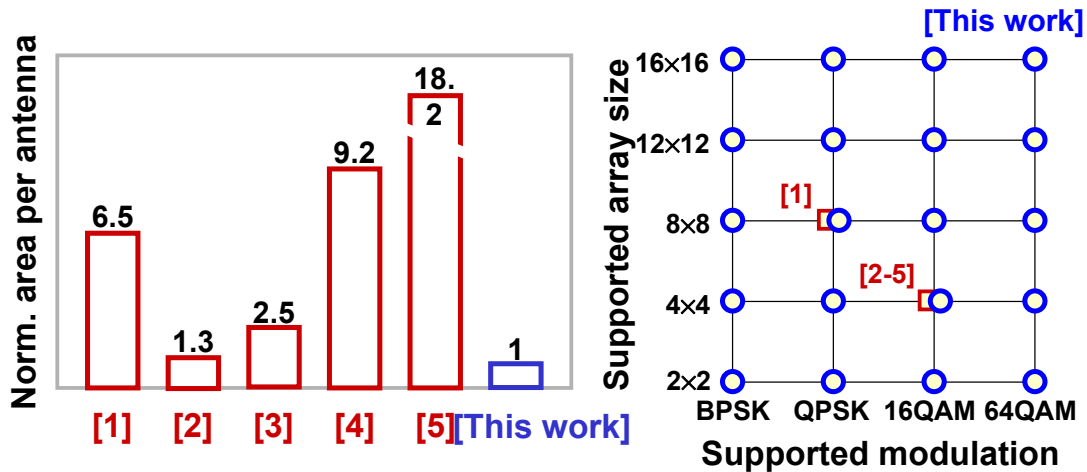


16

Comparison to Prior Work

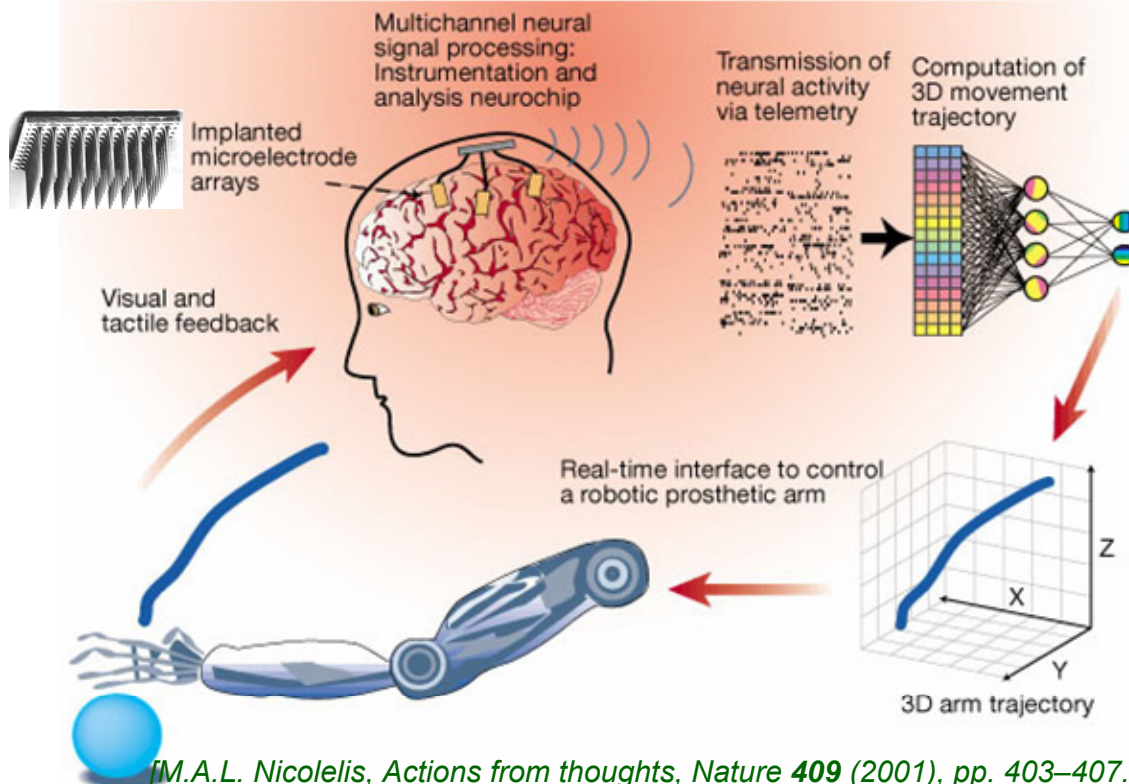
◆ Key features of the proposed architecture

- Highest reported area efficiency (normalized area per antenna)
- Highest reported antenna array size and constellation size
- Supports multiple sub-carriers and search methods
- The first hardware design to deploy the diversity-multiplexing tradeoff

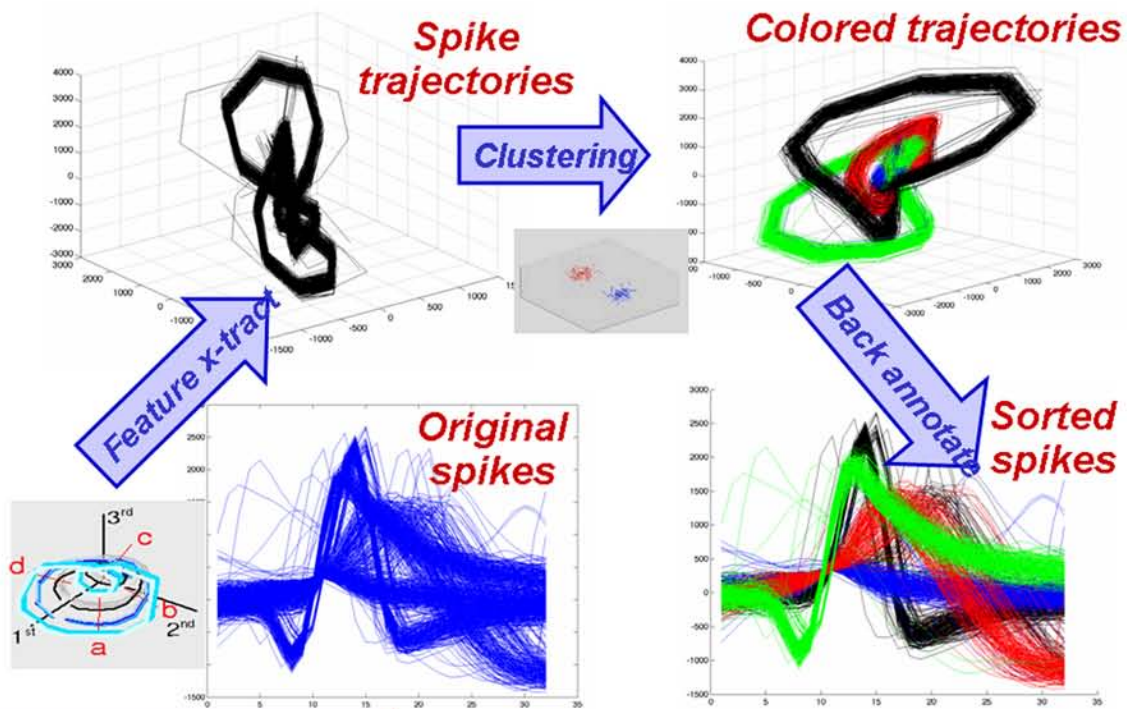


17

MIMO in Neuroscience

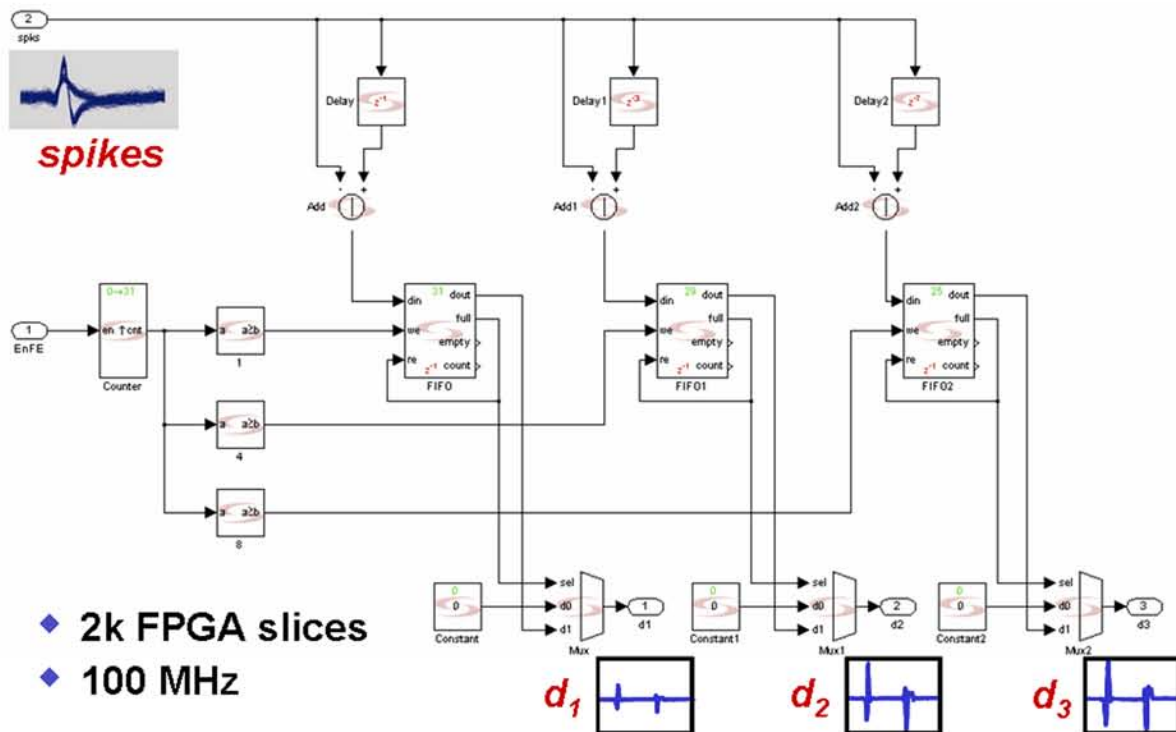


Spike Sorting in a Nutshell



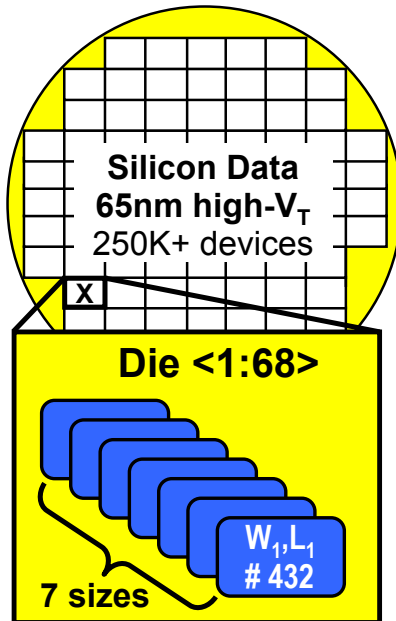
Poster: Sarah Gibson

Feature Extraction Algorithm in Hardware



- ◆ 2k FPGA slices
- ◆ 100 MHz

Mitigating Random Variation Effects



◆ Our approach

- Forget about curve fitting and **start with** the most accurate data provided: **current measurements**
- Develop **models for current variation** with respect to actual **design variables**: W , L and operating points (V_{gs} , V_{ds})

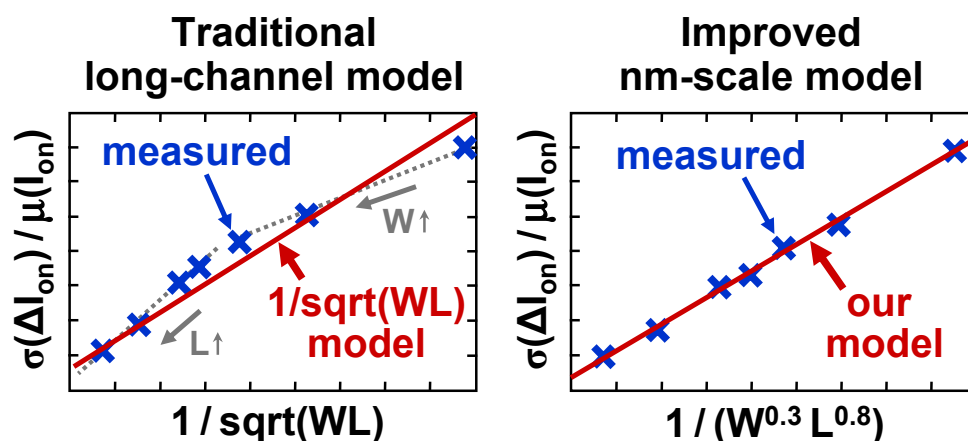
◆ Measured data

- 68 dies measured across a single wafer
- Measured transistors vary in width from 120-500 nm and length from 60-150 nm
- There are 7 unique W, L combinations
- Complete I-V curves were measured for 432 devices per W, L combination per die

21

Size Dependence of $\sigma(I_{on})$

- ◆ **Size dependencies:** standard deviation of current has a much stronger dependence on L than on W



- ◆ **PCA:** Principal Component Analysis is used to graphically explore size and operating point dependencies

[to appear at ISQED'08 & ISSCC'08 student forum]

Poster: Victoria Wang

22

Model for Random $\sigma(\Delta I_{on})$

$$\sigma(\Delta I_{on}) / \mu(I_{on}) = \frac{A}{W^\alpha L^\beta V_{gs}^\eta V_{ds}^\zeta}$$

$$\alpha = 0.3$$

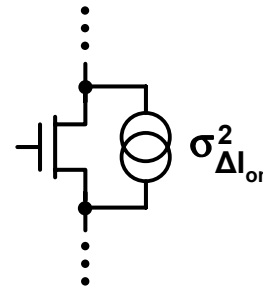
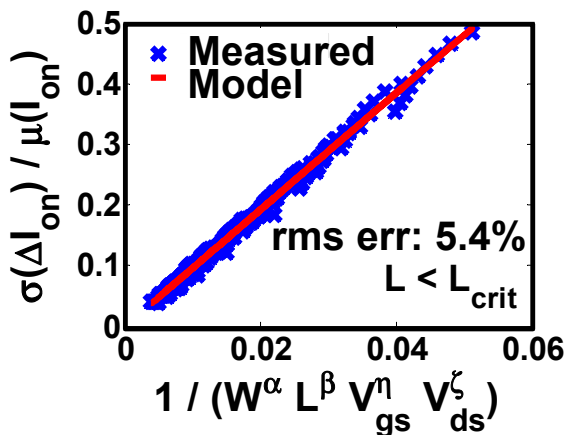
$$\beta = 0.8$$

$$\eta = 2.1$$

$$\zeta = \begin{cases} 0.15, & L < L_{crit} \\ 0, & L \geq L_{crit} \end{cases}$$

- Model has <5.4% error for all I-V points of all sizes

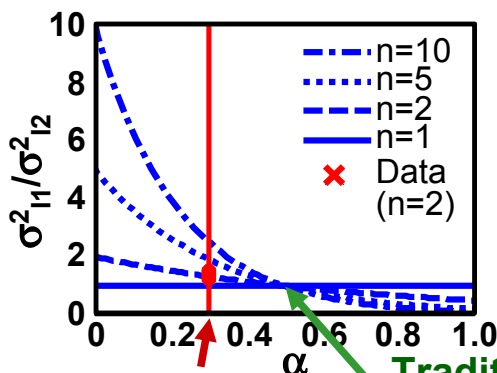
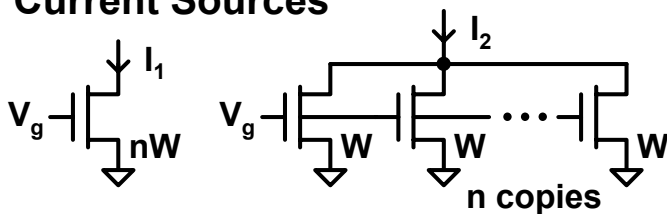
- Current variation can be modeled by a variability "noise" current source



23

Applications

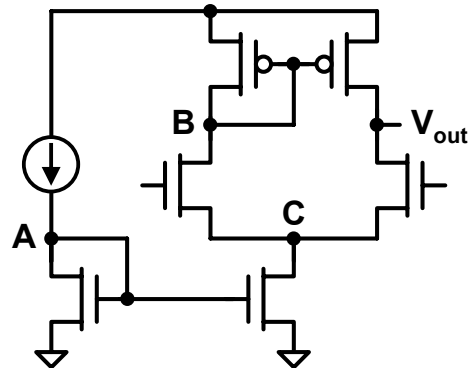
Current Sources



Measured silicon
 $\alpha = 0.3, \beta = 0.8$

Traditional
 $1/\sqrt{WL}$
formula

Differential Amplifier



Node (σ)	MC (mV)	Tool (mV)	Err (%)
A	16.0	16.0	0
B	26.0	26.57	2.2
C	20.9	21.08	0.86
Vout	196.4	187.3	4.6

24

Posters



◆ **Rashmi Nanda**

- High-Level Synthesis Automated Data Flow Graph Scheduling and Retiming



◆ **Chia-Hsiang Yang**

- A Flexible VLSI Architecture for Extracting Diversity and Spatial Multiplexing Gains in MIMO Channels



◆ **Sarah Gibson**

- A Hardware Architecture for Neural Spike Sorting



◆ **Victoria Wang**

- A Design Model for Random Process Variability and Applications