Looking Beyond CMOS: Integrated Circuit Design with Nano Electro-Mechanical Switches

Vladimir Stojanović (MIT) in collaboration with Tsu-Jae King Liu, Elad Alon (UC Berkeley) Dejan Marković (UCLA)
Acknowledgements

- Circuit design
  - Fred Chen, Hossein Fariborzi
  - Matthew Spencer, Abhinav Gupta
  - Cheng Wang, Kevin Dwan

- Device design
  - Hei Kam, Rhesa Nathanael, Vincent Pott, Jaeseok Jeon

- Sponsors
  - DARPA NEMS program
  - FCRP (C2S2, MSD)
  - MIT CICS
  - Berkeley Wireless Research Center
  - NSF
CMOS is Scaling, Power Density is Not

- Since ~2000 supply voltage (Vdd) stuck at ~1V
  - Leakage stops you from lowering threshold (Vth)
- $V_{dd}$ and $V_t$ not scaling well $\rightarrow$ power/area not scaling

Performance limited by power, not number of transistors
Parallelism to the Rescue

- Parallelism allows slower, more efficient units
  - Helps scale performance for fixed power
- Will this last forever?

IBM CELL processor

![Diagram showing normalized energy per operation (E/op) versus 1/throughput, with points indicating parallelization to recover performance and lower supply, balanced $V_{th}$, leading to reduced energy per operation.]

Today: parallelism lowers E/op
Leakage and sub-threshold slope define minimum energy/op for CMOS

Parallelism cannot reduce power if already operating at minimum energy
NEM Switches to the Rescue

- NEM switches show zero leakage & sharp sub-threshold slope
- Could potentially enable reduced E/op with scaling

NEM Switch Structure and Operation

ON Switch:

\[ |V_{gb}| > V_{pi} \] (pull-in voltage)

OFF Switch:

\[ |V_{gb}| < V_{po} \] (pull-out voltage)

<table>
<thead>
<tr>
<th>Poly-SiGe Gate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poly-SiGe Anchor</td>
</tr>
<tr>
<td>Poly-SiGe Beam</td>
</tr>
<tr>
<td>/Flexure</td>
</tr>
<tr>
<td>Tungsten Body</td>
</tr>
<tr>
<td>Tungsten Source/Drain</td>
</tr>
<tr>
<td>Tungsten Channel</td>
</tr>
<tr>
<td>Drain</td>
</tr>
<tr>
<td>Body</td>
</tr>
<tr>
<td>Source</td>
</tr>
<tr>
<td>Substrate</td>
</tr>
<tr>
<td>Insulator</td>
</tr>
<tr>
<td>Gate Oxide</td>
</tr>
<tr>
<td>Gate</td>
</tr>
<tr>
<td>Drain</td>
</tr>
<tr>
<td>90nm</td>
</tr>
<tr>
<td>Source</td>
</tr>
<tr>
<td>A-A' cross-section: off-state</td>
</tr>
<tr>
<td>Channel</td>
</tr>
<tr>
<td>A-A' cross-section: on-state</td>
</tr>
</tbody>
</table>
NEM Switch as a Logic Element

- 4-terminal design mimics MOSFET operation
  - Electrostatic actuation is ambipolar
- Non-inverting logic is possible
  - Actuation independent of source/drain voltages
Lumped Verilog-A model for circuit design and simulation:

- Mechanical dynamics: spring (k), damper (b), mass (m)
- Electrical parasitics: non-linear gate-body ($C_{gb}$), gate-channel ($C_{gc}$), and source/drain-body cap ($C_{s,db}$), contact ($R_{cs,d}$)
Simple electro-mechanical model matches experiment well
- Mechanical “pull-in” delay scales with geometry and overdrive voltage
- “Pull-in” voltage (threshold) scales with switch geometry

Constant E-field scaling
- Mechanical delay and $V_{PI}$ scale linearly

Digital Circuit Design with NEMS

- **CMOS**: delay set by electrical time constant
  - Quadratic delay penalty for stacking devices
  - Buffer & distribute logical/electrical effort over many stages

- **NEMS**: delay dominated by mechanical movement
  - Can stack ~100-200 devices before $t_{d,\text{elec}} \approx t_{d,\text{mech}}$
  - So, want all to switch *simultaneously*
  - $\rightarrow$ Implement logic as a single complex gate
Need to Compare at Block Level

- **Delay Comparison vs. CMOS**
  - Single mechanical delay vs. several electrical gate delays
  - For reasonable load, NEMS delay unaffected by fan-out/fan-in

- **Area Comparison vs. CMOS**
  - Larger individual devices
  - But often need fewer devices to implement same function

Example: NEMS Adder

- Full adder cell:
  - 12 NEMS vs. 24 transistors
  - XOR “free”
  - Complementary signals avoid extra mechanical delay (to invert)

- NEMS all sized minimally
N-bit NEMS Adder

- Ripple carry configuration
  - Cascade full adder cells to create larger complex gate

- Stack N NEMS, but still single mechanical delay

- Performance gap to CMOS shrinks to \(~10x\) for 32-bit add
  (10ns delay at 90nm)
Scaled NEMS vs. CMOS Adders

For similar area: >9x lower E/op, >10x greater delay

- Compare vs. Sklansky CMOS adder*
- 30x less capacitance
  - Lower device $C_g$, $C_d$
  - Fewer devices
- 2.4x lower $V_{dd}$
  - No leakage energy

For similar area: >9x lower E/op, >10x greater delay

F. Chen et al., “Integrated Circuit Design with NEM Relays,” *ICCAD 2008*

Parallelism helps

- Can extend energy benefit up to GOP/s throughput
  - As long as parallelism is available
- Area overhead bounded
  - CMOS needs to be parallelized at some point too
Contact Resistance

Energy/op vs. Delay/op across $V_{dd}$ & $C_L$

- Low contact R not critical
- Good news for reliability…
NEMS Circuit Demonstration

- Test Devices
- Logic
  - Adders, Compressors
- Timing Elements
  - Latches, FFs
- Memory
  - SRAM, DRAM
- I/O
  - ADC, DAC

Things We Learned…

- **Layout Matters**
  - Unbalanced current flow: flexures burn up!
  - Parasitic capacitors: can affect $V_{pi}$ (DIBL-like effect)
VTC looks digital, suggests composability
NEMS Latch Shows Composability

- Designed as if we used MOSFETS, but that is not always optimal...
Switch-based Carry Generation

- Demonstrates propagate-generate-kill logic as a single complex gate
Simultaneous read and write

Read latency = $\tau_{\text{mech, on (decoder)}} + \tau_{\text{mech, off (WL}_{\text{RD}}\text{ switch)}}$
Resistive divider based DAC
- 2-bit thermometer coded output
NEM switch scaling – next design

1um litho

Scaled NEMS size 20um x 20um

0.25um litho

NEM switch size 120um x 150um
Conclusions

- NEMS unique features enable scaling post-CMOS
  - Nearly ideal $I_{on}/I_{off}$
  - Switching delay largely independent of electrical $\tau$
  - Need to adapt circuit design style

- Reliability improving
  - Circuit level insights critical (contact R)
  - Demonstrated simple circuits
  - Can start thinking about building more complex systems

- Potentially order of magnitude lower E/op than CMOS
  - Next steps: scaling and improved device design, testing larger digital blocks