ECE 598NS: Machine Learning in Silicon Fall 2017

Energy-Delay Trade-offs

Naresh Shanbhag
Professor
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
the goal of this lecture is to understand the energy-delay tradeoffs of an architecture:

– computation
– storage & interconnect
• deep memory hierarchy
• high density → more latency, energy
• data movement → dominant cause of energy dissipation in machine learning applications

[Wong, Salahuddin, Nature Nanotechnology, 2015]
• cost per bit increases as memory approaches CPU
• energy-delay models of memory and interconnect needed

The Memory-Processor Bottleneck

Data access cost: ~4-5 pJ/word
Data processing cost: ~1 pJ/MAC

• large data volumes & large data access energy cost
• processor-memory bandwidth limitations

Energy estimates based on: 65nm CMOS, 8b word, 8b ALU, 512X256 array, 4:1 Col MUX) & [Horowitz, ISSCC 2015]
Power Breakdown

[Eyeriss, Y.-H. Chen, JSSC 2017]

- Data movement costs 45% of total power
- Computation costs about 10%
- Clock network = 33% (hidden cost of digital realizations)
Energy-Delay-Robustness Trade-off

- energy consumption vs. delay vs. robustness trade-off
- energy vs. delay trade-off captured via energy-delay product $EDP$
- energy/delay vs. robustness trade-off is more subtle (future lecture)
Energy-delay Trade-offs in Computation
- power-delay trade-off via supply voltage $V_{dd}$ scaling
- what is the relationship between $V_{dd}$, delay, power & energy?
\[ V_{dd} \text{ vs. } T_{cp} : T_p = \frac{C_{cp} V_{dd}}{k(V_{dd} - V_t)^\beta} \approx \frac{C_{cp}}{kV_{dd}} \] (for \( \beta = 2 \) and \( V_{dd} \gg V_t \))

\[ V_t = 0.8V \]

- \( C_{cp} \): total load capacitance of the critical path; \( V_t \): threshold voltage of the transistor
- \( P_D = \alpha C_L V_{dd}^2 f_{clk} \); (power-delay product) \( PDP = \alpha C_L V_{dd}^2 \)
- \( C_L \): total load capacitance of the circuit; \( \alpha \): activity factor
Pipelining for Low-Power

• basic idea: trade-off throughput increase from pipelining with power via supply voltage $V_{dd}$ reduction
\[ P_{\text{ref}} = C_{\text{ref}} V_{\text{ref}}^2 f_{\text{ref}} \]

delay doubles when supply voltage is reduced from 5V to 2.9V

\[ P_{\text{pipe}} = C_{\text{pipe}} V_{\text{pipe}}^2 f_{\text{pipe}} \]
\[ = (1.15C_{\text{ref}})(0.58V_{\text{ref}})^2 f_{\text{ref}} = 0.39P_{\text{ref}} \rightarrow 2.5X \text{ reduction} \]
• Other pipelining benefit: energy reduction due to glitches and hazard reduction.

• analysis ignores leakage power, and already reduced nominal supply voltages in modern process nodes

• in practice, this idealized scenario possible only for small values of $M$

• as $M$ increases and $V_{dd}$ reduces
  – exponential increase in delay $\rightarrow$ near/subthreshold operation
  – energy and delay overhead due to pipelining registers
Parallel Processing as a Low-Power Technique

- $P$-parallel blocks can operate at $\frac{V_{dd}}{L}$ supply. Therefore, the dynamic power is given by
- If $P=L \rightarrow L^2$ power reduction feasible

$$T_{cp} = \frac{C_L L}{k_n V_{DD}} = T_{block} = LT_s$$

$$f_{block} = \frac{f_s}{L} = P_{D, serial} \frac{P}{L^3}$$

$$P_{D, par} = PCL \left( \frac{V_{DD}}{L} \right)^2 f_{block}$$
\[ P_{ref} = C_{ref} V_{ref}^2 f_{ref} \]

Delay doubles when supply voltage is reduced from 5V to 2.9V

\[ P_{par} = C_{par} V_{par}^2 f_{par} \]

\[ = \left(2.15 C_{ref}\right) \left(0.58 V_{ref}\right)^2 \frac{f_{ref}}{2} = 0.36 P_{ref} \rightarrow 2.8X \text{ reduction} \]

Future computer systems promise to achieve an energy reduction of 100 or more times with memory design, device structure, device fabrication techniques, and clocking, all optimized for low-voltage operation.

By Ronald G. Dreslinski, Michael Wieckowski, David Blaauw, Senior Member IEEE, Dennis Sylvester, Senior Member IEEE, and Trevor Mudge, Fellow IEEE

• delay penalty from NTV to SubT is huge (100X)
• most gains energy savings obtained in going from nominal→NTV
• but….NTV (and SubT) suffer from increased sensitivity to variations → logic errors due to timing violations can occur
Near Threshold Computing

- Energy efficiency
- 5X increase in delay variation
- $10^5$X increase in memory failure rate
- Increased logic failures

400 mV $\leftarrow$ 1.1V

[Dreslinski-ProcIEEE10]
Barriers to NTV Operation

• Performance loss: 10X

• Performance variation
  – Defined as $\frac{\sigma}{\mu}$ of $T_p$
  – 1.3X (nominal) $\rightarrow$ 5X (NTV)
  – 2X (ripple) $\times$ 2X (temp)
  – Total: $5 \times 2 \times 2 = 20X$

variation in delay
Barriers to NTV Operation

- increased functional failures
- reduced $I_D$ due to increase in $V_t$ (timing failure)
- skewed $I_{DP}$ vs. $I_{DN}$ (write failure)
- local $V_t$ mismatch (read upset)

SRAM (memory) bit-cell failure rates

#### 4%

ECE ILLINOIS
Department of Electrical and Computer Engineering
NTV & Subthreshold Processors

- impact on frequency much more dramatic than energy
- nominal → NTV (0.5V): 7X energy & 11X frequency reductions

minimum energy operating point (MEOP) Subliminal Processor
• can these trade-off be captured analytically?

• need relationship between energy, delay, and $V_{dd}$
Minimum-Energy Operation Via Error Resiliency

Rami A. Abdallah and Naresh R. Shanbhag

• Total energy

\[ E_{tot} = N_g C_g V_{dd}^2 (\alpha + \beta L e^{-\frac{V_{dd}}{nV_T}}) \]

• \( N_g \): # of gates/blocks in the architecture
• \( C_g \): average load capacitance per block
• \( \alpha \): average activity;
• \( L \): number of gates/blocks in the critical path
• Constants: \( V_T = 26\text{mV} \) (thermal voltage); \( \beta \) and \( n \)
• Total energy

\[ E_{tot} = N_g C_g V_{dd}^2 (\alpha + \beta L e^{-\frac{V_{dd}}{nV_T}}) \]

• first term reduces quadratically with \( V_{dd} \)
• second term increases exponentially as \( V_{dd} \) reduces
• \( E_{tot} \) reduces initially with \( V_{dd} \) and then increases \( \rightarrow \) MEOP
• the MEOP 3-tuple \((V_{dd}^*, f^*, E^*)\) can be calculated numerically
• dynamic energy $E_{dyn}$ reduces as $V_{dd}$ decreases

$$E_{dyn} = \alpha N_g C_L V_{dd}^2$$

• leakage energy $E_{lkg}$ increases as $V_{dd}$ decreases

$$E_{lkg} = N_g V_{dd} I_s \left( \frac{W}{L} \right) e^{-V_t/nV_T} \left( 1 - e^{-V_{dd}/V_T} \right) T_p$$

• delay $T_p$ increases exponentially as $V_{dd}$ decreases

$$T_p = \beta \frac{C_{cp} V_{dd}}{I_s \left( \frac{W}{L} \right) e^{-V_{dd}-V_t/nV_T} \left( 1 - e^{-V_{dd}/V_T} \right)}$$
MEOP for a MAC

- $10b \times 10b$ MAC; 130nm CMOS;
- MEOP = $(V_{dd}^* = 330mV, f^* = 1.2MHz, E^* = 300fJ)$ for $\alpha = 0.3$
MEOP for FFT


- 16b, 1024-pt FFT processor, 180nm CMOS
- \( \text{MEOP} = (V_{dd}^* = 400mV, f^* = 10kHz, E^* = 1 \text{ (normalized)}) \)
Derivation of Energy Expression

- let the dynamic energy be given by
  \[ E_{dyn} = \alpha N_g C_g V_{dd}^2 \]
  where \( C_g \) is the average load capacitance per gate

- the leakage energy is given by
  \[ E_{lkg} = N_g I_{off} V_{dd} T_p \]
  where \( I_{off} = I_{D(sub)}(V_{GS} = 0, V_{DS} = V_{dd}) \)

- set \( T_p = T_{cp} \); i.e., the clock period is equal to the critical path delay, and \( C_{cp} = L C_g \), where \( L \) is the number of logic stages in the critical path
the delay is given by

$$T_p = \beta \frac{L C_g V_{dd}}{I_{on(sub)}}$$

where $I_{on(sub)} = I_{D(sub)}(V_{dd}, V_{dd})$, and $\beta$ is a fitting parameter needed to account for finite rise and fall times.

thus,

$$E_{lkg} = \beta L N_g V_{dd}^2 C_g \left( \frac{I_{off}}{I_{on}} \right)$$

where $\frac{I_{off}}{I_{on(sub)}} = e^{-\frac{V_{dd}}{nV_T}}$ in the subthreshold region

adding $E_{dyn}$ and $E_{lkg}$, we get

$$E_{tot} = N_g C_g V_{dd}^2 (\alpha + \beta Le^{-\frac{V_{dd}}{nV_T}})$$
Energy-delay Trade-offs in Storage
Energy & Delay Cost of Data Access

- large data volumes & large data access energy cost
- processor-memory memory bandwidth limitations

Static Random Access Memory (SRAM)

- used when large storage capacity is needed near processor
- slower than registers but energy efficient/bit
- column muxing (4:1, 8:1, 16:1) reduces bandwidth further
SRAM Bit-Cell (BC)

- cross-coupled inverter with access switches (MN3, MN4)
- small swings on bit-lines ($B$ and $\bar{B}$)
- sense-amplifier (SA) resolves voltage drop
SRAM Energy Model (Active Mode)

\[ E_{READ} = (C_{WL} + C_{cSEL})V_{dd}^2 + MC_{BL}V_{dd}\Delta V_{BL} + \left(\frac{M}{L}\right)C_{SA}V_{dd}^2 \]

\[ E_{WRITE} = (C_{WL} + C_{cSEL})V_{dd}^2 + \left(\frac{M}{L}\right)C_{BL}V_{dd}^2 + \frac{M(L-1)}{L}C_{BL}V_{dd}\Delta V_{BL} \]

\[ E_{LKG} = NMI_{LKG-BC}V_{dd}T_{ACC} \]
Delay Model

- delay is a complex function
- \( T_p = T_{ROW} + T_{BL} + T_{COL} \)
- \( T_{ROW} \): delay of row decoder and row drivers
- \( T_{BL} \): bitline discharge time
- \( T_{COL} \): delay of sense amplifiers and column decoder
- \( T_{BL} = \frac{C_{BL}\Delta V_{BL}}{k(V_{dd} - V_t)^2} \)
Energy-delay Trade-offs in On-chip Interconnect
• On-chip interconnect:
  – *local* interconnect for gate-to-gate or within compact circuit macros) level
  – *global* interconnect (busses, clock distribution networks)

• Off-chip interconnect → Chip I/O or serial links

• focus on on-chip interconnect

• global interconnects do not scale with technology node → chip sizes are steady (more functions are packed into the die)

• need to model delay and energy of interconnect

• Overall goal of system design should be to minimize data transfers over long interconnects.
Interconnect Stack

- Metal (Al/Cu) stack employed as interconnect
- Lower layers (wide & thin), upper layers (thicker)
  - Lower layers used for local connections. Upper layers for global connections

Intel’s 65nm node 8 metal layers; $\frac{t}{w} = 2$

[P. Bai, IEDM 2004]
Isolated Wire Delay & Energy Model

- simple RC model for an isolated interconnect
- delay is proportional to length squared ($l^2$):

$$T_p = 0.38\tau; \quad \tau = 0.5RC = 0.5rcl^2$$

- energy model of an isolated interconnect

$$E = CV_{dd}^2$$
Bus Delay & Energy Models

[Sotiriadis-ASPDAC01]

- multiple wires/lines running in parallel
- capacitance to ground \((C_L)\) and interwire capacitance \((C_I)\)
- \(C_I\) makes the delay of a bus data-dependent
Max delay case: $\uparrow\downarrow\uparrow\uparrow$ and $\downarrow\uparrow\downarrow\downarrow$

Min delay case: $\uparrow\uparrow\uparrow\uparrow$ and $\downarrow\downarrow\downarrow\downarrow$

- $T_{p0} = 0.38\tau = \text{delay of an isolated wire.}$

- $T_k = (1 + p\lambda)T_{p0}$: delay of the $k^{th}$ wire ($T_k$) depends upon the data transitions on the $(k - 1)^{th}$ and $(k + 1)^{th}$ wire (Miller effect)

- $\lambda = \frac{C_I}{C_L}$ is the coupling ratio which varies from 3 to 6 in 180nm CMOS.
Bus Delay Model

\[ T(\Delta) = \begin{cases} 
(1 + \lambda)\Delta_1^2 - \lambda\Delta_1\Delta_2, & \ldots \quad k = 1 \\
(1 + 2\lambda)\Delta_k^2 - \lambda\Delta_k(\Delta_{k-1} + \Delta_{k+1}), & \ldots \quad 1 < k < N \\
(1 + \lambda)\Delta_N^2 - \lambda\Delta_N\Delta_{N-1}, & \ldots \quad k = N 
\end{cases} \]

transitions on the \( k^{th} \) wire in the bus

\[ \Delta_k = 1 \ (0 \rightarrow 1) \]

\[ \Delta_k = -1 \ (1 \rightarrow 0) \]

\[ \Delta_k = 0 \ (1 \rightarrow 1 \text{ or } 0 \rightarrow 0). \]
Bus Energy Model

[Sotiriadis-TCAS03]

\[ E = \frac{C_L V_{dd}^2}{2} (u^f - u^i)^T C (u^f - u^i) = \frac{C_L V_{dd}^2}{2} \Delta^T C \Delta \]
• bus energy consumption is data dependent much like logic
• possibility of using coding techniques to reduce bus energy consumption
Summary

• energy and delay models of on-chip computation, storage, and interconnect help us evaluate energy efficiency and throughput of architectural alternatives
References
