Cerebras’ Third-Gen Wafer-Scale Chip Doubles Performance

March 13, 2024, Sally Ward-Foxton, EETimes. Cerebras has unveiled a third generation of its wafer-scale chip, offering 125 PFLOPS (at FP16 precision) from a single device. Given a single day, a four-chip installation could fine-tune Llama2-70B, while the biggest installations of 2,048 chips would be able to train it from scratch in the same time.

The wafer-scale engine 3 (WSE3) doubles the large language model (LLM) training speed of the WSE2, in the same 15kW power envelope and at the same cost point, Cerebras CEO Andrew Feldman told EE Times.

...The WSE also features 42 GB of SRAM with 21 PBytes/s memory bandwidth.

(Source: Cerebras)
Announcements

• Project
  • Midterm reports due next week
  • Preliminary design review after Spring break

• Homework 3 due next week
  • Quiz 3 after Spring break
Design Variability Sources and Impact on Design
## Systematic and Random Device Variations

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Random</th>
<th>Systematic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel Dopant Concentration Nch</td>
<td>Affects $\sigma_{VT}$ (^{[1]})</td>
<td>Non uniformity in the process of dopant implantation, dosage, diffusion</td>
</tr>
<tr>
<td>Gate Oxide Thickness Tox</td>
<td>Si/SiO$_2$ &amp; SiO$_2$/Poly-Si interface roughness(^{[2]})</td>
<td>Non uniformity in the process of oxide growth</td>
</tr>
<tr>
<td>Threshold Voltage $V_T$ (non Nch related)</td>
<td>Random anneal temperature and strain effects</td>
<td>Non-uniform annealing temperature(^{[5]}) (metal coverage over gate)</td>
</tr>
<tr>
<td>Mobility $\mu$</td>
<td>Random strain distributions</td>
<td>Systematic variation of strain in the Si due to STI, S/D area, contacts, gate density, etc</td>
</tr>
<tr>
<td>Gate Length L</td>
<td>Line edge roughness (LER)(^{[3]})</td>
<td>Lithography and etching: Proximity effects, orientation(^{[4]})</td>
</tr>
<tr>
<td>Fin geometry/ film thickness variations</td>
<td>Rounding, etc, $\sigma_{VT}$, mobility.</td>
<td>Systematic fin thickness Systematic Si film/BOX variations</td>
</tr>
</tbody>
</table>

Dealing with Systematic Variations

- Model-to-hardware correlation classifies unknown sources

Diagram:
- Systematic effect
  - Improve process: Limited options
  - Tighten a design rule: Density loss
- Model/Design-in: Extraction/Compact modeling/Design techniques
Chip Yield Depends on Inter-Gate Correlation

- Yield = Pr (sum of n delays < clock period)
- \( \rho = 0 \) gives highest yield through averaging

Non-correlated gates in a path reduce impact of variation

Bowman et al, JSSC, Feb 2002.
Chip Yield Depends on Inter-Path Correlation

Yield = Pr (max delay of K paths < clock period)

K = 1 gives highest yield

Correlated paths reduce impact of variation

Bowman et al, JSSC, Feb 2002.
Design Variability
Some Systematic Effects
Layout: Poly Proximity Effects

- Gate CD is a function of its neighborhood

Gate length depends on

- Light intensity profile falling on the resist
- Resist: application of developer fluid\(^1\), post exposure bake (PEB) temperature\(^2\)
- Dry etching: microscopic loading effects\(^3\)

---

Layout: Proximity Test Structures

- 90nm experiments
  - Single gate inverter layout
  - Stacked gates
  - L.T. Pang, VLSI’06

- Dummy poly

- 45nm experiments
  - No single gates allowed
  - P1 min max2
  - P2 mid1
  - P3 max1
  - P4 max2
  - L.T. Pang, CICC’08

- Ring oscillators and individual transistor leakage currents
Results: Single Gates in 90nm

- Max $\Delta F$ between layouts $> 10\%$
- Within-die $3\sigma/\mu \sim 3.5\%$, weak dependency on density
Results: Single Gates in 45nm

- Weak effect on performance. $\Delta F \sim 2\%$
- Small shifts in NMOS leakage and bigger shifts in PMOS leakage
Impact of Longer Diffusion in 45nm

- Strongest effect measured in 45nm, $\Delta F \sim 5\%$
- No significant shift in $I_{LEAK}$

**Figures:**
- RO Frequency
- NMOS $I_{LEAK}$
- PMOS $I_{LEAK}$

**Legend:**
- Slowest chip
- Fastest chip
- 22 chips from 2 wafers

**Diagrams:**
- P3
- D1

- Longer diffusion

**Notes:**
- Max1
Impact of Shallow Trench Isolation (STI)

- $\Delta F \sim 3\%$, small changes in $I_{\text{LEAK}}$
- Due to STI-induced stress

**Graphs:**
- RO Frequency
- NMOS $I_{\text{LEAK}}$
- PMOS $I_{\text{LEAK}}$

**Legend:**
- Yellow: Slowest chip
- Green: Fastest chip
- Blue: 22 chips from 2 wafers
Patterning and process impact on FinFETs

\[ \text{Delay} = R_{\text{wire}} \times C_{\text{pin}} + R_{\text{tr}} \times C_{\text{pin}} + R_{\text{tr}} \times C_{\text{wire}} + 1/2 \times R_{\text{wire}} \times C_{\text{wire}} \]

(a) Aggressive use (>90%) of 1X M routing

(b) Less aggressive use (~50%) of 1X M routing

Non-conformal Air Spacer Seal
In-situ doped regrown S/D epitaxial

Air Spacer
Metal Gate
Wrap-Around Contact

N7
Rtr
Cpin
Rwire
Cwire

Relative value to 14nm
0 1 2 3 4 5

14nm 10nm 7nm

10nm 7nm 7nm

14nm 10nm 7nm


Opt. Air Spacer

Remaining Battery Energy (Normalized)

Power at target perf. (arb. unit)

14nm
7nm opt.
7nm not opt.
0.3 0.5 0.7

Mobile SOC application usage

Video Stream (1.0 Hz)
Web Browse (7.5 Hz)
Email & Messaging (3.6 Hz)
Phone Call (1.0 Hz)
Gaming (1.5 Hz)
Standby (14 Hz)

78%

Song VLSI’15
Design Variability
Some Random Effects
Random Dopant Fluctuations

- Number of dopants is finite

Frank, IBM J R&D 2002
Processing: Line-Edge Roughness

• Sources of line-edge roughness:
  • Fluctuations in the total dose due to quantization
  • Resist composition
  • Absorption positions

Effect:
• Variation (random) in leakage and power
Transistor Matching

- $V_{Th}$ matching of geometrically identical transistors varies with size $\sim \sqrt{WL}$ and distance.


Pelgrom parameter $A_{VT}$
- Scales with technology (EOT)

$A_{VT} = 1.4 \text{mV} \cdot \mu \text{m}$
FDSOI example

“Short channel” sources: $\sigma_L$, $\sigma_{T_{Si}}$

“Surface” sources:
$Q_{ox}$, $\sigma_{\Phi_m}$, $\sigma_{T_{ox}}$, $\sigma_{T_{Si}}$, $\sigma_{\varepsilon_{ox}}$

- All the effects follow $1/\sqrt{WL}$ dependence

\[ V_t, \text{ long channel } [18] : \]
\[ V_t = \Delta \phi_{int} + \frac{kT}{q} \ln \left( \frac{2C_{ox}kT}{q^2n_iT_{Si}} \right) + \frac{\hbar^2 \pi^2}{2q\mu m^*T_{Si}^2} \]
with $\Delta \phi_{int}$ the gate WF with respect to intrinsic Si.

Oxide charges:
\[ \sigma_{V_t, Q_{ox}} \approx \frac{qT_{ox}}{\varepsilon_{ox}^{\frac{1}{2}} \sqrt{N_p + N_{ox}}} \]
\[ \sigma_{V_t, T_{ox}} \approx \frac{kT}{q} \alpha \left( \frac{\varepsilon_{ox}}{\varepsilon_{ox}} \right)^2 + \left( \frac{T_{ox}}{T_{ox}} \right)^2 \]

Oxide thickness and permittivity [19]: $\sigma_{T_{ox}}$, $\sigma_{\varepsilon_{ox}}$

$T_{Si}$ thickness: $\sigma_{T_{Si}}$
\[ \sigma_{H, T_{Si}} \approx \frac{kT}{q} \frac{\beta}{\sqrt{W.L}} \sigma_{T_{Si}} \]

Metal gate workfunction: $\sigma_{\Phi_m}$
\[ \sigma_{H, \Phi_m} \approx \frac{\gamma}{\sqrt{W.L}} \sigma_{\Phi_m} \]
with $a$, $\beta$, $\gamma$ spatial correlation lengths of the fluctuations
Negative Bias Temperature Instability

• PFET $V_{Th}$'s shift in time, at high negative bias and elevated temperatures

• The mechanism is thought to be the breaking of hydrogen-silicon bonds at the Si/SiO2 interface, creating surface traps and injecting positive hydrogen-related species into the oxide.

• Also other charge trapping and hot-carrier defect generation

• Systematic + random shifts

Tsujikawa, IRPS’2003
Random Telegraph Signal (RTS)

- Trapping of a carrier in oxide traps modulates $V_{th}$ or $I_{ds}$
- $\tau_e$ and $\tau_c$ are random and follow exponential distributions

RTS and Technology Scaling

- RTS exceeds RDF at 3 sigma with 20nm gates

\[ \Delta V_{th, \, RTS} \sim \frac{1}{WL} \]

\[ \Delta V_{th, \, RDF} \sim \frac{1}{\sqrt{WL}} \]

Tega et. al, VLSI Tech. 09
Memory
Random Access Memory Architecture

• Conceptual: Linear array of addresses
  • Each box holds some data
  • Not practical to physically realize
    – millions of 32b/64b words

• Create a 2-D array
  • Decode Row and Column address to get data
Basic Memory Array (From 151/251A)

• Core
  • Wordlines to access rows
  • Bitlines to access columns
  • Data multiplexed onto columns

• Decoders
  • Addresses are binary
  • Row/column MUXes are ‘one-hot’ - only one is active at a time

• Important to optimize the aspect ratio to balance the delays
Memory Banks

• Traditionally addressed by the LSB
  • Example two-bank memory
  • Odd and even banks
SRAM Cell Trends

![Graph showing trends in SRAM cell size and technology node](image)

- **ITRS Cell**
- **ITRS Eff. Cell**
- **Individual Cell**
- **Array Cell**
- **Eff. Cell**
SRAM Scaling or Not?

- TSMC at IEDM’19

- TSMC at ISSCC’20

**Bora’s spreadsheet/WikiChip**

![Graph showing SRAM cell size vs. year of publication/production](image1)

![Graph showing cell size vs. technology node](image2)

![Graph showing bit cell area vs. year](image3)
Summary

• Variability: Systematic and random
• Random, uncorrelated variations average out
• Identified random and systematic sources of variability
Next Lecture

- Memory