Micron Exits 3D XPoint Market, Eyes CXL Opportunities

Company looking at other technologies to build out a cost-effective memory hierarchy. Micron Technology’s exit from the 3D XPoint market should come as little surprise. The memory maker’s decision follows on several years of not talking much about a technology it jointly developed with Intel Corp.

Launched with a great deal of fanfare in July 2015, 3D XPoint sparked a lot of conversation about what the new class of memory technology might be able to do, as well as what it actually was.

By Gary Hilson, EETimes, 03.25.2021

Announcements

- Lab 5 due this week
- Assignment 3 due next week
Outline

• Memory wrap-up
• Power performance tradeoffs

4J 6T SRAM Alternatives
### 8-SRAM

- Dual-port read/write capability (register-file-like cells)
- \( N_0, N_1 \) separates read and write
  - No Read SNM constraint
  - Half-selected cells still undergo read
- Stacked transistors reduce leakage

---

### eDRAM

- Process cost: Added trench capacitor

---

L. Chang, VLSI Circuits 2005

Barth, ISSCC'07, Wang, IEDM'06
Crosspoint Memories


Fig. 2—Memory structure. $I_2$ and $I_4$ are access drive currents to core-selection switch. Presence or absence of a magnet over a twistor-strip solenoid crosspoint yields a “zero” or “one.” Signals observed between twistor and return wire.

Crosspoint Memories

- Neale, Nelson, Moore, Electronics’70
  - 16 x 16 array (256b) of ‘read-mostly memory’
Crosspoint Memory

- Four modes
  - Form
  - Set
  - Reset
  - Read

Endurance

3D Crosspoint Arrays

- Kau, IEDM’09
- Yeh, JSSC’15
- Ou, JSSC’11
Crosspoint Arrays

• Read and sneak currents

In the News…

• Intel Optane = 3D XPoint

- Protocol on top of electrical/mechanical interface for DSIM
- Allows for asynchronous command/data timing
- Controller uses request/grant scheme to communicate with host
- Data bus direction and timing controlled by host
- Command packet can request zero wait states for Intel® Optane DC persistent memory controller
- Transaction can be re-ordered in the Intel Optane
- Persistent memory controller
- 64/4 cache line granularity (similar to DDR4)
5. Low-Power Design
Importance of Power Awareness

• Energy: Crucial for Portable Applications
  • Determines battery lifetime
  • Amount of computation that can be performed
  • Performance is what sells products

• Power: Crucial for High-Performance Applications
  • Determines cooling and energy costs
  • Most designs today are power limited
  • Still need maximum performance

The Old Design Philosophy

• Maximum performance is primary goal
  • Minimum delay at circuit level
  • Architecture implements the required function with target throughput, latency
  • At circuit level, supplies, thresholds set to achieve maximum performance, subject to reliability constraints
  • Performance achieved through optimum sizing, logic mapping, architectural transformations
Constant Field Scaling Model

Traditional scaling model

\[
\text{If } V_{DD} = 0.7, \text{ and } \text{Freq} = \left(\frac{1}{0.7}\right),
\]
\[
\text{Power} = CV_{DD}^2 f = \left(\frac{1}{0.7}\right) \times (1.14^2) \times (0.7^2) \times (\frac{1}{0.7}) = 1.3
\]

Maintaining the frequency scaling model of 1990s

\[
\text{If } V_{DD} = 0.7, \text{ and } \text{Freq} = 2,
\]
\[
\text{Power} = CV_{DD}^2 f = \left(\frac{1}{0.7}\right) \times (1.14^2) \times (0.7^2) \times (2) = 1.8
\]

While slowing down voltage scaling

\[
\text{If } V_{DD} = 0.85, \text{ and } \text{Freq} = 2,
\]
\[
\text{Power} = CV_{DD}^2 f = \left(\frac{1}{0.7}\right) \times (1.14^2) \times (0.85^2) \times (2) = 2.7
\]

2001 Picture: Power As a Problem

Power delivery and dissipation will be prohibitive
The New Design Philosophy

- Maximum performance is too power-hungry, and/or not even practically achievable
- Extract maximum performance under a power/energy envelope
- Excess performance (as offered by technology) to be used for energy/power reduction

Trading off speed for power

5.A Power and Energy Basics
Portability: Battery Limits

- Little change in basic technology
  - store energy using a chemical reaction
- Battery capacity doubles every 10 years
  - Has slowed down
- Energy density/size, safe handling are limiting factor

<table>
<thead>
<tr>
<th>Energy density of material</th>
<th>KWH/kg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gasoline</td>
<td>14</td>
</tr>
<tr>
<td>Lead-Acid</td>
<td>0.04</td>
</tr>
<tr>
<td>Li polymer</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Battery Progress

- First Commercial Use
- NiCd
- SLA
- NiMH
- Li-Ion
- Reusable Alkaline
- Li-Polymer
- Energy Density (Wh/kg)
- Trend Line
5.B Power-Performance Tradeoffs
Know Your Enemy

- Where does power go in CMOS?
- Switching (dynamic) power
  - Charging capacitors
- Leakage power
  - Transistors are imperfect switches
- Short-circuit power
  - Both pull-up and pull-down on during transition
- Static currents
  - Biasing currents

Summary of Power Dissipation Sources

\[ P \approx \alpha \cdot (C_L + C_{CS}) \cdot V_{swing} \cdot V_{DD} \cdot f + (I_{DC} + I_{Leak}) \cdot V_{DD} \]

- \( \alpha \) – switching activity
- \( C_L \) – load capacitance
- \( C_{CS} \) – short-circuit “capacitance”
- \( V_{swing} \) – voltage swing
- \( f \) – frequency
- \( I_{DC} \) – static current
- \( I_{Leak} \) – leakage current

\[ P = \frac{\text{energy}}{\text{operation}} \times \text{rate} + \text{static power} \]
CMOS Performance Optimization

- Reminder - sizing: Optimal performance with equal fanout per stage

- Extendable to general logic cone through ‘logical effort’
- Equal effective fanouts \( (g_{i+1}C_i) / C_i \) per stage
- Optimal fanout is around 4

![Diagram of CMOS circuit with predecoder and word driver](image)

[Ref: I. Sutherland, Morgan-Kaufman'98]

Performance Optimization

Energy

Increasing performance increases power!

Delay = \( 1 / \text{Performance} \)
**Performance Optimization**

Energy

- Microarchitecture A
- Microarchitecture B

\[ \text{Delay} = \frac{1}{\text{Performance}} \]

**Power-Performance Optimization**

Achieve the highest performance under the power cap

![Diagram showing energy per operation (E_{\text{op}}) vs. delay (D)]

- \( E_{\text{max}} \)
- \( E_{\text{min}} \)
- \( D_{\text{min}} \)
- \( D_{\text{max}} \)

Unoptimized design
Achieve the highest performance under the power cap
How far away are we from the optimal solution?

Global optimum – best performance
Power-Performance Optimization

Maximize throughput for given energy or Minimize energy for given throughput

• There are many sets of parameters to adjust
  • Tuning variables
  • Circuit
    (sizing, supply, threshold)
  • Logic style
    (std. cell library, …)
  • Block topology
    (adder: CLA, CSA, …)
  • Micro-architecture
    (parallel, pipelined)
Power-Performance Optimization

- There are many sets of parameters to adjust
  - Tuning variables
  - Circuit (sizing, supply, threshold)
  - Logic style (std. cells, custom, …)
  - Block topology (adder: CLA, CSA, …)
  - Micro-architecture (parallel, pipelined)

Globally optimal power-performance curve for a given function

Energy-Delay Sensitivity

\[ S_A = \frac{\partial E/\partial A}{\partial D/\partial A} \bigg|_{A=A_0} \]
\[ \Delta E = S_A \cdot (-\Delta D) + S_B \cdot \Delta D \]

At the solution point all sensitivities should be equal.