CHALMERS
UNIVERSITY OF TECHNOLOGY

Using Many Small 1T1C Memory Arrays in a Large and Dense Multicore Processor

Using Many Small 1T1C Memory Arrays in a Large and Dense Multicore Processor

Gunnar Carlstedt
Department of Computer Science and Engineering,
Chalmers University of Technology
gunnar@carlstedt.se

Mats Rimborg
Department of Computer Science and Engineering,
Chalmers University of Technology
mats@rimborg.se

ABSTRACT

A memory system for multicore processors with a large number of processing elements (PE) is presented. Each PE has a local memory implemented in one-transistor one-capacitor (1T1C) DRAM technology, and these local memories contain many small memory arrays. The energy consumption and access time are reduced compared to state-of-the-art dynamic memories, while the aggregate bandwidth is increased by orders of magnitude.

The memory arrays are composed of $4F^2$ cells, with ≤128 bit-lines and word-lines. The area overhead for peripheral circuitry is minimized. The word-line driver is a two-transistor demultiplexer, and due to how short the word-lines are, small transistors can be used. The sense amplifiers are multiplexed 4-to-1 for use by several bit-lines.

The sense amplifiers are controlled with low voltage current injection to a bus. The address is represented as a combination of eight 1-of-N encoded parts, and their cross products select memory bank, sector, array half, and word-line.

The design space is explored by varying the number of banks, sectors, bit-lines and word-lines to find the most ideal combination for a 14 nm technology with 52 nm pitch. Considerable differences in performance measured as area, energy and access time have been found. The optimal constellation provides high area utilization (63 %), low energy consumption (25 fJ/bit) and short access time (515 ps).

Finally, a future memory cell at the end of Moore's law is predicted, and its implications for a new emerging computer paradigm, the Surface Based Processor (SBP).

CCS CONCEPTS

• **Hardware**; • **Integrated circuits**; • **Semiconductor memory**; • **Dynamic memory**;

KEYWORDS

dynamic memory, memory system, multicore processor, sense amplifier, word-line driver

ACM Reference Format:
Gunnar Carlstedt and Mats Rimborg. 2022. Using Many Small 1T1C Memory Arrays in a Large and Dense Multicore Processor. In *2022 International Syposium on Memory Systems (MEMSYS 2022), October 03–06, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3565053.3565057

MEMSYS 2022, October 03–06, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9800-8/22/10.
https://doi.org/10.1145/3565053.3565057

1 INTRODUCTION

This article presents a memory system adapted to a computing device that is based on a multitude of active memories with processing capability. One-transistor one-capacitor (1T1C) DRAM technology is currently mainly used for large memories, where low cost per bit is a major design goal. Such memories use long bit-lines and wide words to reduce overhead for sense amplifiers and word drivers.

There is a dichotomy between processor and memory, where the processor generally requests data accesses much faster than a DRAM can fulfil. To rectify this, several levels of cache memory, provided with complex coherence mechanisms, are used. By reducing the memory array size, it can be made faster and better adapted to the compute characteristics, thereby eliminating the need for a cache. This is almost necessary in multiprocessors with thousands of cores.

There are few reports about the use of small 1T1C memory arrays. Some such have been used as eDRAM in L3 caches [1, 2]. Small memory arrays are otherwise almost absent in the literature. With $4F^2$ architecture, there are even higher requirements on the peripheral circuits.

In this article, a memory system that consists of a huge number of small memory arrays is evaluated. It may be used to replace conventional DRAM chips, or, even better, be part of a multicore processor with a huge number of nodes.

The Surface Based Processor (SBP) is a new such emerging general purpose computer architecture [3]. The SBP may contain hundreds of thousands of nodes, thus providing a huge computing capability but also requiring an extremely high memory bandwidth. The proposed implementation is a mesh-based package switching network (NoC) between small local memories, see Figure 1a. These are implemented by very small memory arrays of 1T1C DRAM cells, see Figure 1b. Accesses are performed in parallel via the nodes in the NoC.

The contribution of this work is an investigation of the design space for very small DRAM memory arrays. To achieve this, the following features have been used:

- A memory architecture using *small arrays*.
- A *multiplexing sense amplifier* for $4F^2$ memory arrays.
- A system to *generate word-line signals*.
- A high speed *memory* with low energy consumption.

The evaluation is performed with a memory described by its parameters and simulated in PSpice. The calculations are based on a 14 nm technology with 52 nm pitch, and the design space of all

parameters is assessed to obtain the memory characteristics. Finally, the parametrized memory model is used to predict the performance at the end of Moore's law, applied to memories.

The remainder of the article is structured in the following way. *Section 2* is an overview of the memory. In *section 3*, the 1T1C memory cell is discussed. The address and data networks are covered in *section 4*.

The memory array is described in *section 5*. *Section 6* is focused on low level control and timing, and *section 7* describes the results.

In *section 8* a future memory cell is predicted, *section 9* contains some related work, and our conclusions are presented in *section 10*.

2 OVERVIEW

2.1 Memory System

A memory area can be organized as a large number of word cells divided into arrays, see Figure 1a. Each cell needs a control signal for word-access, and input/output connections. A set of processing elements (PE) can be distributed over the same area, with connections between the PEs and the memory arrays. On this level of abstraction, a package switching network (NoC) transports a set of control signals to the memory cells, and data to and from these. A PE may send an address that activates the control-signal of a specific memory word. In the case of a *read*, the content of that word is then transported back to the PE on the network. For a *write*, the PE also provides the data to be written.

On an intermediate level, there are transports over hierarchical buses within the arrays, see Figure 1b where a *read* is shown.

On the lowest level, the transports are separated through address decoders and bit-lines, see Figure 1c. Analog signals are amplified and adjusted due to charge sharing on the second lowest bus level *M*. The smallest possible memory array consists of the particular set of memory cells, the bit-lines *BL*, the word-lines *WL*, the word drivers, an intermediate bus *M*, and the sense amplifiers connected to a bus *S*.

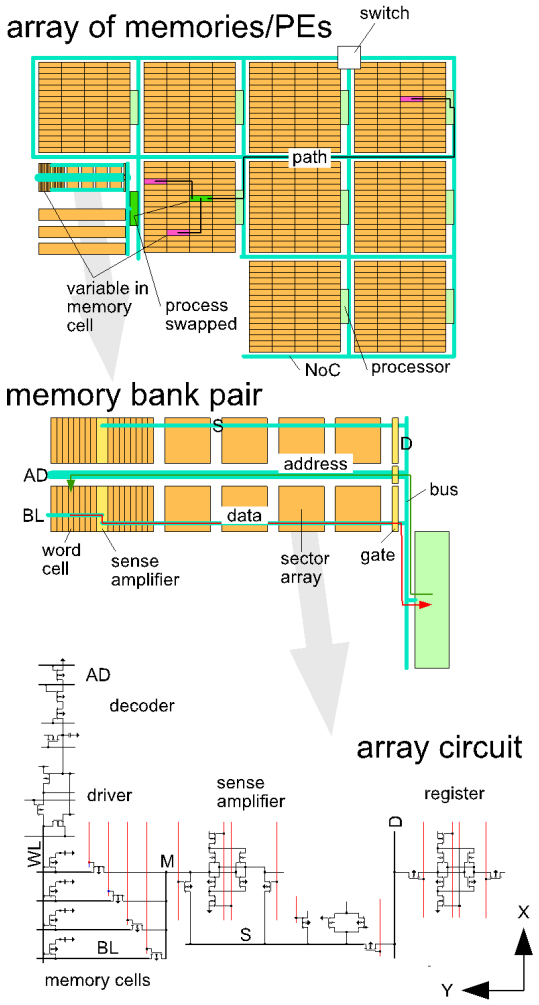

Figure 1: The NoC and the memory cells with the memory array. From top to bottom: (a) Memory cells, (b) Sense amplifier, and (c) Word drivers and wires.

2.2 Maximizing Performance

The maximum number of parallel accesses in a memory system is equal to the minimum of the number of local bus networks and the number of PEs. In some systems several processors may share access to the memory, which limits the parallelism, but in other cases (such as the SBP), each PE has its own local memory.

To enable accesses, there has to be a *path* along links and buses between a PE and a word cell, see Figure 1a. In the bus part, there is only one possible way between a leaf (memory word cell) and the root (node). In the package switching network, there are many alternative paths, and the shortest paths are monotonic in both X and Y directions. For access, an address is transported along a path from the PE to the memory array, and in either direction, a word read or to be written. Within the memory array, there is a continuation to the word-lines and along the parallel bit-lines.

Assume that all buses and links except for their wires are ideal, using no area, with no delay time, and no energy consumption. The energy on a bus/link is then used to charge the entire bus/link (or link segment), and thus proportional to its length. In the bus network, the mean transport length is half the bus length. Therefore, the energy consumption is double compared to a link. The transport delay is also proportional to the length of the bus/link, including the bit-lines. However, the switching speed is limited by the conductor time constant (resistance × capacitance), which is proportional to the square of the wire length.

The bus system really only needs to have a two-dimensional structure, in the X and Y directions, and both directions should reach the memory array periphery. An additional level of smaller buses will increase the path length, and therefore the local bus system should preferably have only one bus in each X and Y direction. The bit-lines constitute such an additional bus, and it should be minimal.

Buses should only be active along the path, to prevent superfluous energy consumption. The energy on the buses and links are proportional to the signal amplitude, and they may use very low voltage. A bit-line of the 1T1C memory cell on the other hand may

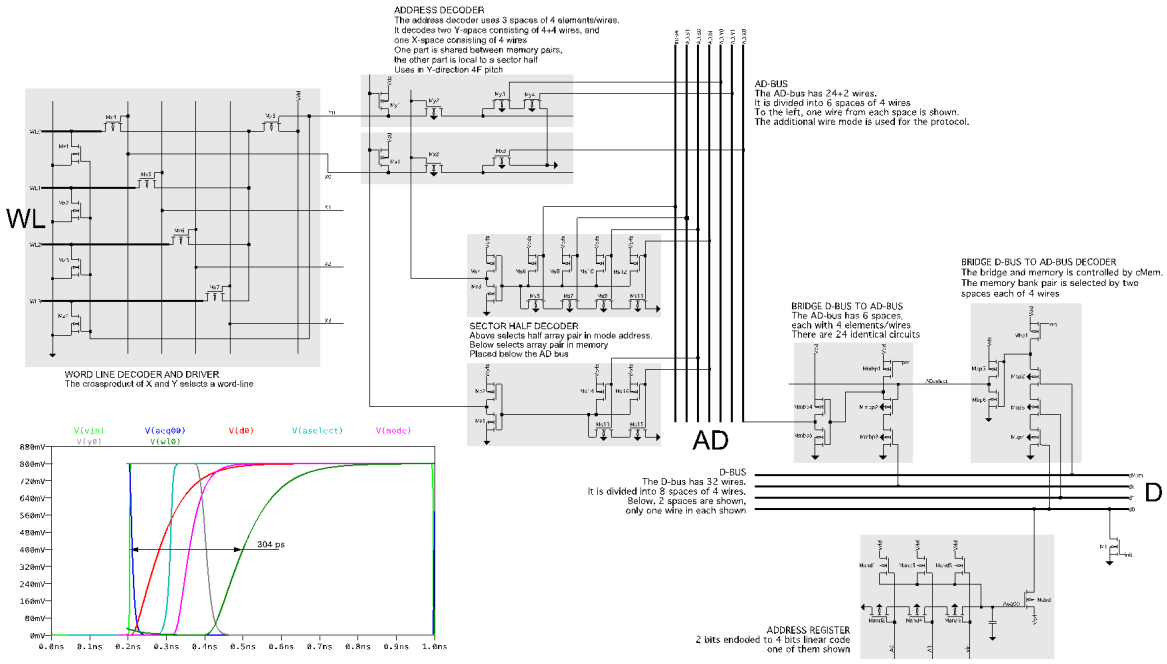

Figure 2: A demultiplexer implementing the driver for 4 word-lines together with the address decoder system. The signal flow is from right to left. The *D* bus is connected to 16 memory bank pairs. Each pair uses a bridge decoder between the *D* and *AD* buses. Each decoder contains 24 bus drivers, in total 384 drivers. There are 16 instances of the *AD* bus. The sector half decoder and address decoder together select one out of 4 word-lines. Each word-line is connected to 64 bit-lines, and each word contains 32 bits. The total size of the memory array is 131,072 bits.

use almost full swing voltage. These should therefore only be a small fraction of the path.

The PEs generate an access pattern, depending on the executed program, with a certain distribution of path lengths. The aggregate performance is the sum of the path lengths, compensated for voltage. The overall performance, the access time and the energy consumption have equal optimization goals, and are not conflicting. An optimal memory system therefore consists of a mesh with short links and many small memory arrays.

2.3 Optimizing Area Utilization

A $4F^2$ architecture is used. The linear dimensions allow only one transistor, one wire, or one via to be packed side by side.

2.3.1 Word-lines. The generally used precharging is an energy consuming method. Instead, the function can be an *AND* between two address parts, where switching should rise the selected wire and reset it afterwards. A single transistor may perform *AND*, and another transistor reset, see Figure 2. Both are placed along the word line continuation. Each driver needs to be reached by an address bus. The word-line length overhead is the sum of the width for these transistors and address wires, each with 2F pitch. The overhead is reduced by interleaving several words along the word-line, and sharing the address between two neighboring banks.

2.3.2 Bit-lines. The dynamic memory requires amplifying and adjusting the bit-line voltage. A six-transistor power-switched flip-flop sense amplifier is smallest and fastest, but still too large here. Several bit-cell contents are fed to their bit-lines, store the analogue voltage in the bit-line capacitances, and subsequently multiplex a single sense amplifier to several bit-lines, see Figure 1c. The multiplexer consists of one transistor per bit-line and may typically be 4-1 or 8-1.

2.3.3 Structure and Buses. The address is transported on a bus *AD* and data on a sector bus *S*, which is parallel to the bit-lines while physically in different places, see Figure 1c. A third perpendicular data bus *D* is used to connect them externally. The *S* bus is placed above the memory cells and *AD* outside, causing overhead. This is reduced by abutting two arrays in the X direction and sharing the address bus *AD* between two banks, as described above.

3 1T1C MEMORY CELL

A conventional DRAM cell is composed of one transistor and one capacitor (1T1C). There have been several DRAM models and simulators published. Some only support existing designs, such as DRAMSim2 [4] and USIMM [5]. The Ramulator [6] was a cycle accurate simulator to be modified for its purpose. DArT [7] is a somewhat more accurate circuit simulator, but is considered discrete.

The 3-D-DATE [8] has a more precise circuit description. It uses models for modern transistors such as VCAT, and wires. It also uses

a single level 1-of-N code. However, the sense amplifier, driver and overall architecture differ from those described in this article.

Therefore, we have used a model described in SPICE. Parameters are extracted and a more accurate description is created from these, and loaded into a PSpice simulator. The design space is evaluated based on a series of individual simulations.

3.1 Wires

The word-line wires are evolving from metal to polysilicon with the wire resistance ~100 Ω/μm and gate capacitance ~70 aF [8]. A 128 bits wide $4F^2$ memory has a wire time constant ~230 ps. This can be reduced to 8 ps by strapping every 16^{th} memory cell [9], *i.e.* using two parallel wires, one metal and one polysilicon.

A bit-line consists of a wire and the drains of the access transistors. A typical total capacitance is ~200 aF per memory cell. A 64-word bit-line has the capacitance ~13 fF. Legacy memory cells may have a storage node capacitance ~15 fF [10, 11]. The quotient of bit-line capacitance and storage node capacitance is ~1, which makes the sense amplifier simple.

3.2 Refresh Rate

The leakage current consists of the sub-threshold current of the access transistor, and tunneling currents through the transistor and the capacitance. The tunneling current is proportional to the area and extremely dependent on how thin the insulator is made. The sub-threshold current depends on the electric field, and is highly affected by the geometric properties [12]. These properties are local to the memory cell and do not depend on the size of the memory array.

The required refresh interval of the worst bit is generally three orders of magnitude less than for an average bit, and proportional to the storage node capacitance [13]. A typical refresh time is 64 ms according to JEDEC specifications. The physical size of a storage capacitance is an obstacle for technology improvement. It may however be reduced considerably, thus making the memory cells much simpler to fabricate, by reducing the size of the memory arrays. This will increase the bit-line refresh rate, but since there are fewer bit-lines still decrease the overall memory refresh rate.

3.3 Transistors

There are no extreme requirements on the transistor conductivity. The entire memory array and its peripheral circuits can be implemented by just three transistor types with the same basic structure, one for the storage node, and where the other two are n-, and p-channel devices.

4 ADDRESS AND DATA NETWORKS

The memory arrays contain two “transport systems” that are very different. One is a decoder for a numeric address and the other transports data between the bit cells and the bit-lines. The decoder uses a set of address parts whose cross products are used both as selectors and for clock distribution to the word-lines. The data is transported on low voltage, low capacitance, wires.

4.1 1-of-N Encoding and Decoding

The decimal number 13 in the range 0–15 is represented by 4 bits as 1101 in ordinary binary code. A 1-of-N code uses one element per state, so the range 0–15 consists of 16 elements. The value 13 is represented as a one in the 14^{th} position and the remaining are all zeroes. When incrementing a number, this code needs fewer switch transitions than a binary code, which results in less energy consumption.

Representing large ranges is impractical due to the required number of positions. Instead, several spaces can be combined where the cross product indicates a value. The range 0–63 can be represented in two spaces as (0–1; 0–31), (0–3; 0–15), or (0–7; 0–7). This will require 2+32=34 wires, 4+16=20 wires, and 8+8=16 wires respectively. Generally, a binary number has fewer bits than a 1-of-N code, but a binary decoder uses both the normal and inverse representation of each bit, and optionally a clock. Binary code therefore needs 3, 5 and 7 wires for 2, 4 and 8 positions 1-of-N codes. In a 1-of-N decoder the cross product can be used as clock, and there will be no glitches since only one wire goes high at any transition.

The set of word-lines is a representation of the memory address. Two spaces 1-of-N codes have been used in [8, 14].

4.2 Address Bus Decoder

A proposed memory system based on 16 bank pairs is shown in Figure 1b. Each bank pair has 16 memory arrays on its left and right sides respectively, *i.e.* 512 arrays. Each array stores 2×64 words for a total storage capacity of 65,536 32-bit words, with 128 bit-lines. Eight 1-of-N spaces are used, see Figure 2.

- 2 spaces (0–3; 0–3) select one of 16 memory bank pairs and drive the remaining spaces to the *AD* bus,
- 2 spaces (0–3; 0–3) select one of 16 vertical pairs of memory arrays,
- 1 space (0–3) selects a word on either side of a sense amplifier and on either side of the *AD* bus,
- spaces (0–3, 0–3, 0–1) select a word for a bit-line.

4.3 Word-Line Electrical Characteristics

The word-line drivers are divided into 16 demultiplexers, where each one uses an element of a space as input, and takes the address from another space. The demultiplexer is implemented with one p-channel transistor *Mx4–Mx7* in series with *My5* used for the rising transition and one n-channel *Mz1–Mz4* for the falling transition to an idle state, see Figure 2. The spaces are controlled by a dynamic register and stored in capacitances. A word-line has the capacitance 14.5 fF, and the propagation delay all the way from the address register is 304 ps, see Figure 2. The total access time is 515 ps.

4.4 Noise Immunity and Crosstalk

The address decoder wires are very long and implemented with low capacitive conductors that are vertical balk structures. Crosstalk is imminent during fast switching. Open circuit (current driven) sources transport the full swing to the neighbors. Low resistance drivers share the cross-talk currents through their mutual capacitance.

A *ONE* signal disturbs a *ZERO* wire with a voltage amplitude less than 33 percent of the swing. At the boundary between two spaces,

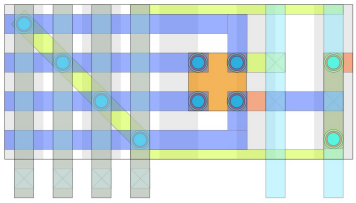

Figure 3: The layout of a demultiplexer that drives four word-lines.

a *ZERO* wire may be surrounded by two *ONEs* causing a noise rate of 67 percent, thus injecting an error. If the width of the CMOS n-transistor is doubled, the noise will be reduced to 40 percent. This avoids crosstalk errors and is the proposed implementation.

4.5 Layout

The transistor *My* is split into one transistor for each word-wire, see Figure 3, and the word-line driver has a width of 18F. The four vertical address wires are routed within the memory array, causing an increased memory array height. The transistors for selecting memory array are placed physically below the *AD* bus.

- The array column has 64 word-lines that are 2F high each, in total 128F.
- The demultiplexer column has 16 demultiplexers, each 8F high and one space for the x0–x3 wires, each 2F high, or in total 136F.
- The address decoder column has 16 address decoders for the Y wires, each 4F high, and 4 other address decoders for X wires, each 4F high, or in total 80F. These are all placed physically below the AD bus.

The three columns are placed beside each other, and the total height is determined by the highest one, *i.e.* the demultiplexer column, which is 136F.

4.6 Data Network

The data network starts at the sense amplifier, passes the *S* and *D* buses to the *A* sensing flip-flop. The *S* and *D* buses are the longest wires in the memory system and have a considerable capacitance. A current around 26 μA is fed into the capacitance during a short time, causing a small ramping voltage, see Figure 4 and Figure 5.

5 MEMORY ARRAY

The energy consumption of a memory array consists mainly of charging all bit-lines and buses. Their aggregated length is proportional to the number of bit cells. Conventional 1T1C memories use square memory array structures with 256–2048 elements per side. The approach here is instead a much smaller 32–128 × 64–128 structure. The charge relation of bit-line to storage node capacitances is reduced considerably by simplifying the sense amplifier. The reduced bit-line capacitance results in fast switching times and low energy consumption.

However, the area overhead for sense amplifier and word driver has to be considered. One sense amplifier is used to sequentially

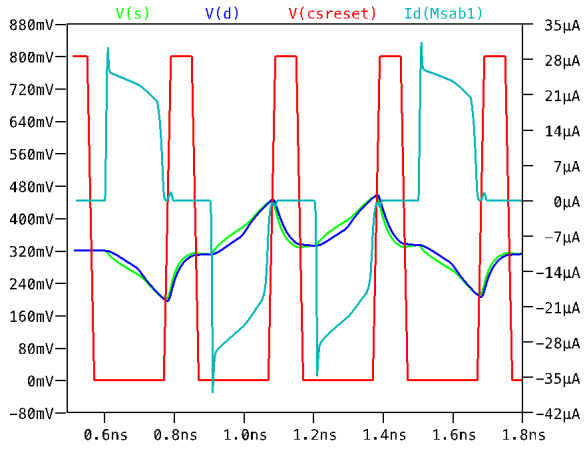

Figure 4: Waveforms on the bit-lines as seen on the *S* and *D* buses.

multiplex several bit-lines. Only one row of sense amplifiers is necessary for $4F^2$ memory arrays. The word-line driver is implemented as a single transistor AND-gate driven by two 1-to-N coded address spaces, as described earlier.

The memory array consists of a central sense amplifier connected to the S bus via two transistors *Msa1* and *Msab1*, a flip-flop *Mff1–Mff4* power controlled by the transistors *MffVdd* and *MffGnd*, on one side of the flip-flop the multiplexing transistors *MbL0–MbL3*, on the other side *MbL4–MbL7* (not shown), and two arrays of memory words consisting of 1T1C bit-cells, see Figure 5. A layout of the central part is shown in Figure 6.

The transistor count has been reduced as much as possible and the circuit is symmetric. Accesses are made to either side of the sense amplifier, and selected transistors are used depending on which side. Equalization is performed symmetrically by connecting all internal nodes via the *S* bus to an intermediate voltage *Vi*. The memory works in two phases, *idle* and *access*, that are divided into one address phase and four (or possibly eight, depending on the interleaving) multiplexing phases. During the access phase one word-line is selected and all of its access transistors are conducting. Charge sharing is performed between the storage node and a bit-line.

5.1 Memory Operations

There are four operations during a multiplexed phase:

5.1.1 Idle. The *D* bus is separated and all other nodes: *BL, M, Mb,* and *S* are set to the intermediate voltage *Vi*. Transistor *MS1* conducts setting the *S* bus, transistors *Msa1, Msab1* and all transistors *MbL0–MbL7* conduct. Only leakage currents are consumed.

5.1.2 Read. Read operates in the phases *reset, charge sharing*, and *amplify*. One word-line is used.

During the read phase transistors *MD1, MS1, Msa1* and *Msab1* conduct, setting the *M, Mb, S* and *D* buses to the intermediate voltage. The duration is determined mainly by discharging the *S* bus. Transistors *MS1* and *MD1* are common for all sectors and memory banks, respectively.

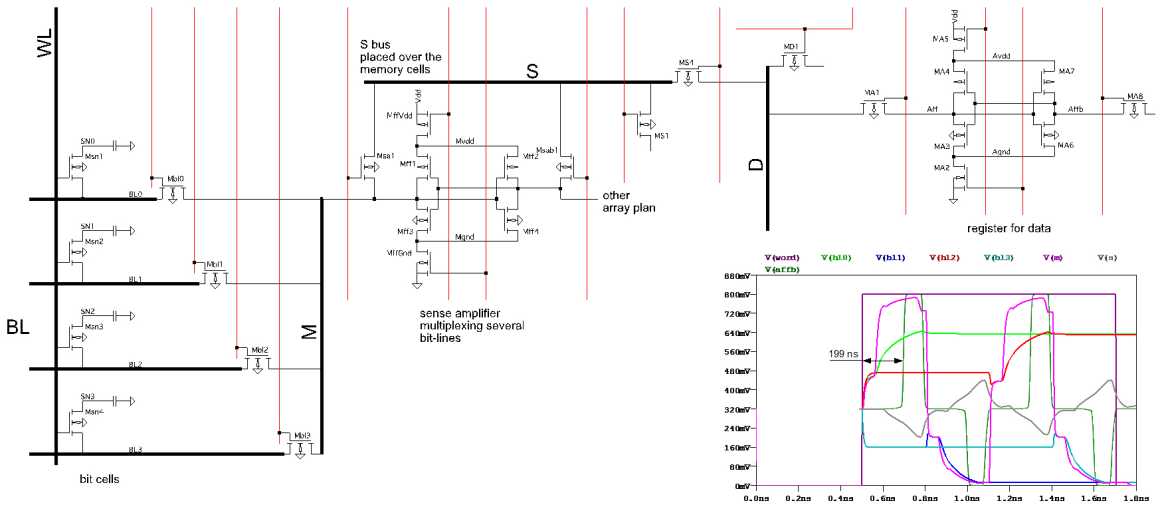

Figure 5: Schematics of an interleaved part of a memory array. The circuit is symmetric, but only the part to the left of the sense amplifier is shown.

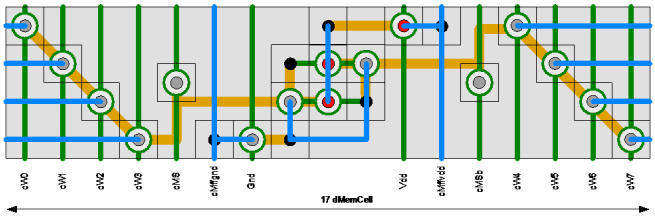

Figure 6: A sketch of the layout of a sense amplifier with multiplexer.

During the *charge sharing* phase one of the transistors *MbL0–MbL3* is conducting.

During the *amplify* phase transistors *MffVdd, MffGnd, Msab1, Ms4* and one of *MbL0–MbL3* are conducting. The *Mb* bus is clamped by a moderate resistance of transistor *Msab1*. During the rise time of *cMffGnd* and fall time of *cMffVdd*, transistors *Mff1–Mff3* increase their conductance. Either transistor *Mff1* or *Mff3* is fully conducting, charging *M*, a bit-line and its storage node. The length of the phase determines the *ONE* voltage level of the storage node. Simultaneously transistor *Mff2* or *Mff4* are conducting together with *Msab1*. An almost constant current flows into the *S* and *D* buses causing a ramping voltage. The shortest duration is set by charging the bit-line, and otherwise by charging the *S* and *D* buses. A *read* is followed by a total of three *refresh* or other *reads*.

5.1.3 Refresh. All upper or lower memory arrays in all memory banks and sectors may refresh one and the same word-line during four multiplexed phases. It is performed almost in the same way as the *read* operation. During the amplification phase, the transistor *Msab1* and *MS4* are not conducting.

5.1.4 Write. *Write* operates (differently from *read*) in the phases *reset, charge sharing, amplify* and *normalize*. One word-line is used.

During the reset phase, transistors *Msa1, Msab1, MS1* and *MD1* are conducting. Transistors *MffVdd* and *MffGnd* are not conducting. The buses *M, Mb, S* and *D* are set to the intermediate voltage *Vi*.

During the *charge sharing* phase, *Msa1, MS1 and MD1* are not conducting. Transistors *MA1, MS4* and *Msab1* are conducting, charging the *D, S* and *Mb* buses with an almost constant current causing a ramping voltage.

During the *amplify* phase *MffVdd, MffGnd* and *Msab1* are conducting. The flip-flop toggles to a proper level.

During the *normalize* phase one of the transistors *MbL0–MbL3* are also conducting. A high current flows into the bit-line.

5.2 Quiescent Voltages

During *read, refresh* and *write* operations, the amplifier causes the storage node voltage to asymptotically go to a quiescent voltage, one for each of the *ZERO* and *ONE* states. The transistor used during the transition to *ZERO* is fast with a quiescent voltage 0 mV. In the other direction the transistors *Msn* and *MbL* are source followers causing a slow charge, see Figure 7. The *M* bus switches fast, and the major delay is caused by the multiplexer transistor *MbL*.

5.3 Read Delay

Figure 8 shows the time delays during a read operation. The period of amplification controlled by *cmffgnd* and *cmffvdd* must be longer than delay #6. The A register may be strobed after the end of delay #5. The access time from the word-line rise depends on the tolerances of the timing of the control wires. It is not greater than the sum of #1, #2, #3, #5 above, 288 ps. The refresh period is longer than the sum of #2 and #3, and longer than #2 and #6, 275 ps. By controlling the pulse generation well, the access time may be shortened.

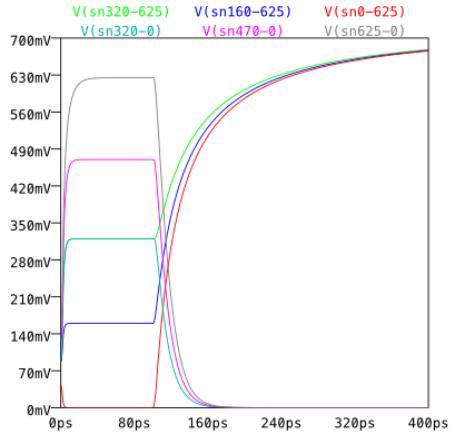

Figure 7: Amplification period time as function of a quiescent voltage. Storage node and bit-line capacitances are 3 fF and 5 fF respectively.

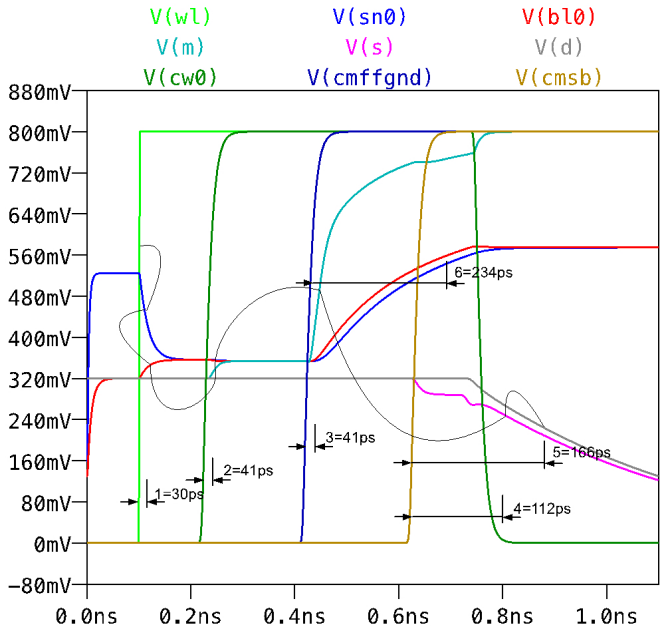

Figure 8: Time delays during a read operation. For clarity, space has been added between the phases.

6 LOW LEVEL CONTROL AND TIMING

The circuits around the sense amplifier use control for *Vdd* and *Gnd*, connection to *S* bus, and the multiplexer, in total 12 wires, see Figure 5. They are all time critical and needed after the word line has become active.

These wires are implemented by letting the wires on the *AD* bus work in two modes, as address decoder and memory array controller.

Each of the control wires has a driver consisting of an inverter with several (typically 4) parallel n- and p-channel transistors (not shown in the figure). It is driven by a gate consisting of only one inverter, with the input from one wire on the *AD* bus and the power from a signal indicating that a sector is selected. The pitch between the control wires is in most cases 2F. A three-dimensional wiring is used.

6.1 Separately Processing Bank Pairs

The access time for a word is short, compared to the long cycle time of a memory bank. Each memory bank pair may execute in semi-parallel, where an access to a particular bank pair is delayed until a complete access cycle has elapsed. Having many bank pairs generally causes a negligible waiting time.

6.2 Time Generator

There is no space for a local time generator. It is instead shared by placing it centrally or within each bank pair. It contains a register that rotates the group of four multiplexer control wires and holds the phases of an access. During each phase it controls three wires to the sense amplifier. The time generator is not further described here.

7 RESULTS

The researched result is the performance measured in area, energy and time. The purpose of this study is not to describe a complete product, but a structure with certain properties.

The discussed design space includes all appropriate combinations of number of bit-lines {64, 128}, interleaving {4}, number of words {32, 64, 128, 256}, number of sectors {8, 16}, and number of bank pairs {8, 16, 32}. The whole memory is analyzed. The drain capacitance of a bit-line depends on the memory cell technology, and 100 aF/bit is presumed. If the X width is greater than 1.4 times the Y height, the memory is “folded” halfway, to make the sides as even as possible.

A fictive theoretical process inspired by [15] is assumed. It has three conductor layers where the topmost is a low resistance and capacitance wire. Their characteristics from bottom to top for pitch (nm) / resistance (Ω/μm) / and capacitance fF/μm are 52/44/0.29, 52/44/0.29, and 120/2.7/0.15. Above these are two layers for *Vdd* and *Gnd*. The n- and p-transistors are symmetrical with threshold voltage 70 mV and saturation current 1.3 mA/μm. The $4F^2$ memory cells have 52 nm pitch. Its implementation is unaffected by the explored design space, and thus not considered here.

Time and energy are almost proportional to the switched capacitive energy. The access time as function of energy consumption for all memory variants is shown in Figure 9. It shows an almost straight line that verifies this property.

7.1 Energy

The energy consumption depends, as earlier discussed, on the wire lengths. Short wire lengths in the memory array results on the other hand in large overhead for control, buses, sense amplifiers and drivers. The energy consumption for the memories in the design space as function of their area utilization is shown in Figure 10a. The cost decreases with high area utilization. The lowest energy increases almost linearly with increasing area utilization. There are many variants with much higher energy, but the energy is still

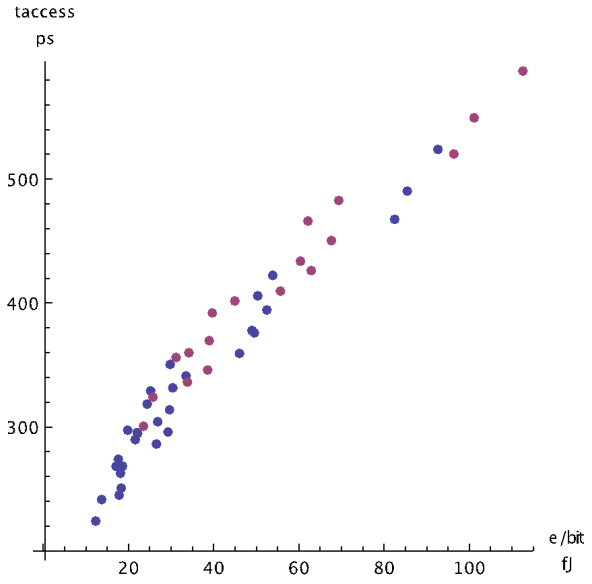

Figure 9: The energy consumption for one memory *read* access as function of the energy consumption for the design space. Bit-line additional drain capacitance is 100 aF/transistor. Red dots indicate an area usage greater than 60 percent, which is desired.

almost three orders of magnitude less than for modern DRAMs [16].

The energy plotted as function of the memory size is shown in Figure 10b. The energy is almost proportional to the square root of the size. For an optimal area of memory cells, the mean path length along X and Y is proportional to the square root of the memory size. This relation holds for any type of planar memory. It may be extended to three dimensional memories. A *read* access has a path going both forth and back.

There is a best low energy wire using full swing that has the energy consumption el per length. The memory cell side is *dmc*. Such an access uses the energy:

$$eref = 2 \times \sqrt{size} \times dmc \times el$$

The energy consumption relative the reference energy is shown in Figure 10c. The outliers are those with extreme values.

7.2 Access Time

The access time is plotted as function of the area utilization in Figure 11a. Because the access time and energy consumption are proportional to the switched charge amount, the access time has the same relationship as the energy consumption. The access time depends on the memory size, see Figure 11b. It is almost two orders of magnitude less than for modern DRAMs.

The memory bandwidth is basically proportional to the inverse of the access time. However, several accesses may be performed in parallel in different parts of the memories. The bandwidth per area as function of the energy consumption is shown in Figure 11c.

7.3 Optimal Memories

Low access time, high bandwidth and low energy consumption are in conflict with high area utilization. Generally, the area utilization should be high for low cost. A requirement has to be set, for example ≥60 % area usage, ≤700 ps access time and ≤40 fJ energy consumption. There are only three combinations that fulfill this, see Table 1. The difference in area usage is not high. Therefore, the first one is preferred since it provides double bandwidth, due to the smaller size of the arrays and that it therefore are twice as many of them.

The energy dissipated in the various nodes is shown in Table 2. The dominating energy is the control of the sense amplifiers via the ıbus followed by the ıbus. The sense amplifiers read 4 words but only one is used. The bit-line energy dissipation is only 14.9 %. The other energies within the memory array are very low. It can be concluded that address decoding uses 242 fJ and sensing 447 fJ, the energy inside the array 134 fJ for a total of 823 fJ. (In Table 2, a few minor contributors have been omitted.)

The power density is the energy divided by the access time and the memory area. For the optimal memory it is 8.8 W/cm^2.

The energy dissipation depends on the bit-line capacitance. In addition to the ideal MOS transistor there is capacitance in the drain region, mainly caused by the drain diffusion. The energy consumption for the memory as function of this capacitance is shown in Figure 12.

8 DISCUSSION

This section contains predictions about the future based on current trends and insights. But as the saying goes, it's difficult to make predictions, especially about the future, so we are speculating and no one really knows for sure.

8.1 1T1C at the End of Moore's Law

The 1T1C memory cell will evolve and survive the end of Moore's law. The electrical components such as wires, transistors and capacitors will asymptotically approach certain limits. The end of the development curve is expected when the dimensions approach the physical limits and it is no longer economically motivated to push even further. The purpose of this section is not to design a new memory cell, but to find proper parameters for the performance of 1T1C memories at that time. While we are speculating, the electrical characteristics of small components are to some extent known, while the structure and fabrication are not fully known.

We believe the main difference from the current trend will be the use of a very small storage node capacitance.

The structure will be based on a checker board layout where every second row is displaced. The memory cell itself is a vertical device. A horizontal footprint should contain a via (with 4 coaxial layers), a transistor (6 layers) and a capacitor (5 layers). They need to be vertical, tube-like, structures with a center, an isolation, an optional wire, and a surrounding. The thickness 5 nm seems to be a limit where dielectrics deteriorate [17], mobility is reduced considerably [18], and lithography is very complex. The transistor size can to some extent compensated for by placing the transistor of every second word-line in another layer.

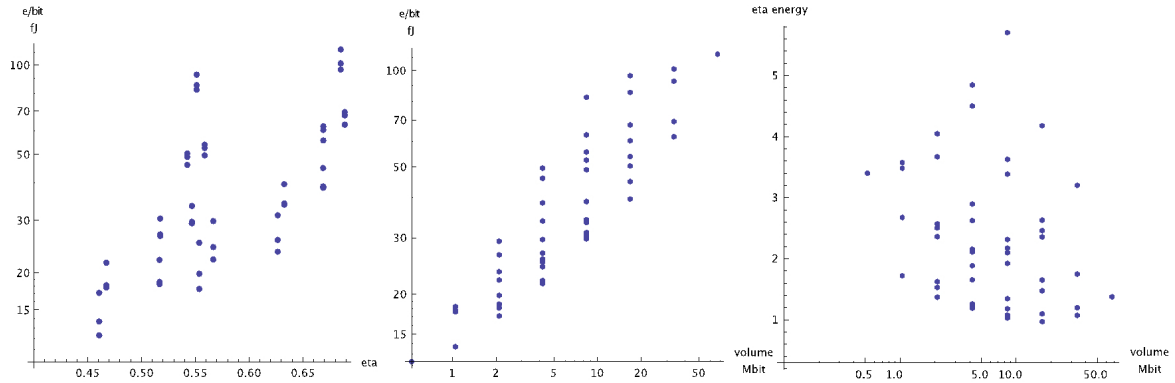

Figure 10: The energy consumption as function of the area utilization, size and energy efficiency for all investigated memories in the design space. From left to right: (a) Energy as function of area utilization, (b) Energy as function of memory size, and (c) Energy as function of the reference energy *eref*.

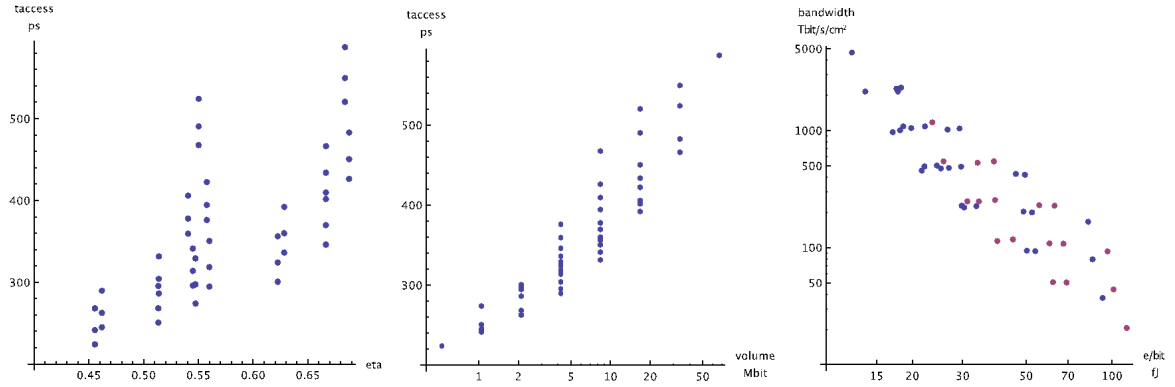

Figure 11: The *read* access time as function of the area utilization and size for all investigated memories in the design space, and the bandwidth as function of the energy consumption. From left to right: (a) Access time as function of area utilization, (b) Access time as function of memory size, and (c) Bandwidth as function of the energy per bit, with an assumed bit-line drain capacitance of 100 aF/transistor.

Table 1: Characteristics for Optimal 32-bit Memories

Utilization	Energy	Access Time	Size	Design Space Variant
62.6 %	25.7 fJ	515 ns	4.2 Mb	{128, 64, 8, 16}
63.2 %	34.2 fJ	569 ns	8.4 Mb	{128, 64, 16, 16}
66.8 %	38.9 fJ	576 ns	8.4 Mb	{128, 128, 8, 16}

8.2 Structure and Fabrication

The Front End of Line (FEOL) structure is based on a bottom layer with a trench capacitor, a buried conductor layer, a transistor layer for the access transistor or n- and p-transistors and word-line wires, and two metal layers, see Figure 13. These are the main steps:

1. on a substrate (1) make a trench capacitor with a planar surface (2–4),
2. the nonmetallic layers (5–8) are built by a damascene process,
3. etching a cavities that form pillars, which may be transistors or vias,
4. deposite a gate insulator over the entire structure (9),
5. make the word-lines within the cavities (10–11),
6. make the metal wire layers (12–13) by a damascene process.

The steps 1, 2, 4, and 5 use Atomic Layer Deposition (ALD) technique to achieve accurate thicknesses [19]. This method allows some offset between the layers.

8.2.1 Capacitors. The conical capacitor has a trench with a diameter of 20 nm and depth 91 nm. It has a conical form where the wall is leaning 1° and where the smallest inner diameter is 9 nm

Table 2: Energy Consumption for Various Parts in Design Space Variant {128, 64, 8, 16}

Node	Energy	Energy Part
SenseCtrl	195.8 fJ	23.8 %
ADCtrl	191.7 fJ	23.3 %
D	142.9 fJ	17.4 %
BL	122.8 fJ	14.9 %
AD	80.9 fJ	9.8 %
Sense	28.9 fJ	3.5 %
C	15.3 fJ	1.9 %
S	11.5 fJ	1.4 %
WL	9.6 fJ	1.2 %
	799.4 fJ	97.2 %

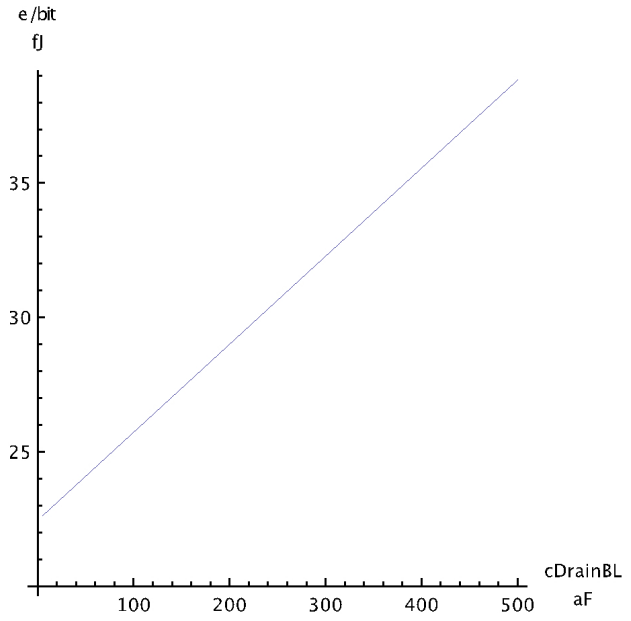

Figure 12: The energy consumption as function of the drain diffusion capacitance on the bit-line for design space variant {128, 64, 8, 16}.

[20]. The trench is covered by a capacitor consisting of ruthenium, Pt/(Al-doped)TiO_2/RuO_2 having a **K** of 100 [17]. The Ru-layer is used as a lattice interface between the surrounding silicon and the high-**K** dielectrics. A central platinum plug is used as lattice interface and low resistance interconnect. The capacitance is 500 aF.

8.2.2 Transistors. The transistor is too large to be placed in a single layer, and every second word-line is therefore implemented in a separate layer. It is conical and consists of a plug with the bottom as a source and the top as a drain surrounded by the gate insulator of silicon nitride HfO_2. The channel width has the circumference 60 nm and length 15 nm. The gate electrode is combined with the word-line conductor. Except for the length, the channel resembles a 14 nm FinFET structure [21], however, its current is only 55 % and the gate capacitance is 6 aF. The drain capacitance of Gate-All-Around (GAA) transistors is generally very low, probably around 14 aF.

8.2.3 Vias. The vias use the same material as the transistor drain.

8.2.4 Conductors. The smallest wire pitch is 12 nm [22]. The word-line strap and bit-line conductors are implemented in tungsten (width 12 nm, pitch 25 nm, 330 Ω/μm and capacitance 1.4 aF per memory cell) [23] and the word-lines in polysilicon (width 12 nm, pitch 25 nm, 11 Ω/μm and capacitance 10 aF per memory cell). Low capacitance Back End of Line (BEOL) wire is in copper (width 24 nm, pitch 48 nm, 46 Ω/μm and capacitance 150 aF/μm [24].

8.3 Processing-in-Memory

Very early, the processors got faster than the memories. The memory interface introduced the cache memory on the same chip to reduce the impact on the von Neumann bottleneck. Because of size, the memory could not be housed on the same chip. With the improvements described in this article, the memory bandwidth exceeds a processor access frequency by orders of magnitude. It is finally time to introduce many small RISC processors within the memory.

In the early introduction of cache memory, the memory was implemented in the same technology as the processor. The state-of-the-art dynamic memories of today have their own technology in which the processor then also has to be implemented, and a such processor can be built. Common for these applications is a VLSI technology with a simple layout and few layers: two layers for Vdd and Gnd, 3–4 layers are used for interconnect where one layer has low capacitance wires.

Dedicated layers are used for transistors and bit-cells, *e.g.* buried wires, buried polysilicon wires, and special material wires.

Conventional processors have been implemented by standard cells using many layers for interconnect. This technique cannot be used. Instead, simple cells that are not based on boolean algebra, *e.g.* instruction fetch and register stack, have to be used. There is a tight relation between layout, circuitry, and architecture. The Surface Based Processor, SBP, is such a design [3].

8.4 Application to a Surface Based Processor

The SBP has a rectangular floor-plan. Surrounding the periphery there are streets for the NoC and clock using 42 wires that are shared with the neighbors. There are an instruction fetch, registers forming a stack, and an 8 μm high arithmetic unit abutted to the D-bus. A control unit occupies 8 μm width.

Using the optimal memory in Figure 13 with an area usage of 46 %, the PE size becomes 154×162 μm^2, energy for the memory read 823 fJ, access time 516 ps and bandwidth 260 Tbit/s/cm^2. The memory size is 4.2 Mbit or 131,072 words. There are 4,043 PEs per cm^2, with 67 % of the area being memory.

The optimal SBP technology at the end of Moore's law has a memory density of 51 %, size 72×144 μm^2, energy 386 fJ, access time 259 ps, and speed 1,200 Tbit/s/cm^2. It has 32 bit-lines with 4-way interleave, 2×64 words, 16 sectors, and 16 memory bank pairs. There are 9,660 PEs per cm^2, with 77 % of the area being memory and density 68 Gbit/cm^2. The maximum power density is 17 W/cm^2. This is not taking into account the further possible improvement of 2.5/3D structures.

9 RELATED WORK

Modern applications need large memories. Thus, energy consumption and high packing density is paramount. Speed is also important.

The trend is a transition from DDRx to GDDRx memories. The control model of these memories relies on an evolutionary slightly altering protocol where internal details are visible, *e.g.* rows, columns, and banks [6, 10]. Another new model based on basic static logic random access memory would influence the industry. The memory architecture described here is a such. Its advantage needs to be high to be accepted.

The von Neumann bottle neck exists in all these applications. It relies on the DRAM interface [11]. Bandwidth is created by the use of many pins or high speed. Pin-count exceeds 4k, *e.g.* JEDEC MO-316B. Speed has reached 18 GHz/pin [25].

To reach this speed microwave wire layout has to be used. Wires are terminated in traditional characteristic impedances, e.g. 48 Ω and 200 mV amplitude [26]. The energy consumption for the interface only is around 0.25 pJ/bit, to be compared with the complete memory consumption of about 25 fJ/bit for the memories described here. The contemporary size is 16 Gbit with 16 channels using a mean area of 2,890 μm^2. The bandwidth is 288 Gbit/s [25] or 580 Gbit/s/cm^2 compared to 260 Tbit/s/cm^2.

The high bandwidth memory HBM uses stacked memory on a substrate increasing the pin-count at moderate speed 1 GHz/pin. The capacitances are reduced. The width of the memory could be increased to 1,024 bits. By stacking this memory by 4, 8 or 12 chips the width could be further increased. The HBM3 consists of 4 TSV connected chips of 16 Gbit. It is 4,096 bits wide with the speed of 5.2 Gbit/s/pin and totally 665 GB/s.

These memories are interconnected by thermo-compression bonding (TCB), using solder capped micro-bumps. It has high thermal resistance and is costly. The next step has been Direct Bond Interconnect (DBI), using a wafer as substrate and stacking up to 16 dies over the entire area [27]. The interconnect pad pitch is 5–20 μm at production speed. With reduced speed, pitch is <1 μm [27].

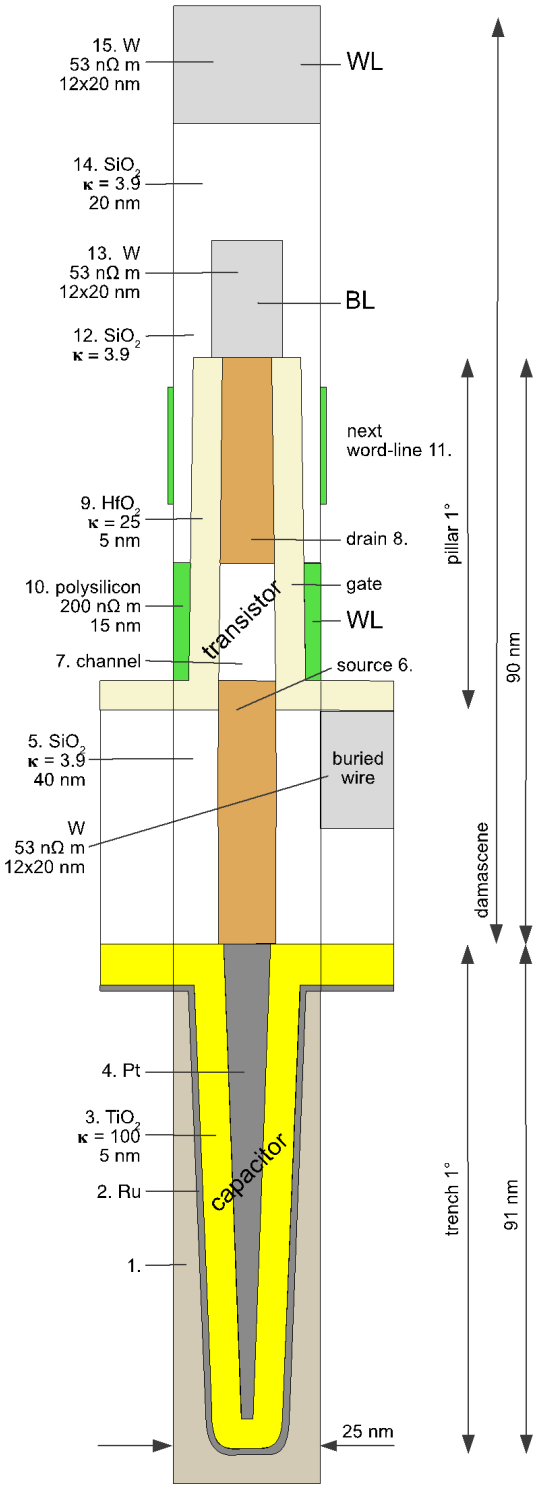

Figure 13: A predicted future 1T1C memory cell.

Subsequently, this has been improved by wafer-to-wafer bonding (W2W) [28].

The direct bonding interconnect uses sub-nm surface roughness causing the oxide surfaces to stick together without any additional material. Cu metal in recesses are annealed to stick together and pads are hermetically sealed. The small and controlled size causes good electrical and thermal properties [29]. Very large memories could be built at moderate cost. Yet, the von Neumann bottleneck prevail for all these technologies. The low pad capacitance 40–100 fF provides a low energy consumption [29], however, one or two magnitudes higher than the memories described in this work.

10 CONCLUSION

Climate change predicts a global warming. However, the current trend in computing is an increasing use of power and energy. Emerging large data centers and AI on consumer level will further increase the energy consumption. These applications are large multiprocessor systems where storage is critical. The world has to reduce the energy use by an order of magnitude; however, this is not yet feasible. New types of low power multiprocessors have to emerge.

Speed and energy are equally important. The trend of CISC processors has been to add more transistors to increase speed and rely on hardware evolution to compensate for the energy requirements. Some other processors guide an opposite trend by reducing complexity, cost and energy consumption. This work is one step towards such processing.

A PE uses energy for arithmetic and memory accesses. In both of them, long wires are the primary energy consumers. The physical size of emerging PEs has to be reduced to minimize energy consumption, especially in the memory. Allocation of processes is made to these small elements. Their working set should be allocated to minimize the mean path length to neighbor elements. The memory size should be minimized, and area utilization should be high.

1T1C 4F^2 memories can be designed for very high performance regarding access time and energy efficiency, with a high area utilization. The peak power density for current and future memories is very low.

ACKNOWLEDGMENTS

The authors wish to thank Prof. Per Stenström at Chalmers University of Technology for his support.

REFERENCES

[1] C. -H. Lin *et al.*, "High performance 14nm SOI FinFET CMOS technology with 0.0174μm^2 embedded DRAM and 15 levels of Cu metallization," *2014 IEEE International Electron Devices Meeting,* 2014, pp. 3.8.1–3.8.3, doi: 10.1109/IEDM.2014.7046977.

[2] E. J. Fluhr *et al.*, "5.1 POWER8TM: A 12-core server-class processor in 22nm SOI with 7.6Tb/s off-chip bandwidth," *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC),* 2014, pp. 96–97, doi: 10.1109/ISSCC.2014.6757353.

[3] M. Rimborg, P. Trancoso and G. Carlstedt, "PHOENIX: Efficient Computation in Memory," *Proceedings of the International Symposium on Memory Systems,* Alexandria, VA, October 2017, pp. 15–25, doi: 10.1145/3132402.3132430.

[4] P. Rosenfeld, E. Cooper-Balis and B. Jacob, "DRAMSim2: A Cycle Accurate Memory System Simulator," in *IEEE Computer Architecture Letters,* vol. 10, no. 1, pp. 16–19, Jan.–June 2011, doi: 10.1109/L-CA.2011.4.

[5] C. Niladrish *et al.*, "USIMM: the Utah SImulated Memory Module," UUCS-12-002, University of Utah, 2012.

[6] Y. Kim, W. Yang and O. Mutlu, "Ramulator: A Fast and Extensible DRAM Simulator," in *IEEE Computer Architecture Letters,* vol. 15, no. 1, pp. 45–49, 1 Jan.–June 2016, doi: 10.1109/LCA.2015.2414456.

[7] H. -C. Shih *et al.*, "DArT: A Component-Based DRAM Area, Power, and Timing Modeling Tool," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 33, no. 9, pp. 1356–1369, Sept. 2014, doi: 10.1109/TCAD.2014.2323203.

[8] J. B. Park, W. R. Davis and P. D. Franzon, "3-D-DATE: A Circuit-Level Three-Dimensional DRAM Area, Timing, and Energy Model," in *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 66, no. 2, pp. 756–768, Feb. 2019, doi: 10.1109/TCSI.2018.2868901.

[9] K. -W. Song *et al.*, "A 31 ns Random Cycle VCAT-Based 4F^2 DRAM With Manufacturability and Enhanced Cell Efficiency," in *IEEE Journal of Solid-State Circuits,* vol. 45, no. 4, pp. 880–888, April 2010, doi: 10.1109/JSSC.2010.2040229.

[10] A. Spessot and H. Oh, "1T-1C Dynamic Random Access Memory Status, Challenges, and Prospects," in *IEEE Transactions on Electron Devices,* vol. 67, no. 4, pp. 1382–1393, April 2020, doi: 10.1109/TED.2020.2963911.

[11] S. Muzaffer A., "Dynamic RAM : Technology Advancement," Boca Raton: CRC Press, 2013, ISBN 978-1-4398-9373-9.

[12] H. Sunami *et al.*, "Scaling Considerations and Dielectric Breakdown Improvement of a Corrugated Capacitor Cell for a Future dRAM," in *IEEE Journal of Solid-State Circuits,* vol. 20, no. 1, pp. 216–223, Feb. 1985, doi: 10.1109/JSSC.1985.1052296.

[13] K. Kim and J. Lee, "A New Investigation of Data Retention Time in Truly Nanoscaled DRAMs," in *IEEE Electron Device Letters,* vol. 30, no. 8, pp. 846–848, Aug. 2009, doi: 10.1109/LED.2009.2023248.

[14] S. J. E. Wilton and N. P. Jouppi, "CACTI: an enhanced cache access and cycle time model," in *IEEE Journal of Solid-State Circuits,* vol. 31, no. 5, pp. 677–688, May 1996, doi: 10.1109/4.509850.

[15] K. Fischer *et al.*, "Low-k interconnect stack with multi-layer air gap and tri-metal-insulator-metal capacitors for 14nm high volume manufacturing," *2015 IEEE International Interconnect Technology Conference and 2015 IEEE Materials for Advanced Metallization Conference (IITC/MAM),* 2015, pp. 5–8, doi: 10.1109/IITC-MAM.2015.7325600.

[16] S. Ghose *et al.*, "What Your DRAM Power Models Are Not Telling You: Lessons from a Detailed Experimental Study," *Proceedings of the ACM on Measurement and Analysis of Computing Systems,* vol. 2, no. 3, pp. 1–41, 2018, doi: 10.1145/3224419.

[17] J. H. Han *et al*, "Improvement in the leakage current characteristic of metal-insulator-metal capacitor by adopting RuO_2 film as bottom electrode," in *Applied Physics Letters,* ISSN 0003-6951, vol. 99, no. 2, pp. 1–3, 2011, doi: 10.1063/1.3609875.

[18] F. Gamiz, J. A. Lopez-Villanueva, J. B. Roldan, J. E. Carceller and P. Cartujo, "Monte Carlo simulation of electron transport properties in extremely thin SOI MOSFET's," in *IEEE Transactions on Electron Devices,* vol. 45, no. 5, pp. 1122–1126, May 1998, doi: 10.1109/16.669557.

[19] J. Niinistö, K. Kukli, M. Heikkilä, M. Ritala and M. Leskelä, "Atomic layer deposition of high-k oxides of the group 4 metals for memory applications," *Advanced Engineering Materials,* vol. 11, no. 4, pp. 223–234, 2009, doi: 10.1002/adem.200800316.

[20] V. Cremers, R. L. Puurunen and Jolien Dendooven, "Conformality in atomic layer deposition: Current status overview of analysis and modelling," in *Applied Physics Reviews,* vol. 6, no. 2, 2019, doi: 10.1063/1.5060967.

[21] S. Natarajan *et al.*, "A 14nm logic technology featuring 2nd-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588 μm^2 SRAM cell size," *2014 IEEE International Electron Devices Meeting,* 2014, pp. 3.7.1–3.7.3, doi: 10.1109/IEDM.2014.7046976.

[22] M. Neisser and S. Wurm, "ITRS lithography roadmap: status and challenges," *Advanced Optical Technologies,* vol. 1, no. 4, 2012, pp. 217–222, doi: 10.1515/aot-2012-0045.

[23] D. Gall, "Metals for Low-Resistivity Interconnects," *2018 IEEE International Interconnect Technology Conference (IITC),* 2018, pp. 157–159, doi: 10.1109/IITC.2018.8456810.

[24] C. Penny *et al.*, "Reliable airgap BEOL technology in advanced 48 nm pitch copper/ULK interconnects for substantial power and performance benefits," *2017 IEEE International Interconnect Technology Conference (IITC),* 2017, pp. 1–4, doi: 10.1109/IITC-AMC.2017.7968970.

[25] Y. -J. Kim *et al.*, "A 16-Gb, 18-Gb/s/pin GDDR6 DRAM With Per-Bit Trainable Single-Ended DFE and PLL-Less Clocking," in *IEEE Journal of Solid-State Circuits,* vol. 54, no. 1, pp. 197–209, Jan. 2019, doi: 10.1109/JSSC.2018.2883395.

[26] T. M. Hollis *et al.*, "Recent Evolution in the DRAM Interface: Mile-Markers Along Memory Lane," in *IEEE Solid-State Circuits Magazine,* vol. 11, no. 2, pp. 14–30, Spring 2019, doi: 10.1109/MSSC.2019.2910617.

[27] G. Gao *et al.*, "Scaling Package Interconnects Below 20μm Pitch with Hybrid Bonding," *2018 IEEE 68th Electronic Components and Technology Conference (ECTC),* 2018, pp. 314–322, doi: 10.1109/ECTC.2018.00055.

[28] Z. Wan, K. Winstel, A. Kumar and S. S. Iyer, "Low-Temperature Wafer Bonding for Three-Dimensional Wafer-Scale Integration," *2018 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S),* 2018, pp. 1–2, doi: 10.1109/S3S.2018.8640155.

[29] A. Agrawal, S. Huang, G. Gao, L. Wang, J. DeLaCruz and L. Mirkarimi, "Thermal and Electrical Performance of Direct Bond Interconnect Technology for 2.5D and 3D Integrated Circuits," *2017 IEEE 67th Electronic Components and Technology Conference (ECTC),* 2017, pp. 989–998, doi: 10.1109/ECTC.2017.341.