A novel DRAM architecture as a low leakage alternative for SRAM caches in a 3D interconnect context.

Anselme Vignon, Stefan Cosemans, Wim Dehaene
K.U. Leuven
ESAT - MICAS Laboratory
Kasteelpark Arenberg 10, Leuven, Belgium
anselme.vignon@esat.kuleuven.be

Pol Marchal, Marco Facchini
IMEC
Kapeldreef 75, B-3001 Leuven, Belgium
pol.marchal@imec.be

***Abstract*—This paper presents a DRAM architecture that improves the DRAM performance/power trade-off to increase their usability on low power chip design using 3D interconnect technology. The use of a finer matrix subdivision and buffering the bitline signal at the localblock level allows to reduce both the energy per access and the access time. The obtained performances match those of a typical low power SRAM, while achieving a significant area and static power reduction compared to these memories.**

The 128 kb memory architecture proposed here achieves an access time of 1.3 ns for a dynamic energy of less than 0.2 pJ per bit. A localized refresh mechanism allows gaining a factor of 10 in static power consumption associated with the cell, and a factor of 2 in area, when compared with an equivalent SRAM.

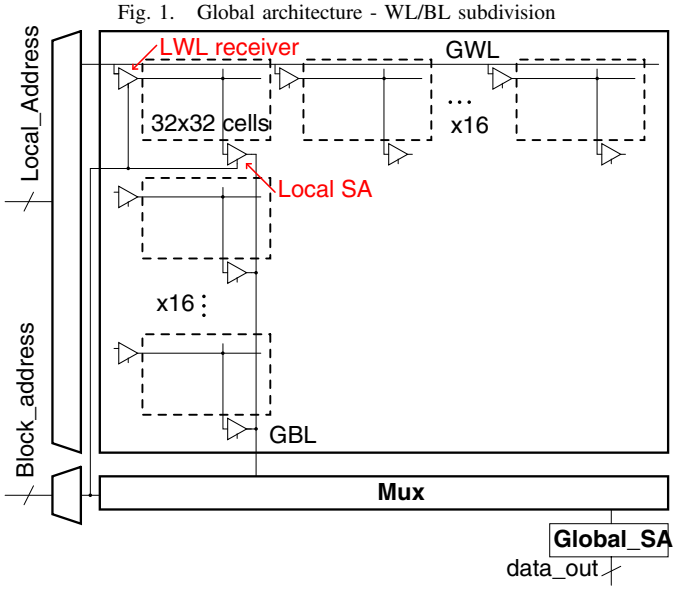
I. Context

As feature size reduces, on-chip memory design is becoming more and more challenging. Reducing the typical dimensions and the supply voltage for SRAM memories degrades the cell stability [1]. The stability is degraded further by intra-die variations which lead in addition to increased average power consumption. Several solutions have been investigated to reduce this issue, from changing the cell topology [2] [3] [4] to modifying the peripheral architecture [5]. However, these solutions increase the memory area and thus compromise scaling. Embedded DRAM (eDRAM) has been proposed for large memory arrays. eDRAM clock speed and access time have been improved to match the SRAM typical behavior [6]. However, using eDRAM requires to integrate more dense capacitors in the logic technology process, and thus needs costly additional process steps.

3D interconnect enables the use of heterogeneous technologies on the same chip. 3D vias are typically smaller and have less parasitic capacitance than off-chip connections [7]. In addition, they can be spread across the chip. This reduces the routing energy, and increases the number of available connections between two stacked dies.

These advantages allow to provide a better bandwidth-energy trade off for the routing between two stacked dies than between two packaged dies. A possible application of 3D interconnect is to separate the logic core of a system from the memory it requires. Such systems have already been studied in [8] [9], with stacks of an SRAM matrix on top of a logic layer. It is also possible to stack DRAM on top of a logic layer.

Fig. 1. Global architecture - WL/BL subdivision



This solution offers numerous other advantages compared to packaged DRAM, including simpler inputs/outputs protocol, and can solve the terminations and clock synchronisation issues by using shorter connections. This allows using conventional DRAM instead of SRAM or embedded DRAM for the largest memories in SOC, bringing a higher density compared to SRAM, without the need to integrate dedicated capacitors in the logic process, as for eDRAM.

However, traditional DRAM is outperformed by SRAM in several domains. The typical access time of a DRAM is still higher than for an SRAM, and the access energy per bit is higher. This makes conventional DRAM not suited for high activity caches, where dynamic access energy consumption and delay are critical.

978-3-9810801-5-5/DATE09 © 2009 EDAA

Fig. 2. Proposed system architecture

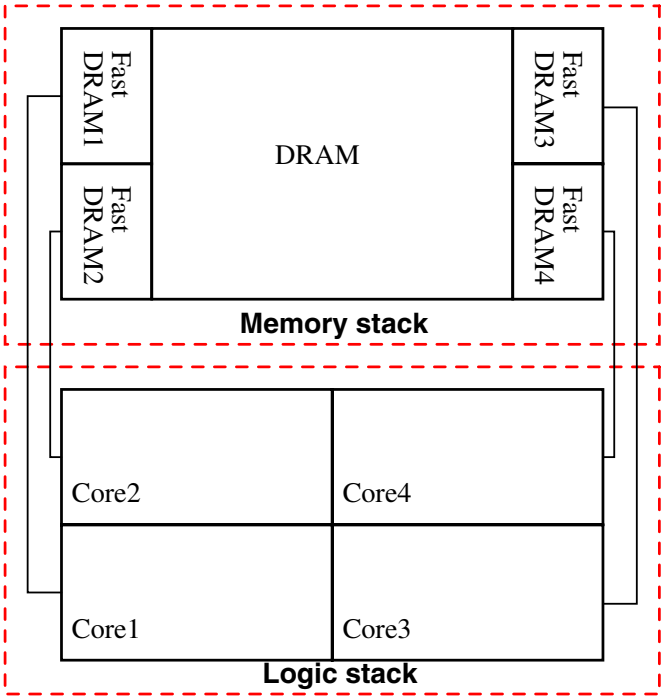


Fig. 3. read and refresh operations

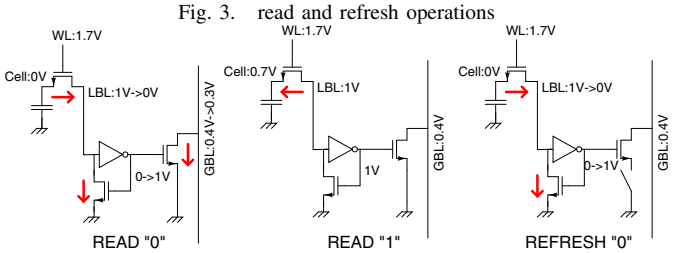
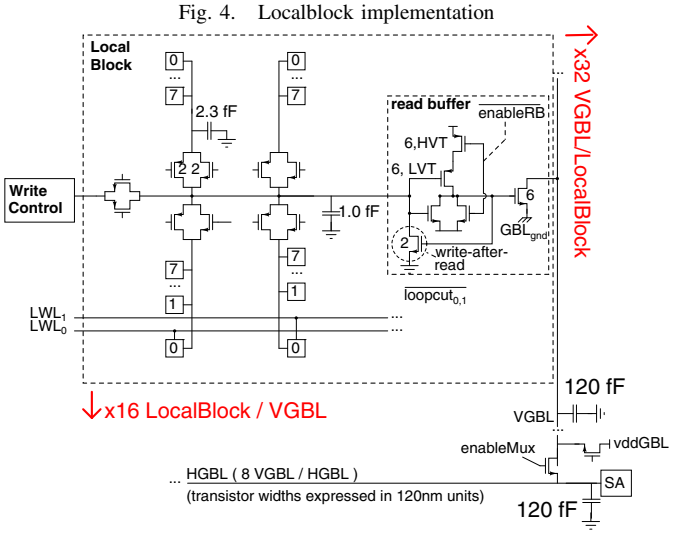


Fig. 4. Localblock implementation



The proposed DRAM architecture provides faster access time as well as lower energy consumption compared to conventional DRAM, while still being more dense than SRAM. This can be achieved by using techniques that are being developed for SRAM [5], i.e. a finer granularity matrix subdivision, as shown in figure 1. In addition, adding a local sense amplifier inside the memory matrix allows to implement the write after read operation at local level. These techniques save on energy and latency, and allows a more efficient refresh scheme, described in the next section.

A 3D hybrid architecture can be used to build the entire cache system on the same memory die, in the context of 3D memory architecture. As shown in figure 2, faster DRAM is used as first cache level, with regular DRAM as second cache level.

II. Fast DRAM circuit topology

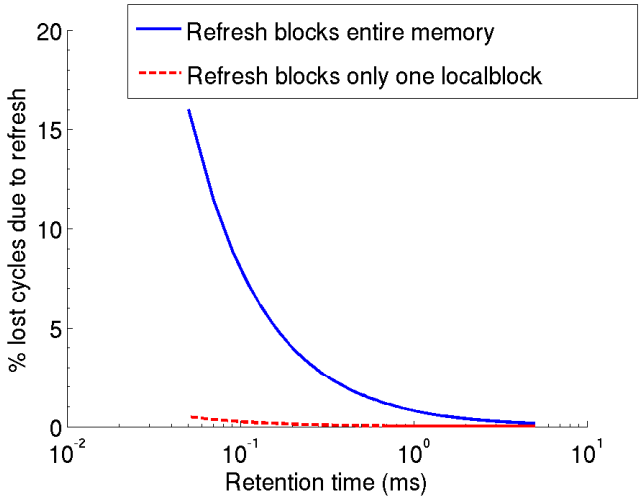
The studied design is a 128 kb memory array with 32 bits word length. It is based on the SRAM memory described in [10], which focuses on reducing the active energy per bit, while operating at a reasonable speed. Three main improvements are introduced, a high granularity memory organisation, tunable sense amplifiers, and tunable delay lines to cope with variability with acceptable delay and energy penalty. These improvements are integrated in the design described here, with some modification at the localblock level and for the global peripherals, in order to deal with DRAM specific issues. This stackable DRAM was designed in a 90*nm* technology.

One of the reasons for the longer access time of a DRAM matrix compared to an SRAM, is the smaller voltage drop that the cell can develop on the bitline. This voltage drop is limited by the ratio between the DRAM cell capacitance and the bitline capacitance. In order to tackle this limitation, very short local bitlines, of an equivalent capacitance of 16 cells per bitline, are used. The voltage drop is first sensed by a local sense amplifier, which will drive the global bitline (GBL) signal. One local wordline (WL) access only one word, as shown in figure 1. This decreases the access energy consumption, as only the accessed cells get read, thus saving the activation cost of opening additional words.

This architecture with enhanced locality, shown in figure 3 brings another advantage for a DRAM matrix. The local sense amplifier restores the data locally after reading. This reduces the latency of a reading operation, as the write after read operation is performed while the *GBL* signal is sensed by the global sense amplifier (SA).

A critical aspect of power consumption and speed for a DRAM based memory is the refresh handling. Compared to the rest of the memory die, the static power associated with the cell often dominates static power consumption. For an SRAM cell, this power is the static cell leakage. The equivalent power for the DRAM cell is the energy associated with refreshing this cell, multiplied by the frequency at which this operation must be repeated. This is because the static leakage of an SRAM is directly consumed, while the leakage of a DRAM cell consumes energy only when the cell is restored. The typical current leakage of a DRAM cell is smaller than for an SRAM, but the total static power consumed is larger than this leakage, due to this refresh operation. In terms of speed, when

Fig. 5. percentage of busy cycles due to refresh, for a monoblock and a 128 localblocks DRAM, running at $500MHz$.



a concurrent access is impossible during the refresh operation, a latency penalty occurs, due to the cycle lost during refresh.

The refresh circuit uses the circuit described in figure 4 to minimize its impact on speed and on energy. During refresh, the data read from the cell is written back into the cell at local level. The GBL_{gnd} node is left floating during this operation, therefore saving the energy of restoring the GBL afterwhile. As a consequence, the energy cost of a refresh cycle is reduced, as neither the global sensing circuit nor the global write circuits are used during the operation. The implementation of this architecture is shown in figure 4

Refresh operations as described above affect the memory only at localblock level thus can be done in parallel with accessing another localblock. This reduces the average latency penalty due to refresh, as shown in figure 5.

The percentage of idle cycles is calculated in two cases, when a refresh operation is performed at memory level, without concurrent access, and when refreshing a localblock keeps only this localblock inaccessible. The refresh timing penalty is negligible in the context of a refresh taking advantage of the memory granularity, especially for high retention time.

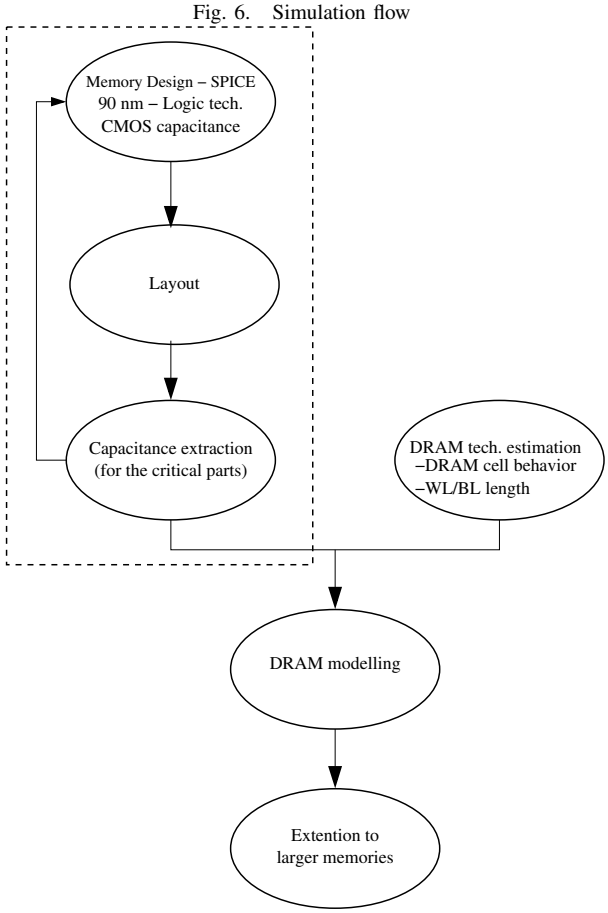
The main advantage of the topology described above, is to decrease the active power and the access time of a DRAM based memory. In addition, using such topology allows also to implement a low power refresh operation, concurrent with write and read access on different localblocks. This reduces the latency penalty due to this additional operation, compared to an SRAM.

III. Methodology

To evaluate the potential of the proposed design considering DRAM technology, a three steps approach was used.

- A scratchpad test memory has been designed using CMOS capacitances for the memory cell.
- Based on the post layout simulation result of this design, the performances of such architecture were estimated when using DRAM technology.

Fig. 6. Simulation flow



- The evolution of the speed, area, static and dynamic power figures were estimated for larger memories.

Voltages used in the design for the scratchpad test memory are limited to 1.2V. The capacitance based cell is a $11fF$ CMOS gate capacitance, with a high threshold transistor used as access transistor. The worst case retention time in 6σ worst case monte-carlo simulation, was found at $200\mu s$. This time was used to estimate the static power consumption, assuming the entire memory array is refreshed. The retention time is especially important to estimate the static power, as the power consumption associated with the cell is proportional to this value. The retention time is dependant on the technology used to build the access transistor of the cell, and the capacitance of this cell. The value used here is very conservative, as the cell described here doesn't use dedicated access transistors, neither deep trench capacitors.

Several differences between DRAM and logic process were taken into account, to get an accurate view on the power consumption and speed of this architecture in a DRAM technology. The DRAM cell is based on deep trench capacitors, which means a higher capacitance value and a smaller area. DRAM cell access transistor gates are also typically overdriven, which is not possible in a logic process, due to the reliability electrical rules restrictions for logic transistors. To take these difference into account, we assumed the cell capacitance value

Fig. 7. Access time (a), Dynamic (b) and static (c) Energy, and area (d), for DRAM and SRAM

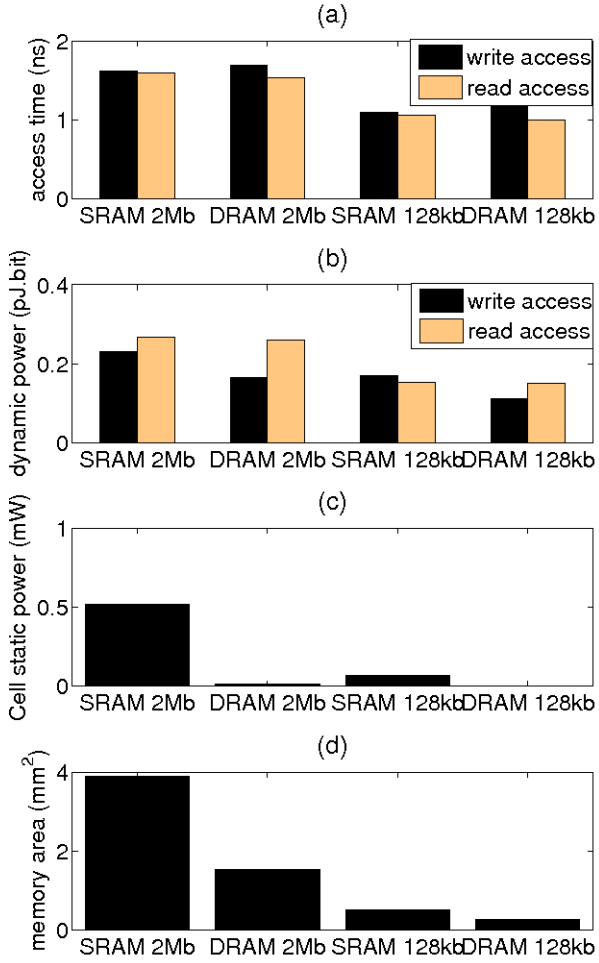


Fig. 8. Energy repartition in fast dram

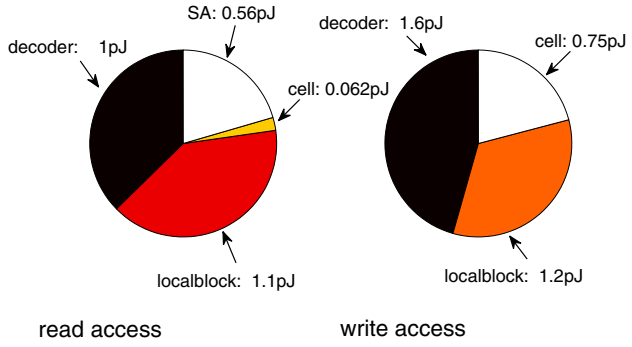
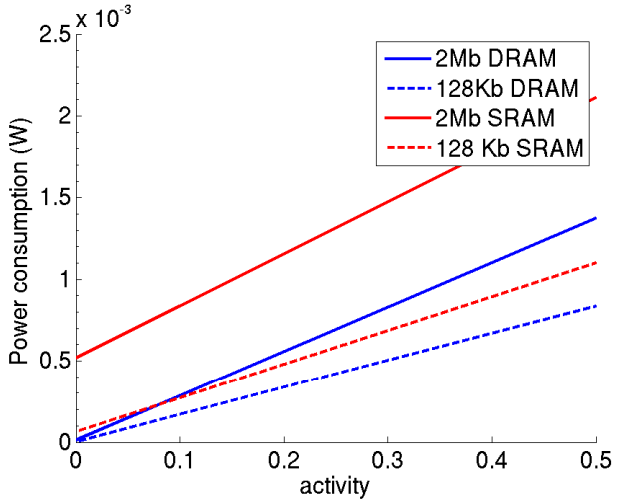


Fig. 9. Total energy consumption, as a function of activity, for different memory sizes



was $30fF$, and the local wordlines were driven at $1.7V$. The modification described above were implemented in circuit simulation.

To evaluate the advantages of using this voltages and capacitance values for the area figure, the number of cells per LBL was increased, in order to get similar timing compared to the original CMOS capacitance based cell. It was found that, using overdrive on the local wordline, it is possible to double this number of cells, from 16 to 32 cells per bitline. The total area was corrected accordingly. A typical area for a $90nm$ DRAM cell is $0.30\mu m^2$. We used this value to estimate the total area obtained by reducing the area of the cell while keeping the peripheral area constant. Finally, the active power consumption consumed on charging the global wordlines and bitlines was refined, according to their size reduction.

The increase in access time as well as in energy consumption for a larger memory array of 2Mb was estimated, using GBL/GWL larger capacitance estimation, with a timing penalty due to larger buffers needed on this signal.

IV. Results

All the results are compared here against the equivalent SRAM matrix, presented in [10]. As shown in figure 7(a), the impact of using this DRAM topology in term of access time is negligible, especially for medium size ($2Mb$) memories. This shows that this architecture is suitable for building faster access cache memories based on DRAM technology.

Mainly three factors impact the active power difference between the two matrices. On one hand, the DRAM active power increases due to a higher WL voltage. More power is also consumed on the local sense amplifiers during read, as the voltage drop is lower for DRAM as for SRAM. On the other hand, the higher density for DRAM leads to lower GWL/GBL energy consumption at constant memory size, due to the shorter signal lines. This leads to a similar read active power for the two matrices, and a significant improvement for the write energy of a large matrix, as shown in figure 7(b). Due to the refresh handling scheme, the cell static power consumption is 10 times less for DRAM than for the SRAM memory, for a 2Mb memory (figure 7(c)). The cell static power consumption is given as the static leakage for the SRAM, compared to the

power consumed by the refresh operation, when all the cells in the matrix are being refreshed.

Figure 9 shows an overall power consumption improvement, especially for large arrays with low activity, for a random access pattern with as much read as write accesses.

TABLE I
MEMORY AREA ESTIMATION FOR SRAM AND PROPOSED DRAM

Size	SRAM (mm^2)	Fast DRAM (mm^2)
128 *kb*	0.5	0.276
2 *Mb*	3.9	1.520

As we can see in figure 8, doubling the number of cells per LBL has a marginal impact on the power consumption, as most of the localblock power consumption is due to the local sense amplifiers. The total area is reduced by a factor of 2.7. This gain could be improved, as the peripheral circuits used here were originally designed for an SRAM. Further gain should be possible by designing peripherals dedicated to a DRAM matrix.

V. CONCLUSION

In this paper, we showed that DRAM memory is an attractive candidate to replace SRAM using 3D interconnects. The proposed architecture allows to replace low memory hierarchy SRAM, and outperforms typical SRAM in density and passive power, while matching active power and speed figures.

The use of 3D interconnects allows to reduce the routing energy and speed penalty between two dies. In this context, typical DRAM matrix presents a speed/energy/area trade off, that makes it attractive for high level memory caches, where similar speed and passive power can be obtained for a lower area per cell. A DRAM architecture was proposed, that brings these advantages to low level memory caches, while consuming less static power. This novel architecture shows that DRAM memory is suitable to build faster memory, at the cost of modifying the memory matrix architecture. Using a finer granularity matrix reduces both the access time and the dynamic power consumption. It also allows to use a new implementation of the refresh operation, that reduces its impact on access delay and passive energy consumption. The proposed architecture allows to replace low level SRAM memories, and offers a better speed versus static power trade off than SRAM, while being more dense. It outperforms typical SRAM in density, by a factor of 2.7, and by a factor of 10 in passive power. The active power and speed figures are similar for both DRAM and SRAM architectures. This makes such DRAM an attractive candidate to replace SRAM based caches, when using 3D interconnects.

REFERENCES

[1] L. Chang *et al.*, "Stable sram cell design for the 32 nm node and beyond," *VLSI Technology, 2005. Digest of Technical Papers. 2005 Symposium on*, pp. 128–129, 14-16 June 2005.

[2] S. Jain and P. Agarwal, "A low leakage and snm free sram cell design in deep sub micron cmos technology," *VLSI Design, 2006. Held jointly with 5th International Conference on Embedded Systems and Design., 19th International Conference on*, p. 4, 3-7 Jan. 2006.

[3] N. Verma and A. Chandrakasan, "A 65nm 8t sub-vt sram employing sense-amplifier redundancy," *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, pp. 328–606, 11-15 Feb. 2007.

[4] Noguchi *et al.*, "A 10t non-precharge two-port sram for 74% power reduction in video processing," *VLSI, 2007. ISVLSI '07. IEEE Computer Society Annual Symposium on*, pp. 107–112, 9-11 March 2007.

[5] S. Cosemans, W. Dehaene, and F. Catthoor, "A low power embedded sram for wireless applications," *Solid-State Circuits Conference, 2006. ESSCIRC 2006. Proceedings of the 32nd European*, pp. 291–294, Sept. 2006.

[6] J. Barth *et al.*, "A 500mhz random cycle 1.5ns-latency, soi embedded dram macro featuring a 3t micro sense amplifier," *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, pp. 486–617, 11-15 Feb. 2007.

[7] M. Kawano, "A 3d packaging technology for high-density stacked dram," *VLSI Technology, Systems and Applications, 2007. VLSI-TSA 2007. International Symposium on*, pp. 1–2, 23-25 April 2007.

[8] Y.-F. Tsai, F. Wang, Y. Xie, N. Vijaykrishnan, and M. J. Irwin, "Design space exploration for 3-d cache," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 16, no. 4, pp. 444–455, April 2008.

[9] K. Puttaswamy and G. Loh, "Implementing caches in a 3d technology for high performance processors," *Computer Design: VLSI in Computers and Processors, 2005. ICCD 2005. Proceedings. 2005 IEEE International Conference on*, pp. 525–532, 2-5 Oct. 2005.

[10] S. Cosemans, W. Dehaene, and F. Catthoor, "A 3.6pj/access 480mhz, 128kbit on-chip sram with 850mhz boost mode in 90nm cmos with tunable sense amplifiers to cope with variability," *Solid-State Circuits Conference, 2008. ESSCIRC 2008.*, Sept. 2008.