Design Optimization of Sense Amplifiers using Deeply-scaled FinFET Devices

Alireza Shafaei[1], Yanzhi Wang[1], Antonio Petraglia[2], and Massoud Pedram[1]
[1] Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089
[2] Federal University of Rio de Janeiro, Brazil
shafaeib@usc.edu, yanzhiwa@usc.edu, antonio@pads.ufrj.br, pedram@usc.edu

***Abstract*—This paper presents the design optimization of sense amplifiers made of deeply-scaled (7nm) FinFET devices in order to improve the energy efficiency of cache memories, while robust operation of the sense amplifier under process variations is achieved. To this end, an analytical solution for deriving the minimum voltage difference that can be correctly sensed between the sense amplifier inputs, considering process variations, is presented. Device parameters and transistor sizing of the sense amplifier are then optimized in order to further increase the cache energy efficiency. The optimized sense amplifier design has 2-fold lower input voltage difference compared with the baseline counterpart, which according to the architecture-level simulations, causes 26% reduction in the total energy consumption of an L1 cache memory.**

I. Introduction

Sense amplifiers are commonly used in the read path of cache memories. Basically, the purpose of the sense amplifier circuit is to sense and then amplify a small voltage difference between the two input nodes, BL and $\overline{BL}$, which prevents a full-swing discharge on the aforesaid interconnects, and hence improves the cache access latency and reduces the dynamic power consumption. On the other hand, the robust operation of the sense amplifier mainly depends on this input difference voltage, denoted by ΔV [1], [2]. More precisely, ΔV should be small enough to reduce the energy consumption, but large enough to ensure the robustness of the sense amplifier (i.e., sensing ΔV correctly) under process variations.

Moving towards deeply-scaled technologies, where extremely small geometries, such as transistors with gate lengths below 10nm, are employed and *short channel effects* (SCE) in bulk CMOS devices are increased, the effect of process variations is becoming more severe. However, quasi-planar FinFETs provide a three-dimensional gate control over the channel which effectively reduces the source and drain controls, thereby suppressing SCE [3]. Moreover, because of undoped channels, FinFETs offer higher immunity to random variations and soft errors [4], [5]. As a result, FinFETs are perceived as the choice of underlying device for technologies beyond the 10nm regime [6].

Due to the benefits of FinFET devices, FinFET-based SRAMs have been proposed as a solution for enhancing the stability and energy efficiency of SRAM cells [7], [8]. Accordingly, sense amplifiers equipped with FinFET devices are shown to function with smaller ΔVs compared with planar CMOS counterparts [2], [9]. This paper thus presents the design optimization of FinFET-based sense amplifiers in order to minimize ΔV such that yield constraints of the sense amplifier under process variations are satisfied. Our designs employ 7nm FinFET devices [10], where the device optimization procedure is carried out using advanced simulators from Synopsys [11].

We also adopt an analytical solution to derive the value of ΔV that guarantees the robust operation of the sense amplifier under variations caused by line edge roughness, which is the main source of statistical variabilities in FinFET devices [5]. Increasing the number of fins or transistor gate length are effective solutions for mitigating process variations [12]. Hence, we optimize gate lengths and numbers of fins of FinFET devices in order to further minimize ΔV, and hence increase the cache energy efficiency. The optimized sense amplifier design has 2-fold lower ΔV compared with the baseline counterpart, which according to the architecture-level simulations, causes 26% reduction in the total energy consumption of a 32KB, 4-way set-associative, L1 cache memory.

The rest of the paper is organized as follows. Section II reviews basic operation of sense amplifiers and introduces our 7nm FinFET devices. Section III presents the yield analysis of FinFET-based sense amplifiers. The proposed design optimization is discussed in Section IV, followed by simulation results in Section V. Finally, Section VI concludes the paper.

II. 7T Sense Amplifier

A latch-type sense amplifier made of seven transistors (7T), as shown in Fig. 1, is adopted in this paper. This 7T sense amplifier contains two isolating transistors (M_1 and M_4), two cross-coupled inverters composed of two pull-up (M_2 and M_3) and two pull-down (M_5 and M_6) transistors, and a footer transistor (M_7). When ΔV is established between BL and $\overline{BL}$, *sense enable* (SE) signal is activated, which in turn triggers the positive feedback provided by the cross-coupled inverters in order to rapidly generate the proper outputs. The performance of a sense amplifier is characterized by the *sensing delay*, denoted by D, and defined as the time from the activation of SE until outputs are ready. On the other hand, the robustness is mainly determined by ΔV, which is defined as the minimum voltage difference between BL and $\overline{BL}$ that can be sensed correctly [1]. Hence, ΔV plays an important role in yield calculations of the sense amplifier.

Furthermore, our sense amplifiers are designed using FinFET devices with a gate length of 7nm [10]. FinFET-specific geometries, including the fin height (H_{FIN}), the fin width, also known as the silicon thickness (T_{SI}), and the gate length (L), of the 7nm FinFET process are reported in Table I. Because of the 3D structure of the FinFET gate, the effective channel width of a single fin device is approximately equal to $2 \times H_{FIN}$. In order to increase the width of a FinFET, more fins are added in parallel, where the spacing between two adjacent pins is determined by the fin pitch (P_{FIN}), whose

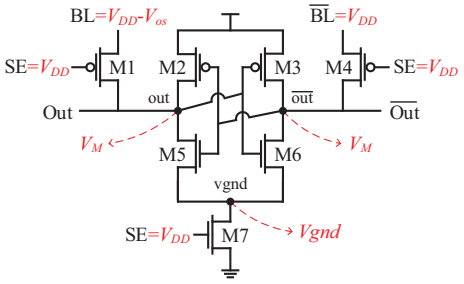

Fig. 1. Circuit structure of the 7T sense amplifier. Red texts show voltage levels when the circuit is in the metastability state.

TABLE I. SPECIFICATIONS OF 7NM FINFET DEVICES [10].

Parameter	Value (nm)	Comment
L	$2\lambda = 7$	Fin or gate length
T_{SI}	3.5	Fin width, also know as silicon thickness
H_{FIN}	14	Fin height
P_{FIN}	$2\lambda + T_{SI} = 10.5$	Fin pitch using spacer-defined lithography
t_{ox}	1.3	Oxide thickness

value is dictated by the underlying FinFET technology. The supply voltage, V_{dd}, of the adopted FinFET devices is 0.45V, and the threshold voltage, V_{th}, is between 0.2V and 0.25V.

III. YIELD ANALYSIS

In this section, sources of process variations in deeply-scaled FinFET technologies are discussed. We then present an analytical solution for deriving ΔV that ensures the robust operation of the sense amplifier.

A. Process Variations in FinFET Devices

The undoped channel of FinFET devices eliminates the *random dopant fluctuation*, making FinFETs less sensitive to process variations compared with planar CMOS counterparts. However, FinFETs suffer from other sources of process variations, particularly under deeply-scaled technologies. The main source is recognized as the *line edge roughness* (LER) [5], which imposes variations on the (effective) channel length, L. The effect of LER on 7nm FinFETs has been studied by measuring the ON and OFF currents of NFET and PFET devices for different values of L by using Synopsys TCAD [11]. Results are illustrated in Fig. 2, which shows that the OFF current is highly sensitive to variations of L, whereas the ON current slightly changes by varying the gate length.

For 14nm FinFET technology, the standard deviation of L is predicted to be 0.8nm [5] [13]. Taking into account scaling trends in FinFET process technology, we assume 0.5nm as the standard deviation of L for 7nm FinFET, which is within the reasonable range. Hence, in this paper, we assume that the gate length has a Gaussian distribution with mean μ_L =7nm, which is the nominal gate length of our FinFET devices, and standard deviation σ_L =0.5nm. Moreover, the gate length of transistor M_i will be denoted by L_i, whereas N_i is used to refer to the number of fins.

B. Deriving ΔV for a Robust Sense Amplifier

Supposing that due to LER, the gate length of M_6 becomes smaller than that of M_5, which essentially increases the current through M_6, then the sense amplifier will be biased to produce

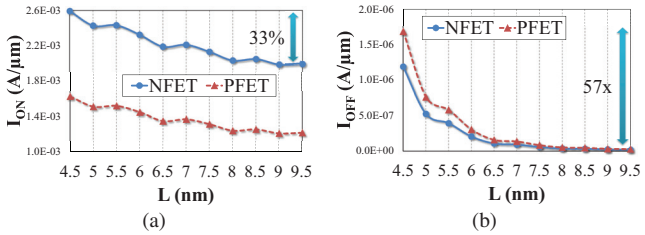

Fig. 2. The effect of LER on 7nm FinFETs: (a) ON and (b) OFF currents as a function of gate length, L.

$V(out) = V_{DD}$ and $V(\overline{out}) = 0$. However, by setting an appropriate ΔV, the effect of process variations can be mitigated. In order to mathematically formulate the problem, the input offset voltage, V_{os}, is defined as the voltage offset between BL and $\overline{BL}$ that leads the sense amplifier to the metastable state, i.e., $V(out) = V(\overline{out}) = V_M$ [1]. Robust operation of the sense amplifier is then achieved by having $\Delta V \geq \mu_{V_{os}} + 3\sigma_{V_{os}}$, where $\mu_{V_{os}}$ and $\sigma_{V_{os}}$ denote the mean and standard deviation of V_{os}, respectively. In other words, as ΔV increases so does the current through M_1, because of increasing V_{gs} of M_1, and subsequently M_5. V_{os} is then the voltage such that $V(out)$ is equal to $V(\overline{out})$, and hence, any $\Delta V > V_{os}$ forces the sense amplifier to generate the correct output. However, due to process variations, we use $\Delta V \geq \mu_{V_{os}} + 3\sigma_{V_{os}}$ to achieve a high yield sense amplifier.

The value of V_{os} is obtained by writing the Kirchhoff's current law equations at out, $\overline{out}$, and $vgnd$ nodes of the sense amplifier (cf. Fig. 1), which will give us V_{os} as a function of gate lengths of sense amplifier transistors. By assuming that the gate lengths are independent and normally distributed random variables and by running Monte Carlo simulations, values of $\mu_{V_{os}}$ and $\sigma_{V_{os}}$ are calculated. However, for analytically solving the resulted equation systems, ON and OFF current equations of FinFET devices are needed, which are modeled as shown next.

C. Modeling FinFET Currents

After 7nm FinFET devices have been designed using the TCAD tool suite, SPICE-compatible Verilog-A models are also extracted to enable fast circuit-level simulations. Using these SPICE models, we measured the V_M value of the sense amplifier using the 7nm FinFET devices, which showed $V_M < V_{th}$. Therefore, all transistors of the sense amplifier, except for M_7, are in the subthreshold mode. Since during the metastable state, V_{ds} of M_1 to M_6 transistors are relatively large (compared with the thermal voltage V_T), and by neglecting the drain voltage dependence coefficient (DIBL coefficient) for FinFET devices, the OFF current is modeled using the following equation:

$$I_{OFF} = \frac{A}{L} \cdot e^{\frac{|V_{gs}| - |V_{th}|}{nV_T}}, \quad (1)$$

where A is a technology-dependent value, L denotes the gate length, n represents the subthreshold slope factor, and V_T is the thermal voltage.

Values of A and n are fitted based on SPICE simulations using the Verilog-A models. Fig. 3 validates the accuracy of the model vs. SPICE simulations. On the other hand, M_7 is turned on and, because of small V_{ds}, lies in the linear region. We therefore use the alpha-power law [14] to model the ON

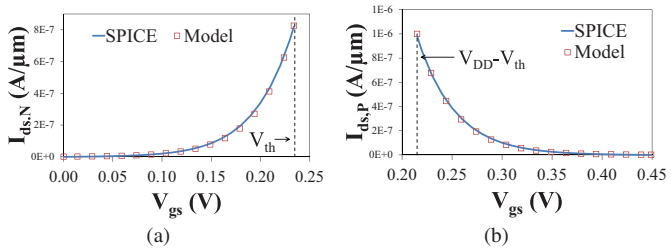

Fig. 3. I_{ds} vs. V_{gs} for 7nm (a) NFET and (b) PFET devices using SPICE simulations and the subthreshold model.

TABLE II. DESIGN VARIABLES.

Transistor	M_1	M_2	M_3	M_4	M_5	M_6	M_7
Gate length	7nm	L_{opt}	L_{opt}	7nm	L_{opt}	L_{opt}	7nm
Number of fins	N_{opt}	1	1	N_{opt}	1	1	1

current of FinFET devices. Based on our curve fitting results, we obtained α=1.3 for our 7nm FinFET devices.

IV. DESIGN OPTIMIZATION

Our objective is to minimize ΔV in order to reduce the cache access latency as well as the dynamic power consumption, and hence improve the energy efficiency. On the other hand, due to the inevitable effect of process variations under deeply-scaled technologies, it is crucial to guarantee the robust operation of the sense amplifier during the design time. That is, for the given design, we should ensure that under process variations $\Delta V \geq \mu_{V_{os}} + 3\sigma_{V_{os}}$ holds. Variations of V_{os} are primarily dependent on the variations of gate lengths (LER variations) of transistors in the cross-coupled inverters, which basically form the positive feedback, the core function of the sense amplifier.

Therefore, M_1, M_4, and M_7, which are not involved in the positive feedback operation, are assumed to have the nominal gate length. For the rest of transistors, an optimal gate length, denoted by L_{opt}, will be derived. Furthermore, the transistor sizing procedure of the sense amplifier is carried out as follows. The number of fins of the transistors of the cross-coupled inverters should be equal, such that the sense amplifier is not biased, and hence are assumed to be single fin devices. As for the transistor M_7, the number of fins mainly impacts the sensing delay, since increasing N_7 allows larger current flow in the circuit. The value of N_7 does not affect V_{os}, so we use $N_7 = 1$. However, the optimal number of fins of isolating transistors M_1 and M_4, which will be referred to as N_{opt}, directly affects the value of V_{os}. More precisely, as N_{opt} increases so does the current through isolating transistors, and as a result, a smaller V_{os} can lead the sense amplifier into metastability condition. For a summary of design variables used during the optimization process, please refer to Table II.

The optimization problem is then formulated as follows.

Find the L_{opt} and N_{opt} values.
Minimize ΔV, subject to $\Delta V \geq \mu_{V_{os}} + 3\sigma_{V_{os}}$.

Increasing the number of fins or transistor gate length are effective solutions to mitigate the effect of process variations [12]. Accordingly, increasing L_{opt} and N_{opt} reduces ΔV. This is verified in Fig. 4 which shows ΔV for various values of L_{opt} and N_{opt}. We can also observe in Fig. 4 the larger impact of N_{opt} compared to that of L_{opt} in reducing ΔV. This is because N_{opt} directly affects the value of V_{os}, whereas

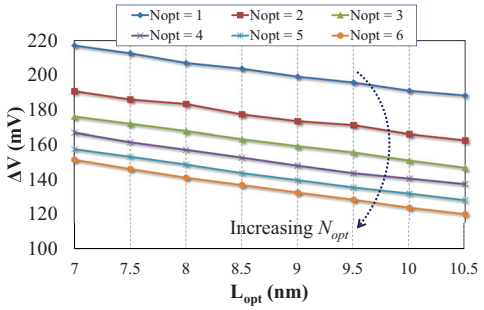

Fig. 4. ΔV for different L_{opt} and N_{opt} values. Increasing L_{opt} and N_{opt} reduces ΔV, but the effect of N_{opt} is more profound.

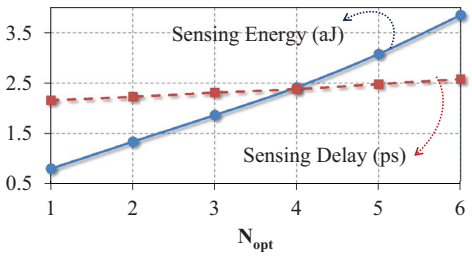

Fig. 5. Delay and energy consumption of sensing 100mV input voltage difference as a function of N_{opt}, assuming 512 SRAM cells on the bitline.

L_{opt} is basically a way by which the cross-coupled transistors alleviate the effect of process variations. On the other hand, increasing N_{opt} slightly increases the sensing delay and, more significantly, increases the sensing energy, as indicated in Fig. 5. However, whereas smaller values of ΔV enhance the cache energy efficiency, delay and energy consumption of the sense amplifier circuit have a negligible impact on the cache access latency and energy consumption, respectively. Other peripheral circuits, especially the row decoder and wordline drivers, are the main dominant contributors to cache access latency and energy consumption. In the next section, the effectiveness of sense amplifier designs are evaluated at the architecture-level.

V. RESULTS

We used a modified version of CACTI with FinFET support [15] in order to assess the effect of FinFET-based sense amplifier designs on cache characteristics. For simulations in

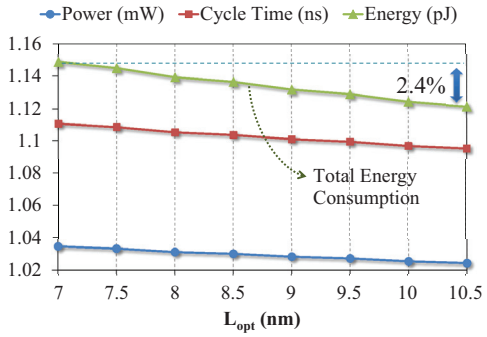

Fig. 6. L1 cache characteristics as a function of L_{opt}, with N_{opt}=1.

TABLE III. COMPARISON OF 32KB L1 CACHE CHARACTERISTICS USING BASELINE AND OPTIMIZED SENSE AMPLIFIER DESIGNS.

Sense Amplifier Design	ΔV (mV)	T_{cycle} (ns)	E_{access} (pJ)	$P_{leakage}$ (mW)	$P_{dynamic}$ (mW)	P_{total} (mW)	E_{total} (pJ)
Baseline (L_{opt}=7nm, N_{opt}=1)	217	1.110	1.479	0.635	1.332	1.034	1.149
Optimized (L_{opt}=10.5nm, N_{opt}=8)	107	1.089	1.150	0.522	1.056	0.838	0.913
Improvement	2×	2%	29%	22%	26%	23%	**26%**

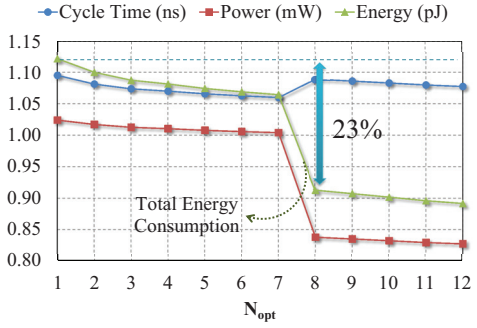

Fig. 7. L1 cache characteristics as a function of N_{opt}, with L_{opt}=10.5nm.

this section, we adopt a 32KB, 4-way set-associative, 64B line, L1 cache memory. We assume 30% of instructions are loads and stores [16], which means the activity factor of the L1 cache is 0.3. Therefore, the total power consumption, P_{total}, and total energy consumption, E_{total}, of the L1 cache memory are calculated as follows:

$$P_{total} = 0.3 \cdot P_{dynamic} + P_{leakage}, \quad (2)$$

$$E_{total} = P_{total} \times T_{cycle}, \quad (3)$$

where $P_{dynamic}$ and $P_{leakage}$ are the dynamic and (active and standby) leakage power consumptions, respectively, and T_{cycle} is the cycle time of the cache memory.

Fig. 6 shows T_{cycle}, P_{total}, and E_{total} of the L1 cache using sense amplifier designs with N_{opt}=1 and different values of L_{opt}, where only 50% increase in the nominal value of L_{opt} is allowed. As can be seen, increasing L_{opt} decreases the cache energy consumption by at most 2.4%. To further reduce the energy consumption, a similar plot, but adopting sense amplifier designs with L_{opt}=10.5nm and different values of N_{opt} is depicted in Fig. 7, where 23% improvement in the energy efficiency is achieved for N_{opt}=8. The sudden decrease of E_{total} in Fig. 7 for N_{opt}=8 is caused by a consequent reduction of ΔV which allows CACTI to find a better cache organization that even improves the cache leakage power. Hence, N_{opt} is an important decision variable for the design of energy efficient cache memories with robust sense amplifiers.

Since a column of SRAM cells share a sense amplifier, the area of a sense amplifier cell is not as critical as that of the SRAM cell. Therefore, we pick N_{opt}=8 and L_{opt}=10.5nm for the optimized sense amplifier design. Table III compares L1 cache characterization results using baseline (L_{opt}=7nm, N_{opt}=1) and optimized (L_{opt}=10.5nm, N_{opt}=8) sense amplifier designs. The optimized design reduces ΔV by a factor of 2 compared with the baseline counterpart. This 2-fold reduction in ΔV finally causes 26% improvement in the energy efficiency of the L1 cache memory.

VI. CONCLUSIONS

We optimized the 7T sense amplifier design for a 7nm FinFET technology in order to improve the energy efficiency of the cache memory. The optimization procedure took into account process variation effects such that the robust operation of the sense amplifier could be achieved. According to our architecture-level simulations on an L1 cache memory, the optimized sense amplifier design has 26% higher energy efficiency compared with the baseline counterpart.

VII. ACKNOWLEDGMENTS

This research is supported by grants from the PERFECT program of the Defense Advanced Research Projects Agency, the Software and Hardware Foundations of the National Science Foundation, and the Brazilian research agencies CAPES, CNPq, and FAPERJ.

REFERENCES

[1] B. Wicht, T. Nirschl, and D. Schmitt-Landsiedel, "Yield and speed optimization of a latch-type voltage sense amplifier," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 39, no. 7, pp. 1148–1158, July 2004.

[2] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "A novel high-performance and robust sense amplifier using independent gate control in sub-50-nm double-gate mosfet," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 14, no. 2, pp. 183–192, Feb 2006.

[3] S. Tang *et al.*, "Finfet - a quasi-planar double-gate mosfet," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2001.

[4] T. Matsukawa *et al.*, "Comprehensive analysis of variability sources of finfet characteristics," in *Symposium on VLSI Technology*, 2009.

[5] X. Wang, A. Brown, B. Cheng, and A. Asenov, "Statistical variability and reliability in nanoscale finfets," in *IEEE International Electron Devices Meeting (IEDM)*, Dec 2011, pp. 5.4.1–5.4.4.

[6] E. Nowak *et al.*, "Turning silicon on its edge [double gate cmos/finfet technology]," *IEEE Circuits and Devices Magazine*, 20(1), 2004.

[7] Z. Guo *et al.*, "Finfet-based sram design," in *International Symposium on Low Power Electronics and Design (ISLPED)*, Aug 2005, pp. 2–7.

[8] F. Moradi *et al.*, "Asymmetrically doped finfets for low-power robust srams," *IEEE Transactions on Electron Devices*, 58(12), 2011.

[9] M.-L. Fan *et al.*, "Variability analysis of sense amplifier for finfet sub-threshold sram applications," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 59, no. 12, pp. 878–882, Dec 2012.

[10] S. Chen *et al.*, "Performance Prediction for Multiple-Threshold 7nm-FinFET-based Circuits Operating in Multiple Voltage Regimes using a Cross-Layer Simulation Framework," in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, Oct. 2014.

[11] Synopsys technology computer-aided design (TCAD). [Online]. Available: http://www.synopsys.com/tools/tcad

[12] J. Kwong and A. Chandrakasan, "Variation-driven device sizing for minimum energy sub-threshold circuits," in *International Symposium on Low Power Electronics and Design (ISLPED)*, Oct 2006, pp. 8–13.

[13] K. Patel, T.-J. K. Liu, and C. J. Spanos, "Gate line edge roughness model for estimation of finfet performance variability," *IEEE Transactions on Electron Devices*, vol. 56, no. 12, pp. 3055–3063, Dec 2009.

[14] T. Sakurai and A. Newton, "Alpha-power law mosfet model and its applications to cmos inverter delay and other formulas," *IEEE Journal of Solid-State Circuits*, vol. 25, no. 2, pp. 584–594, Apr 1990.

[15] A. Shafaei, Y. Wang, X. Lin, and M. Pedram, "Fincacti: Architectural analysis and modeling of caches with deeply-scaled finfet devices," in *IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, July 2014, pp. 290–295.

[16] G. Reinman *et al.*, "Classifying load and store instructions for memory renaming," in *Proceedings of the 13th International Conference on Supercomputing (ICS)*, 1999, pp. 399–407.