A 48,000 pixel, 590,000 transistor silicon retina in current-mode subthreshold CMOS

Andreas G. Andreou

Electrical and Computer Engineerng
Johns Hopkins University
3400 N. Charles Street, Baltimore, MD 21218.

Kwabena A. Boahen

CNS Program
Caltech
Pasadena, CA 91125.

Abstract

A second generation, contrast sensitive silicon retina is reported in this paper. The architecture and organization is inspired by the outer plexiform processing in the vertebrate retina. Current-mode subhtreshold MOS design techniques are employed to obtain high performance and energetic efficiency. The system has been fabricated with 230×210 *pixels on a* 1×1 *cm die in a* $1.2\mu m$ *n–well double metal, double poly, digital oriented CMOS technology. The chip incorporates 590,000 transistors, 48,000 pixels, operating in subthreshold/transition region with power dissipation of 50 mW when powered from a 5V power supply. The pixel has a frequency response of 100Khz.*

1 Introduction

Biological organisms excel at solving problems in sensory communication and motor control, by sustaining high computational throughput with minimal energy dissipation. Their effectiveness stems partly from exploiting *prior* knowledge about the problems that they encounter [1]. Such information in the form of *internal models,* reflects the statistical properties of the natural environments in which the systems function. Since the environment is rarely fixed, model *adaptation* and *self-organization* is necessary [2, 3]. However, the algorithms that are employed in neural computation must be implemented on a computational substrate that is *analog* in nature and therefore elementary computational primitives must emerge out of the physics of the computational substrate. Motivated by the principles of organization and function in neural systems, analog VLSI [4, 5] is a synthetic approach aimed at exploring the possibility of synthesizing complex analog information processing systems on silicon.

In this paper, we discuss how such a methodology has led to the development of an *analog* VLSI silicon system for a second generation, contrast sensitive, silicon retina [6]. The architecture is inspired by the processing performed at the outer plexiform layer of the vertebrate retina. It is mapped onto silicon using circuits of minimal complexity that exploit native properties of subthreshold MOS transistors. High computational throughput at low levels of energy dissipation is achieved by employing analog processing in a massively parallel architecture; a paradigm that minimizes the "mismatch" between the physics of the problem and the physics of the computational substrate. We begin with a discussion of the diffusor, the basic building block for current–mode analog networks.

2 The Diffusor

Key to the success in the realization of the system is the *diffusor* [6]; the MOS transistor operating in subthreshold ohmic regime where the current is an exact difference of exponential functions of the drain and source voltages [5] and for an NMOS the current is given by:

$$I = I_0 \, S \, exp(\kappa \, V_{GB}) \, [exp(-V_{SB}) - exp(-V_{DB})] \quad (1)$$

The terminal voltages V_{GB}, V_{SB}, V_{DB} are referenced to the substrate and are normalized to the thermal voltage (kT/q). The constant I_0 depends on mobility (μ) and other silicon physical properties. S is a geometry factor, the width W to length L ratio the device. The constant κ takes values between 0.6 and 0.9.

The exponential functions of voltage in the square brackets of Equation 1, correspond to Boltzmann distributed charges at the source and drain.

$$I \propto [\mathcal{Q}_S - \mathcal{Q}_D] \quad (2)$$

0-7803-2428-5/95 $4.00

The charge-based representation depicted in Equation 2, suggests that the MOS transistor in subthreshold is a highly linear device; a property that finds many uses in analog circuit design. This property was first observed by Kwabena Boahen and discussed in [6] where the concept of a *diffusor* was introduced. The view of an MOS transistor in subthreshold as a basic diffusive element allows for the effective implementation of systems that exploit properties of elliptic partial differential equations. With the appropriate logarithmic loads connected to the source and drain, linear networks can be obtained. The same idea was more recently revisited by Vittoz and Arreguit [7]. The system described in this paper is what we believe the first large scale application of this novel circuit design technique.

To demonstrate how computational primitives emerge at the network level from device physics of the underlying technology, let us consider an example of a summing operation, *local aggregation*. Such linear addition of signals over a confined region of space occurs throughout the nervous system. Aggregation was discussed in Chapter 6 of [4], and it is the basis for many neuromorphic silicon VLSI systems described therein. Here we take a close look at *diffusion*, the physical process that underlies local aggregation in the nervous system, contrast it with the process of diffusion in MOS transistors and come up with a novel network design technique.

The diffusion process is described by the following equation:

$$\frac{dN}{dt} = D\nabla^2 N(x,y) \tag{3}$$

N is the concentration of the diffusing species and D is their diffusivity. Equation 3 applies to the 2-D case where the concentration is assumed uniform in the third dimension and N is the number of particles per unit area. Two alternative analog simulations of this process on a discrete grid are shown in Figure 1.

The first network uses voltages and currents (Figure 1a). Its node equation is

$$\frac{dV_n}{dt} = \frac{4G}{C}\left(\frac{1}{4}(V_j + V_k + V_l + V_m) - V_n\right) \tag{4}$$

which is homologous with Equation 3 since the term in large parenthesis is a first-order approximation to the Laplacian. However, this solution is not amenable to VLSI integration because transconductances (G) with a large linear range consume large amounts of area and power.

The second network uses charges (positive) and currents (Figure 1b). Its node equation is

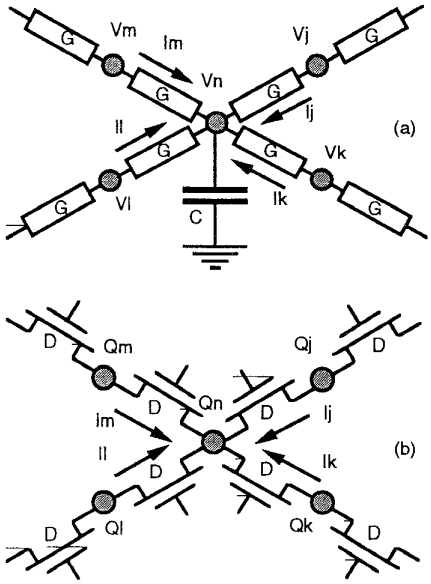

Figure 1: Simulating diffusion with (a) conductances and voltage/current variables or (b) diffusors and charge/current variables.

$$\frac{dQ_n}{dt} = 4D\left(\frac{1}{4}(Q_j + Q_k + Q_l + Q_m) - Q_n\right) \tag{5}$$

Note that dQ_n/dt is the same as the current supplied to node n by the network. This solution is easily realized by exploiting diffusion in subthreshold MOS transistors. It was shown earlier that in the MOS transistor, the current is linearly proportional to the charge difference across the channel (See Equation 2). Therefore, the diffusion process may be modeled using devices with identical geometry S and identical gate voltages. The former guarantees they have the same diffusivity and the latter guarantees that the charge concentrations at all the source/drains connected to node n are the same and equal Q_n.

In both of these networks, the boundary conditions may be set up by injecting current into the appropriate nodes. In the voltage-mode network, the solution is the node voltages. They are easily read without disturbing the network. On the otherhand, the network in Figure 5 represents the solution by charge concentrations $\mathcal{Q}_S$ and $\mathcal{Q}_D$ at source/drains—not the charge on the node capacitance. The source/drain charge cannot be measured directly without disturbing the network. It may be inferred from the node voltage.

When high conductances are attached to the nodes where currents are injected, the network is said to be

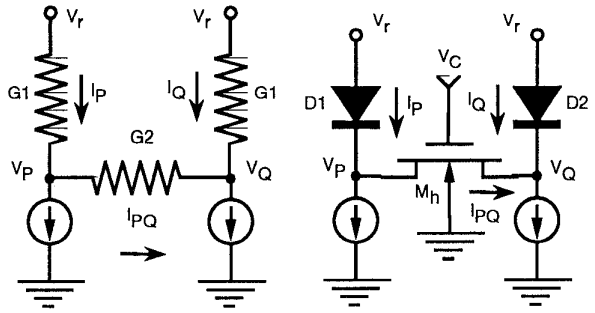

Figure 2: Building blocks for linear networks. Using segments that employ **ideal** (left) linear and (right) non-linear elements.

loaded and its characteristics change. To observe the behaviour of a loaded network, we begin with a small segment of a one dimensional network (Figure 2).

A voltage mode circuit model for a loaded network is shown in Figure 2(left) for which:

$$I_{PQ} = (G_1/G_2)(I_Q - I_P)$$

This is a lumped parameter model where G_1 and G_2 correspond to resistances per unit length. The voltages on nodes P and Q referenced to ground, represent the state of the network and can be read out using a differential amplifier with the negative input grounded.

The equivalent circuit using idealized non-linear conductances is shown in Figure 2(right). The difference in currents through the diodes D_1 and D_2 are linearly related to the current through the diffusor MOS transistor. This relationship can be derived from Equation 1 describing subthreshold conduction, and the ideal diode characteristics where $I_D = I_S \exp[qV_D/(kT)]$. An expression can be derived for the current I_{PQ} in terms of the currents I_P and I_Q, the reference voltage V_r and the bias voltage V_C, where:

$$I_{PQ} = \left(\frac{SI_{on}}{I_S}\right) \exp\left[\frac{\kappa V_C - V_r}{(kT/q)}\right](I_Q - I_P) \qquad (6)$$

The current I_{on} and S is the zero intercept current and geometry factor respectively for the diffusor transistor M_h. I_S is the reverse saturation current for the diode that is assumed to be ideal. The currents in these circuits are identical if

$$\frac{G_1}{G_2} = (\frac{SI_{on}}{I_S}) \exp\left[\frac{\kappa V_C - V_r}{(kT/q)}\right]$$

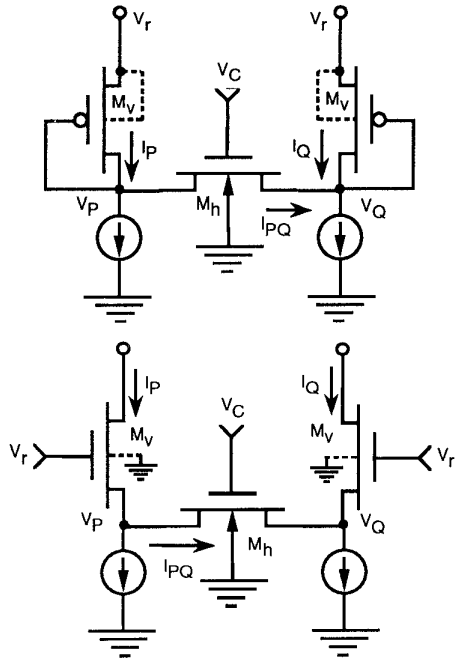

Figure 3: Current–mode building blocks for linear networks using (top) PMOS transistor implementation, (bottom) NMOS single transistor current-conveyor implementation.

Increasing V_C or reducing V_r has the same effect as increasing G_1 or reducing G_2. The state of this network is represented by the charge at the nodes P and Q. Since the anode of a diode is the reference level (zero negative charge), the currents I_P and I_Q represent the result. Unfortunately, the anode of a diode or a diode connected transistor is not a good current source. When diodes are not explicitely available in the process, diode connected PMOS or NMOS transistors can be used as shown in Figure 3. When the loads are PMOS, the current current I_{PQ} is:

$$I_{PQ} = \left(\frac{S_h I_{onh}}{S_v I_{onv}}\right) \exp\left[\frac{\kappa_h V_C - \kappa_v V_r}{(kT/q)}\right](I_Q{}^{1/\kappa_v} - I_P{}^{1/\kappa_v}) \qquad (7)$$

When NMOS transistors are used as loads, there is the additional benefit, that of exploiting the current conveying properties of a single transistor [5], to obtain the current outputs I_P and I_Q, on nodes that are low conductance (the drain terminal are now excellent outputs for the currents). Using Equations 8.45 in [5], the current I_{PQ} is given as:

$$I_{PQ} = \left(\frac{S_h I_{onh}}{S_v I_{onv}}\right) \exp\left[\frac{\kappa_h V_C - \kappa_v V_r}{(kT/q)}\right] (I_Q - I_P) \quad (8)$$

where S_h, I_{onh}, κ_h and S_v, I_{onv}, κ_v are geometry, zero bias intercept current and subthreshold slope parameters for transistors M_h and M_v, respectively.

3 A Contrast Sensitive Silicon Retina

The analog silicon system is modeled after neurocircuitry in the distal part of the vertebrate retina—called the outer-plexiform layer. Figure 4 illustrates interactions between cells in this layer [9]. The well-known center/surround receptive field emerges from this simple structure, consisting of just two types of neurons. Unlike the ganglion cells in the inner retina and the majority of neurons in the nervous system, the neurons that we model here have graded responses (they do not spike); thus this system is well-suited to analog VLSI.

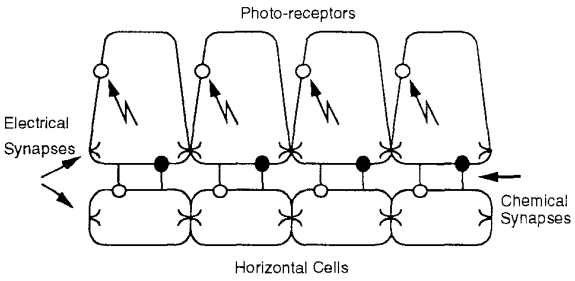

Figure 4: One-dimensional model of neurons and synapses in the outer-plexiform layer. Based on the red-cone system in the turtle retina.

The photoreceptors are activated by light; they produce activity in the horizontal cells through excitatory chemical synapses. The horizontal cells, in turn, suppress the activity of the receptors through inhibitory chemical synapses. The receptors and horizontal cells are electrically coupled to their neighbors by electrical synapses. These allow ionic currents to flow from one cell to another, and are characterized by a certain conductance per unit area.

In the biological system, contrast sensitivity –the normalized output that is proportional to a local measure of contrast– is obtained by shunting inhibition. The horizontal cells compute the local average intensity and modulate a conductance in the cone membrane proportionately. Since the current supplied by the cone outer-segment is divided by this conductance to produce the membrane voltage, the cone's response will be proportional to the ratio between its photoinput and the local average, i. e. to contrast. This is a very simplified abstraction of the complex ion-channel dynamics involved. The advantage of performing this complex operation at the focal plane is that the dynamic range is extended (local automatic gain control).

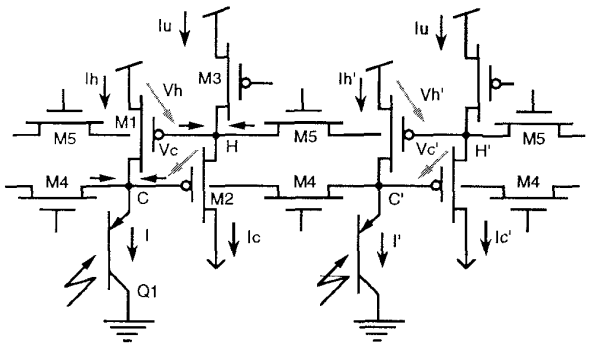

Figure 5: One-dimensional implementation of outer-plexiform retinal processing. There are two diffusive networks implemented by transistors M_4 and M_5, which model electrical synapses. These are coupled together by controlled current-sources (devices M_1 and M_2) that model chemical synapses. Nodes H in the upper layer correspond to horizontal cells while those in the lower layer (C) correspond to cones. The bipolar phototransistor Q_1 models the outer segment of the cone and M_3 models a leak in the horizontal cell membrane. Note that the actual system has a six neighbor connectivity.

The basic analog MOS circuitry for a one dimensional pixel with two neighbor connectivity is shown in Figure 5. The analysis of the system can be found in [6, 5], here we present an outline and approximations to the main results.

We begin with the non-linear aspects of system operation, its *contrast sensitivity*. The non-linear operation that leads to a local gain-control mechanism in the silicon system is acheived through a mechanism that is qualitatively similar to the biological counterpart, but quantitatively different (see discussion in [6]). Refering to Figure 5, the output current $I_c(x_m, y_n)$ at each pixel, can be given (approximately) in terms of the input photocurrent $I(x_m, y_n)$ and a local average of this photocurrent in a pixel neighborhood (M, N). This region may extend beyond the nearest neighbor. The fixed current I_u supplied by transistor M_3 normalizes the result.

$$I_c(x_m, y_n) = I_u \frac{I(x_m, y_n)}{\left(I(x_m, y_n) + \sum_{\mathrm{M,N}} I(x_i, y_j)\right)} \quad (9)$$

At any particular intensity level, the outer-plexiform behaves like a linear system that realizes a powerful second-order regularization algorithm for edge detection. This can be seen by performing an analysis of the circuit about a fixed operating point. To simplify the equations we first assume that $\hat{g} = \langle I_h \rangle g$, where $\langle I_h \rangle$ is the local average. Now we treat the diffusors (devices M_4) between nodes C and C' as if they had a fixed diffusitivity $\hat{g}$. The diffusitivity of the devices M_5 between nodes H and H' in the horizontal network is denoted by h. Then the simplified equations describing the full two-dimensional circuit on a square grid are:

$$I_h(x_m, y_n) = I(x_m, y_n) + \hat{g} \sum_{\substack{i=m\pm1 \\ j=n\pm1}} \{I_c(x_i, y_j) - I_c(x_m, y_n)\}$$

$$I_c(x_m, y_n) = I_u + h \sum_{\substack{i=m\pm1 \\ j=n\pm1}} \{I_h(x_m, y_n) - I_h(x_i, y_j)\}$$

Using the second-difference approximation for the laplacian, we obtain the continuous versions of these equations

$$I_h(x, y) = I(x, y) + \hat{g}\nabla^2 I_c(x, y) \quad (10)$$

$$I_c(x, y) = I_u - h\nabla^2 I_h(x, y) \quad (11)$$

with the internode distance normalized to unity. Solving for $I_h(x, y)$, we find

$$\hat{g}h\nabla^2\nabla^2 I_h(x, y) + I_h(x, y) = I(x_i, y_j) \quad (12)$$

This is the *biharmonic* equation used in computer vision to find an optimally smooth interpolating function $I_h(x, y)$ for the noisy, spatially sampled data $I(x_i, y_j)$; it yields the function with minimum energy in its second derivative [10]. The coefficient $\lambda = \hat{g}h$ is called the regularizing parameter; it determines the trade-off between smoothing and fitting the data.

A one dimensional solution to this equation can be obtained using Green's functions valid for vanishing boundary conditions at plus and minus infinity; this has the characteristic mexican hat shape.

$$I_h(x, \lambda) = \frac{1}{2\lambda^{1/4}} \exp(-|x|/\sqrt{2}\lambda^{1/4}) \cos\left(\frac{|x|}{\sqrt{2}\lambda^{1/4}} - \frac{\pi}{4}\right) \quad (13)$$

In the original work [6], the chip was fabricated with 90×92 pixels on a 6.8×6.9 mm die in a 2μm n-well double metal, double poly, garden variety digital oriented CMOS technology and was fully functional. More recently the same system has been fabricated with 230×210 pixels on a 1×1 cm die in a 1.2μm n-well double metal, double poly, digital oriented CMOS technology. The chip incorporates 590,000 transistors, 48,000 pixels, operating in subthreshold/transition region with power dissipation on the order of a few mW when powered from a 5V power supply. Temporal response is in the order of a few microseconds. The chip incorporates a video pre-amplifier and some digital circuitry for scanning the processed images out of the array. Standard NTSC video is produced off-chip using an FPGA controller and a video amplifier.

An image captured through the silicon retina is shown in Figure 6. Note the edge enhancement properties of the system and the absence of a dynamic range (flat image).

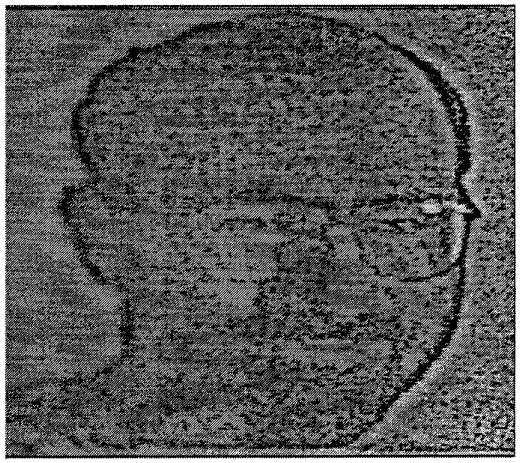

Figure 6: An image of the author as captured by the retina system.

4 Discussion

The analog VLSI system presented in this paper is essentially an *analog floating point* processor. As a first step, the system computes the range (the voltages in the horizontal cell synthycium correspond to the value of the exponent). This is the operating point of the system; that is also how the automatic gain control is achieved. At an operating point, sophisticated spatial filtering is performed to smooth the sampled data and enhance the edges. Having separated the problem

of precision and dynamic range, the signal processing within the range can be done with low precision analog hardware. The issue of precision versus dynamic range in analog circuits was addressed by Barrie Gilbert, 10 years ago with his elegant implementation of an "array normalizer" that used bipolar transistors and current–mode translinear circuits [8]. The system presented here is similar in two ways to Gilbert's array normalizer. First there is *local* normalization of the input current signals. Second, all processing is done in the current domain where the translinear properties of MOS subthreshold devices are exploited to implement the required functions. For a detail discussion on Translinear circuits in subthreshold MOS please refer to [5].

A conservative estimate for the energetic efficiency can be obtained by assuming that a total of 18 low precision operations (OP) are performed per pixel. Six operations are necessary for the convolution with with bandpass kernel of Equation 13, six for the Laplacian operator (Equation 11) and six for the local gain control computation (Equation 9). If the system is biased so that at the pixel level the frequency response is 100Khz, approximately 1×10^{12} low precision calculations per second are performed in the (210×230) pixels. The power dissipation under the above biasing conditions is about 50mW when operating from 5 Volt power supplies. This is equivalent to 0.05 pW/OP. This performance is a result of an optimization done at the system level, by mapping the problem on an effective physical computational model, rather than trying to optimize the energetic efficiency of an individual gate.

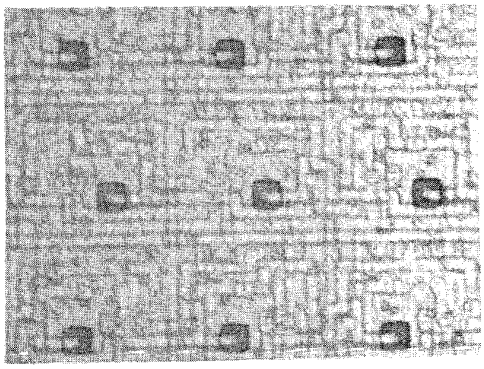

Figure 7: Photomicrograph of the chip. The surface is covered by second metal except where there are openings for the phototransistors (the dark square areas). Note the hexagonal connectivity between the pixels.

Acknowledgments: One of the authors (AGA) was supported in part by Nissan corporation and by a Research Initiation Award from NSF (MIP-9010364). We thank Professor Carver Mead for making us believe in the power of low energy analog computation and for his encouragement over the last 6 years.

References

[1] H.B. Barlow, "Unsupervised Learning," Neural Computation, Vol. 1, No. 3, pp. 295-311, Fall 1989.

[2] T. Kohonen, *Self–Organization and Associative Memory,* Springer Verlag, (2nd edition), Berlin, Heidelberg New York, 1988; A.L. Gorin, S. Levinson, A. Gertner and E. Goldman, " Adaptive Aqcuisition of Language," Computer Speech and Language, Vol. 5, No. 2, pp. 101-132, April 1991; S. Haykin, *Neural Networks; A comprehensive foundation,* McMillan College Publishing, New York, 1994.

[3] C.A. Mead, "Neuromorphic electronic systems," *Proceedings IEEE,* vol. 78, no. 10, pp. 1629-1636, Oct. 1990.

[4] C.A. Mead, *Analog VLSI and Neural Systems,* Reading, MA: Addison-Wesley, 1989.

[5] A.G. Andreou and K.A. Boahen, "Neural Information Processing (II)," Chapter 8, in *Analog VLSI Signal and Information Processing,* M. Ismail and T. Fiez eds., McGraw-Hill, 1994.

[6] K.A. Boahen and A.G. Andreou, "A contrast sensitive silicon retina with reciprocal synapses," *Advances in Neural Information Processing Systems 4,* Moody, J.E., Hanson, S.J. and Lippmann, R.P. (eds.), Morgan Kaufmann Publishers, San Mateo, CA 1992.

[7] E. Vittoz and X. Arreguit, "Linear networks based on transistors," Electronics Letters, vol. 29, pp. 297-299, Feb. 4th, 1993.

[8] B. Gilbert, "A Monolithic 16-Channel Analog Array Normalizer," *IEEE Journ. of Solid-State Circuits,* vol. SC-19, No. 6, 1984.

[9] J. E. Dowling, "The retina: an approachable part of the brain," The Belknap Press of Harvard University, Cambridge, MA, 1987.

[10] T. Poggio, V. Torre and C. Koch, "Computational vision and regularization theory," *Nature,* 317, pp. 314-319, 1985.