Low Power Multiplier-Accumulator

*K.N Varaprasad, Dr.Nisha Sarwade and Ch. M Krishn*a

ABSTRACT: Power dissipation is recognized as a critical parameter in modern VLSI design field. To satisfy MOORE'S law and to produce consumer electronics goods with more backup and less weight, low power VLSI design is necessary. High speed and low power Multiplier-Accumulator(MAC) units are required for applications of digital signal processing like Fast Fourier Transform, Finite Impulse Response filters, convolution etc. The core of every microprocessor, DSP, and data-processing ASIC is its data path. Statistics showed that more than 70% of the instructions perform additions and multiplications in the data path of RISC machines. At the heart of data-path and addressing units in turn are arithmetic units, such as comparators, adders, and multipliers. Digital multipliers are the most commonly used components in any digital circuit design. Multiplication based operations such as Multiply and Accumulate and inner product are among some of the frequently used Computation-Intensive Arithmetic Functions, currently implemented in many DSP applications such as convolution, fast Fourier transform, filtering and in microprocessors in its arithmetic and logic unit. Since multiplication dominates the execution time of most DSP algorithms, so there is a need of low power and high speed multiplier. A review of recent trends in MAC are presented here.

***Keywords*—MAC, Partially Guarded Computation(PGC), Spurious-Power Suppression Technique(SPST).**

I. INTRODUCTION:

ONE OF THE accompanying challenges in designing ICs for portable electrical devices is lowering down the power consumption to prolong the operating time on the basis of given limited energy supply from batteries. with the recent rapid development in multimedia and communication systems, digital signal processing are increasingly being demanded. The multiplier and multiplier-and-accumulator (MAC)[1] are the essential elements of the digital signal processing such as filtering, convolution, and inner products. Most digital signal processing methods use nonlinear functions such as discrete cosine transform (DCT)[2] or discrete wavelet transform (DWT). Because they are basically accomplished by repetitive application of multiplication and addition, Multiplication is an important operation in digital signal processing algorithms. It should be small in area, and consumes minimum power. Therefore, there is need of designing low power high speed multiplier.

K.N Varaprasad and ,Dr.Nisha Sarwade are with Department of Electrical Engineering, VJTI, Mumbai, Maharashtra, INDIA. And Ch. M Krishna is with Department of Electronics Engineering, COEP ,Pune, Maharashtra, INDIA, Emails: varaprasad.konakalla@gmail.com, nishasarvade@vjti.org.in, challagollamkrishna@gmail.com

Extensive research has been carried out on low power and high speed multipliers at technology, physical, circuit and logic levels. These low-level techniques are not unique to multiplier modules and they are generally applicable to other types of modules. Moreover, power consumption is directly related to data switching patterns. However, it is difficult to consider application-specific data characteristics in low-level power optimization. Various techniques have been developed for reducing the power consumption of VLSI designs, including voltage scaling, switched-capacitance reduction, clock gating, power-down techniques, threshold-voltage controlling, multiple supply voltages, and dynamic voltage frequency scaling . These low-power techniques have been proven to be efficient at certain expense and are applicable to multimedia/DSP designs. Among these low-power techniques, a promising direction for significantly reducing power consumption is reducing the dynamic power which dominates total power dissipation.

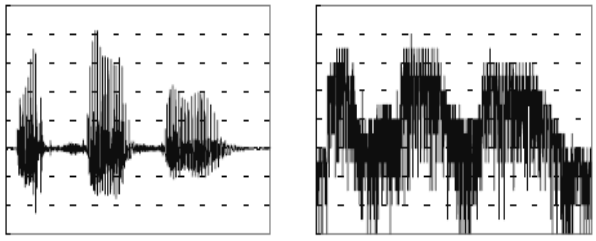

Fig. 1. Example speech data and associated range.

This paper presents a review of Low power MACs. The organization of this paper is as follows Section II gives Architecture of Mac followed by Section III which provides a review of recent trends of Low power Mac concluded by Section IV.

II. MAC ARCHITECTURE:

This chapter introduces the basics of binary multiplication, partial product generation, reduction and techniques to make the multiplication process faster. The multiplication and accumulation is the main computational kernel in Digital Signal Processing architectures. The MAC unit determines the speed of overall system as it is always lies in the critical path. Developing high speed MAC is crucial for real time DSP application. In order to improve the speed of the MAC unit, there are two major bottlenecks that need to be considered. The first one is the fast multiplication network and the second one is the accumulator. Both of these stages require addition of large operands that involve long paths for carry propagation. In recent Mac accumulation and addition are merge to save the time and power. The MAC unit basically do the multiplication of two umbers multiplier and

multiplicand and add that product in result stored in the accumulator.

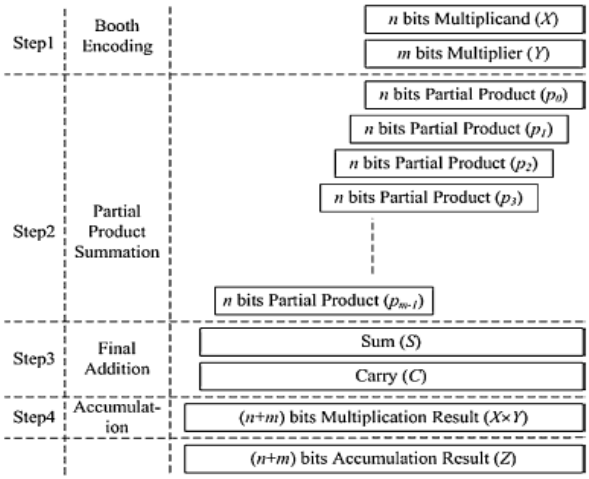

Figure 2: Basic arithmetic steps of multiplication and accumulation

The Fig.2 shows Basic arithmetic steps of multiplication and accumulation. The general construction of the MAC operation can be represented by this equation: Z = X *Y + Z
Where the multiplier A and multiplicand B are assumed to have n bits each and the addend Z
has (2n+1) bits. A basic MAC unit can be divided into two main blocks .

1.Multiplier
2.Accumulator

A Fast Multiplication process consists of three steps

- Partial Product Generation.
- Partial Product Reduction.
- Final stage Carry Propagate Adder.

To generate the number of partial product Radix-4 Modified booth encoding techniques have been used. The Modified Booth Encoding (MBE) or Modified Booth's Algorithm (MBA) was proposed by O. L. Macsorley in 1961 . Booth's radix-4 algorithm is widely used to reduce the area of multiplier and to increase the speed. The booth encoding algorithm is a bit-pair encoding algorithm that generates partial products which are multiples of the multiplicand. The booth algorithm shifts and/or complements the multiplicand (X operand) based on the bit patterns of the multiplier (Y operand). Essentially, three multiplier bits [Y (i+1) ,Y (i) and Y (i-1)] are encoded into eight bits that are used to select multiples of the multiplicand [-2X,-X,0,+X,+2X]. The three multiplier bits consist of a new bit pair [Y (i+1) and Y (i)] and the leftmost bit from the previously encoded bit pair [Y (i-1)]. Grouping the three bits of multiplier with overlapping has half partial products which improve the system speed Multiplier require high amount of power and delay during the partial products addition. At this stage, most of the multipliers are designed with different kind of multi operands adders that are capable to add more than two input operands and results in two outputs, sum and carry. The number of adders will be minimized by Wallace Tree.

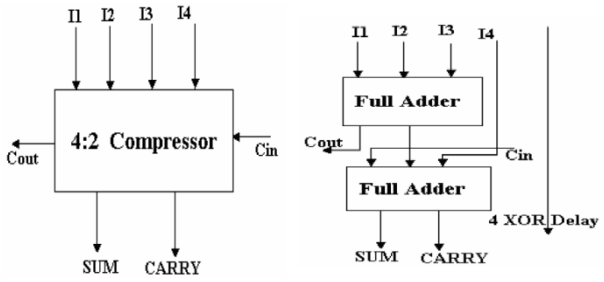

Figure 3: Block Diagram of 4:2 Compressor

In addition stage 4-2 compressors are used as carry save adders. The 4-2 and 5-2 compressors have been widely employed in the high speed multipliers to lower the latency of the partial product accumulation stage. Owing to its regular interconnection, the 4-2 compressor is ideal for the partial products addition stage. The 4:2 compressor structure actually compresses five partial products bits into three. The architecture is connected in such a way that four of the inputs are coming from the same bit position of the weight j while one bit is fed from the neighbouring position j-1(known as carry-in). The outputs of 4:2 compressor consists of one bit in the position j and two bits in the position j+1.This structure is called compressor since it compresses four partial products into two parts. A 4-2 compressor can also be built using 3-2 compressors. It consists of two 3-2 compressors (full adders) in series and involves a critical path of 4 XOR delays. The output C_{out}, being independent of the input C_{in} accelerates the carry save summation of the partial products. Fig.3 shows Hardware Architecture of general MAC Array Multiplier. Fig.3 shows the block diagram of 4:2 compressor and compressor with full adder.

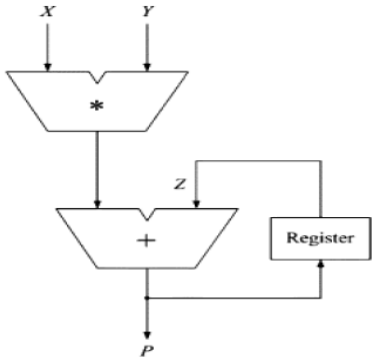

Figure 3: Hardware Architecture of general MAC Array Multiplier

III.POWER OPTIMAZATION:

Power dissipation has emerged as an important design parameter in the design of microelectronic circuits, especially in portable computing and personal communication applications. Addition is very important operation in any digital design. If we can make adder to work with minimum delay and minimum power it will reflect on final design. Many adders are introduced but there is lot to be implanted in addition. If we use ripple carry adder delay will be propagated through-out the process of addition which is undesirable. The delay is proportional to width of operand which is undesirable. Carry Look ahead

Adder(CLA) which somewhat better compared to RCA. If the length of operand is very high CLA is also not recommended because of its complex structure due to which there is large area overhead and power consumption. So later some of adders like Carry skip adder, Carry save adder and Carry select adders were introduced. But each of adder has its advantages and drawbacks. But according our requirement we are going to use these adders in our design. Since core power consumption must be dissipated through the packaging, increasingly expensive packaging and cooling strategies are required as chip power consumption increases. In addition to cost, there is the issue of reliability. High power systems often run hot, and high temperature tends to exacerbate several silicon failure mechanisms. Another crucial driving factor is that excessive power consumption is becoming the limiting factor in integrating more transistors on a single chip or on a multiple-chip module.

(A)SOURCES OF POWER DISSIPATION: Power dissipation in CMOS circuits is caused by three sources: 1) the leakage current which is primarily determined by the fabrication technology, consists of reverse bias current in the parasitic diodes formed between source and drain diffusions and the bulk region in a MOS transistor as well as the sub-threshold current that arises from the inversion charge that exists at the gate voltages below the threshold voltage, 2) the short-circuit current which is due to the DC path between the supply rails during output transitions and 3) the charging and discharging of capacitive loads during logic changes.

(B)LOW POWER DESIGN SPACE: The previous section revealed the three degrees of freedom inherent in the low-power design space are voltage, physical capacitance, and data activity. Optimizing for power entails an attempt to reduce one or more of these factors. But we are going to discuss more about switching activity in this paper switching activity also influences dynamic power consumption. A chip may contain an enormous amount of physical capacitance, but if there is no switching in the circuit, then no dynamic power will be consumed. The data activity determines how often this switching occurs. There are two components to switching activity: fclk which determines the average periodicity of data arrivals and E(sw) which determines how many transitions each arrival will generate. For circuits that do not experience glitching, E(sw) can be interpreted as the probability that a power consuming transition will occur during a single data period. Even for these circuits, calculation of E(sw) is difficult as it depends not only on the switching activities of the circuit inputs and the logic function computed by the circuit, but also on the spatial and temporal correlations among the circuit inputs. The data activity inside a 16-bit multiplier may change by as much as one order of magnitude as a function of input correlations For certain logic styles, however, glitching can be an important source of signal activity and, therefore, deserves some mention here. Glitching refers to spurious and unwanted transitions that occur before a node settles down to its final steady-state value. Glitching often arises when paths with unbalanced propagation delays converge at the same point in the circuit. Since glitching can cause a node to make several power consuming transitions, it should be avoided whenever possible. The data activity E(sw) can be combined with the physical capacitance C to obtain switched capacitance, Csw=C.E(sw), which describes the average capacitance charged during each data period 1/fclk. It should be noted that it is the switched capacitance that determines the power consumed by a CMOS circuit. In high-level synthesis domain, there have been quite a few studies devoted to minimize transitions in functional units, registers, multiplexers, and buses [3 - 11]. Many of them focus on minimizing transition activity in functional units because they are the main source of power dissipation in data dominated applications [3 - 8]. The most effective method to reduce the number of transitions in functional units is increasing the correlation of input data. Therefore, many of the previous work focus on increasing input data correlation by changing operation binding [3],[8] loop pipelining [7], loop interchange, operand reordering, operand sharing, unrolling [5], and guarded evaluation[11].The existing works that reduce the dynamic power consumption by minimizing the switched capacitance include the designs in [13]–[18]. The design in [13] proposes a concept called partially guarded computation (PGC), which divides the arithmetic units, e.g., adders and multipliers, into two parts and turns off the unused part to minimize the power consumption. The reported results show that the PGC can reduce power consumption by 10%–44% in an array multiplier with 30%–36% area overheads in speech-related applications. However, the PGC technique cannot gain any power reduction when applied on adders because of the overhead-augmented circuitry. The design in [14] proposes a 32-bit 2's complement adder equipping a two-stage (master and slave stages) flip-flop at each of the two inputs, a dynamic-range determination (DRD) unit and a sign-extension (SE) unit, which tends to reduce the power dissipation of conventional adders for multimedia applications. Additionally, the design in [15] presents a multiplier using the DRD unit to select the input operand with a smaller effective dynamic range to yield the Booth codes. However, the DRD unit induces additional delay and area overheads. Besides, the input data flows are also frequently switched if the input operands with a smaller effective dynamic range often change between operands A and B, and vice versa. In such cases, the power dissipation of the designs in [14] and [15] is increased rather than decreased. The design in [16] incorporates a technique for glitching power minimization by replacing some existing gates with functionally equivalent ones that can be frozen by asserting a control signal. This technique can be applied to replace layout-level descriptions and guarantees predictable results. However, it can only achieve savings of 6.3% in total power dissipation, since it operates in the layout-level environment which is tightly restricted. One of the most advanced types of MAC for general-purpose digital signal processing has been proposed by Elguibaly [19]. It is an architecture in which accumulation has been combined with the carry save adder (CSA) tree that compresses partial products. In the architecture proposed in [12], the critical path was reduced by eliminating the adder for accumulation and decreasing the number of input bits in the final adder. While it has a better performance because of the reduced

critical path compared to the previous MAC architectures

	Resource	Input Data Set	w/o PGC	w/ PGC	w/ BIND & PGC (n_D=2)	power reduction w.r.t basic MAC(%)
			nJ	*nJ*	*nJ*	
fir11	+1 *3 (shared)	*speech1*	54.2	44.6	39.1 (37.4)	**31.1**
		speech2	55.0	43.0	37.7 (34.6)	**37.1**
		normrand	54.7	43.4	37.4 (37.4)	**31.8**
	+10, *11 (fully parallel)	*speech1*	26.1	16.8	N/A	**35.5**
		speech2	26.4	14.8	N/A	**43.9**
		normrand	25.8	15.3	N/A	**40.5**
Wavelet	+2 *4	*speech1*	82.8	76.2	75.6 (71.5)	**13.6**
		speech2	83.4	69.9	67.0 (55.9)	**20.4**
		normrand	83.7	77.5	74.5 (69.1)	**17.4**
nc	+2 *4	*speech1*	255.3	233.6	233.6 (230.2)	**9.8**
		speech2	263.7	229.1	229.1 (235.1)	**13.1**
lattice	+1 *2	*speech1*	46.4	39.5	39.5 (37.9)	**18.3**
		speech2	38.1	31.0	31.0 (24.0)	**34.0**
		normrand	47.7	42.4	42.4 (40.7)	**11.6**
Average						**23.87**

Table.1: Comparison of power consumption in multipliers

Table 1 gives the brief history of power reduction by using PGC. There is a need to improve the output rate due to the use of the final adder results for accumulation. Architecture to merge the adder block to the accumulator register in the MAC operator was proposed in [19] to provide the possibility of using two separate N/2-bit adders instead of one -bit adder to accumulate the N–bit MAC results.

Design	Feature	Tech.	Power (mW)	Delay (ns)
Huang [21]	(1) 32b × 32b	0.18-μm	(1) 19.65 for Djpeg (2) 40.65 for Random data @ 100MHz	7.25
Liao [22]	(1) Coprocessor (2) SIMD (3) 32b MAC	0.18-μm	900@1.6V, 800MHz	1.25
Wang [23]	(1) 32b × 32b (2) Fixed-width	0.35-μm /3.3V	79.86	14.01
Chen[24]	(1) 16b × 16b	0.25-μm	17.30 for *Normal distribution* inputs	8.3
Lee [25]	Scalable length of 4b, 8b, 16b	0.13-μm /1.2V	1.04@100MHz for random data	N.A.
Hsu [26]	(1) 16b × 16b (2) Sleep mode (3) Duel V_T	90nm	(1) 9@1.3V, 1GHz (1) 7.9×10^{-2} @50MHz, 0.57V	1
Spst	(1) 16b × 16b (2) Versatile functions	0.18-μm	(1) 4.2@100MHz,1.8V (2) 1.25×10^{-1}@25MHz, 0.7V for H.264 IQ (3) 6.3@125MHz,1.8V for *Normal distribution* inputs	8

Table.2: performance comparison of existing multipliers and spst.

Recently, Zicari proposed an architecture that took a merging technique to fully utilize the 4–2 compressor [20]. It also took this compressor as the basic building blocks for multiplication circuit. There are several techniques in reducing power using techniques like increasing the correlation of input data, Partially Guarded Computation and Spurious power suppression technique. But each of these techniques have their advantages and drawbacks. But depending on our application requirement we should choose our suitable technique to reduce power.

IV.CONCLUSION:

By using the PGC technique we can reduce power consumption in an array multiplier by about 10 to 44%. This method can effectively reduce power consumption even after minimization of power by using high-level power minimization technique. However, the PGC technique cannot gain any power reduction when applied on adders because of the overhead-augmented circuitry. Equipping the SPST can save 24% power dissipation at the cost of only 15% area increment, which is a valuable trade-off especially for modem CMOS technologies. We can still reduce the power by using different adders in accumulation stage and innovative multiplication algorithms. Reduction of power is possible by using more than 2 stages of SPST functional block with some increase in area.

REFERENCES:

[1] J. J. F. Cavanagh, Digital Computer Arithmetic. New York: McGraw-Hill, 1984.

[2] Information Technology-Coding of Moving Picture and AssociatedAutio, MPEG-2 Draft International Standard, ISO/IEC 13818-1, 2, 3,1994.

[3] A. Raghunathan and N. K. Jha, "Behavioral synthesis for low power," Proceedings of International Conference on ComputerDesign, pp. 318-322, Oct. 1994.

[4] A. Raghunathan, S. Dey, N. K. Jha, "Controller re-specification to minimize switching activity in controller/data path circuits," Proceedings of International Symposium on Low Power Electronics and Design, pp. 301-304, Aug. 1996.

[5] E. Musoll and J. Cortadella, "High-level synthesis techniques for reducing the activity of functional units," Proceedings of International Symposium on Low Power Design, pp. 99-104, Nov. 1995.

[6] L. Benini, P. Vuillod, G. D. Micheli, and C. Coelho, "Synthesis of low power selectively-clock systems from high-level specification," Proceedings of International Symposium on System Synthesis, pp. 57-63, Nov. 1996.

[7] D. Kim and K. Choi, "Power conscious high level synthesis using loop folding," Proceedings of Design Automation Conference, pp.441-445, 1997.

[8] D. Shin and K. Choi, "Lower power high level synthesis by increasing data correlation," Proceedings of International Symposium on Low Power Electronics and Design, Aug. pp. 441-445, Aug. 1997.

[9] R. Mehra, L. M. Guerra, and J. Rabaey, "Low-power architectural synthesis and the impact of exploiting locality," Journal of VLSI Signal Processing, 1996.

[10] A. Dasgupta and R. Karri, "Simultaneous scheduling and binding for power minimization during micro architecture synthesis," Proceedings of International Symposium on Low Power Design, 1995.

[11] V. Tiwari, S. Malik, and P. Ashar, "Guarded Evaluation: Pushing Power Management to Logic Synthesis/Design," IEEE Trans. on Computer Aided Design of Integrated Circuits and Systems, vol. 17, no. 10, pp.1051-1060, Oct. 1998.

[12] Young-Ho Seo and Dong-Wook Kim,"A new VLSI architecture of parallel multiplier-accumulator based on radix-2 modified Booth algorithm", in IEEE Trans. on Very Large Scale Integration (VLSI) Systems, vol. 18, no. 2, pp.201-208, February 2010.

[13] J. Choi, J. Jeon, and K. Choi, "Power minimization of functional units by partially guarded computation," in Proc. IEEE Int. Symp. Low power Electron. Des., 2000, pp. 131–136.

[14] O. Chen, R. Sheen, and S. Wang, "A low-power adder operating on effective dynamic data ranges," IEEE Trans. Very Large Scale Integr (VLSI) Syst., vol. 10, no. 4, pp. 435–453, Aug. 2002.

[15] O. Chen, S.Wang, and Y. W.Wu, "Minimization of switching activities of partial products for designing low-power

multipliers," IEEE Trans.Very Large Scale Integr. (VLSI) Syst., vol. 11, no. 3, pp. 418–433, Jun.2003.
[16] L. Benini, G. D. Micheli, A. Macii, E. Macii, M. Poncino, and R. Scarsi,"Glitch power minimization by selective gate freezing," IEEE Trans.Very Large Scale Integr. (VLSI) Syst., vol. 8, no. 3, pp. 287–298, Jun.2000.
[17] S. Henzler, G. Georgakos, J. Berthold, and D. Schmitt-Landsiedel,"Fast power-efficient circuit-block switch-off scheme," Electron. Lett.,vol. 40, no. 2, pp. 103–104, Jan. 2004.
[18] T. Xanthopoulos and A. P. Chandrakasan, "A low-power DCT core using adaptive bit width and arithmetic activity exploiting signal correlations and quantization," IEEE J. Solid-State Circuits, vol. 35, no. 5,pp. 740–750, May 2000.
[19] F. Elguibaly, "A fast parallel multiplier–accumulator using the modified Booth algorithm", IEEE Trans. Circuits Syst., vol. 27, no. 9, pp.902–908, September 2000.
[20] A. R. Cooper, "Parallel architecture modified Booth multiplier",Proc.Inst.Electr.Eng.G, vol.135, pp.125–128, 1988.
[21] Z. Huang and M. D. Ercegovac, "High-performance low-power left-to right array multiplier design," IEEE Trans. Computers., vol. 54, no. 3, pp.272–283, Mar. 2005.
[22] Y. Liao and D. B. Roberts, "A high-performance and low-power 32-bit multiply-accumulate unit with single-instruction–multiple-data (SIMD) feature," IEEE J. Solid-State Circuits, vol. 37, no. 7, pp.926–931, Jul. 2002.
[23] J. S.Wang, C. N. Kuo, and T. H. Yang, "Low-power fixed-width array multipliers," in Proc. IEEE Symp. Low Power Electron. Des., Aug.9–11, 2004, pp. 307–312.
[24] O. Chen, S.Wang, and Y. W.Wu, "Minimization of switching activities of partial products for designing low-power multipliers," IEEE Trans.Very Large Scale Integr. (VLSI) Syst., vol. 11, no. 3, pp. 418–433, Jun.2003.
[25] H. Lee, "A power-aware scalable pipelined Booth multiplier," in Proc.IEEE Int. SOC Conf., Sep. 2004, pp. 123–126.
[27] S. K. Hsu, S. K. Mathew, M. A. Anders, B. R. Zeydel, V. G. Oklobdzija,R. K. Krishnamurthy, and S. Y. Borkar, "A 110 GOPS/W 16-bit multiplier and reconfigurable PLA loop in 90-nm CMOS," *IEEE J.Solid-State Circuits*, vol. 41, no. 1, pp. 256–264, Jan. 2006.

Author's Profile:

1.K.N Varaprasad is doing M.Tech in Electronics Engineering from Veermata Jijabai Technological Institute and his area of interests are VLSI and Nanotechnology.

2. Dr.Nisha Sarwade is professor in Department of Electrical Engineering in Veermata Jijabai Technological Institute and her area of interests are VLSI and Nanotechnology.

3. Ch.M Krishna is doing M.Tech in VLSI and Embedded Systems from College of Engineering Pune and his area of interests are VLSI and Embedded Systems.