[[1]](#footnote-2)

Low power cache using four transistor SRAM

**kirti S Pande(1), Sai Sundeep I(2), Revanth M(2) and Nikhilanjali J(2)**

**(1) Asst. Professor-ECE Department, Amrita University, Bangalore.**

**(2) B-tech-ECE Department, Amrita University,Bangalore.**

**kirtispande@yahoo.co.in, sundeep.innamuri@gmail.com, marrirevanth.121@gmail.com, nikhila.135@gmail.com**

*Abstract—* **-***This paper presents a CMOS four-transistor(4T) SRAM cell for low power embedded SRAM applications as well as for stand-alone SRAM applications. The new cell consumes % less power than a conventional six transistor(6T) SRAM cell but delay is % more. This delay in 4T SRAM cell must be decreased such that 6T SRAMs can be replaced by them for low power cache. Cache memories of size 128 bits are designed using both 6T and 4T SRAM to compare the power and delay parameters. Divided bit line technique is used to decrease the delay in 4T SRAM cache. this technique provides 40% delay reduction and also 16.6% reduction in power. Microwind2 simulation in standard 45nm CMOS technology confirms all results from this paper.*

*Index Terms*— Average delay, average dynamic power, leakage current, SRAM.

# INTRODUCTION

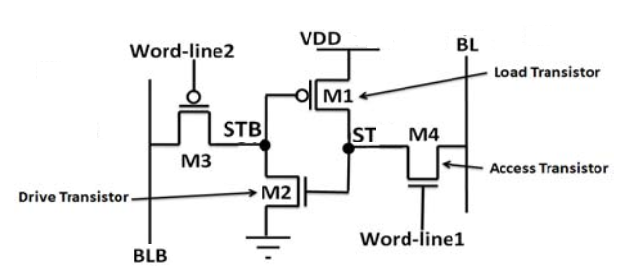
SoCs integrate multiple functions on a single silicon die. As process geometries have scaled, designs which use more and more of the additional silicon real estate available on chips to integrate embedded memories evolved. These embedded memories allow for significantly better system performance and lower power compared to a solution where off-chip memories are used. Most current designs have over 50% of their area used by embedded memories and these memories account for 50-70% of the total SoC power dissipation. Clearly, any attempt to reduce SoC power is incomplete if it does not attempt to reduce the power consumed by the embedded memories in the design.

Embedded cache memories are implemented using SRAM cells. SRAM is a type of semiconductor memory that uses stable latching to store each bit. A conventional SRAM has a 6T design. It is volatile in the conventional sense that data is eventually lost when the memory is not powered. So designing the SRAMs which consume less power becomes crucial. The cache is a smaller, faster memory which stores copies of the data from the most frequently used main memory locations. As long as most memory accesses are cached memory locations, the accesses time of memory will be closer to the accesses time of cache than to the access time of main memory. Hence improving the speed of cache becomes important.

The objective of this paper is to implement a cache memory using 4T SRAM and compare it with conventional 6T SRAM cache for parameters like power and speed of operation, then implement divided bit-line technique on 4T SRAM cache to get an optimized design with low power and lesser delay. All the modules are designed in standard 45nm technology for this technology node is 0.5V and is 0.2V.

# 4T SRAM

4T SRAM cell is shown in Fig(2). It uses two word lines.



Fig(2) 4T SRAM Cell

## Write and read operation

In 4T SRAM bit line(BL) is pre-charged to GND and bit bar line(BLB) is pre-charged to . Here .

When a write operation is issued the memory cell will go through the following steps.

1) Bit-line driving: For a write data placed on BL and then word-line1 asserted to , but voltages on word-line2 and BLB maintained at idle mode ().

2) Cell flipping: this step includes two states as follows.

(a) data is one: in this state, ST node pulled up to by NMOS access transistor, and therefore the drive transistor will be ON , and STB node will be pulled down to GND, thus load transistor will be ON and STB node pulled up to VDD.

(b) data is zero: in this state, ST node pulled down to GND by NMOS access transistor, and therefore the drive transistor will be OFF, and STB node will be floated and then pulled up to voltage of BLB () by leakage current (most of this current is sub-threshold current) of PMOS access transistor, and thus load transistor will be OFF.

3) Idle mode: At the end of write operation, cell will go to idle mode and word-line1 and BL asserted to and GND respectively.

When a read operation is issued the memory cell will go through the following steps.

1) Bitbar-line Pre-charging: For a read, BLB pre-charged to and then floated. Since, in idle mode BL maintained at . This step didn’t include any dynamic energy consumption.

2) Word-line activation: in this step word-line1 asserted to GND and two states can be considered:

(a) Voltage of ST node is low: when voltage of ST node is low the voltage of BL and ST node equalized. (we refer to voltage of BL in this state as ). Since in this state there is very small different between BL and ST node dynamic energy consumption is very small.

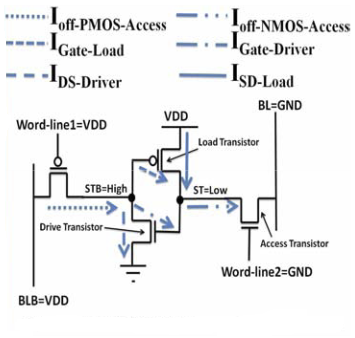
(b) Voltage of ST node is high: when voltage of ST node is high, the voltage of BL pulled up to high voltage by NMOS access transistor. (we refer to voltage of BL in this state as ).

3) Sensing: the sense amplifier is turned on to read data on BL.

4) Idle mode: At the end of read operation, cell will go to idle mode and word-line1 and BL asserted to .

## Cell Design Concepts

When “1” is stored in the cell, load and drive transistors are ON and there is a positive feedback between ST and STB node. Therefore STB node is pulled to GND by drive transistor and ST node is pulled to VDD by load transistor.

**Fig (2.2) 4T SRAM Cell in idle mode when “0” is stored**

When “0” is stored in the cell both load and drive transistors are OFF and for data retention without refresh cycle, the following conditions must be satisfied.

IOff-NMOS-access > ISD-Load + IGate-Driver + IGate-Load **Eq (2.1)**

IOff-PMOS-access > IDS-Load + IGate-Driver + IGate-Load **Eq (2.2)**

Fig (2.2) shows leakage current of cell during idle mode for data retention when “0” is stored in the cell. For satisfying the above condition when “0” is stored in the cell we use leakage current of access transistors, especially sub-threshold current of access transistors (IOff-NMOS-access and IOff-PMOS-access ).

To achieve this we can use high threshold voltage for load and drive transistors to reduce sub-threshold currents of these transistors. Low threshold voltage can be used for access transistors; hence leakage current of access transistors will be greater than leakage currents of load and drive transistors. With this threshold voltage assignments above conditions can be satisfied and “0” can be stored in a cell successfully. Leakage currents also depend on the width of the transistors. As width of a transistor increases, leakage currents also increase. As the leakage currents of access transistors should be greater than leakage currents of load and drive transistors, the width of access transistors should be greater than the width of load and drive transistors.

From the equations 2.1 and 2.2 if the leeakge currents on the right side of the equations are large IOff –access In the both equations will be high. If say 1 is written ,which is a stable state in the cell, stb node will be pulled to node ground due to 0 at st,but IOff-PMOS-access is high and tries to pull to vdd,in this process a high power consumption takes place.so there is a necessacity to reduce the leekage currents on the right side of the equation to reduce IOff –access current.

To decrease the power consumption problem VDD is reduced to 0.45V and VSS is increased to 0.05V. Due to this bit-line swing reduces resulting in less power consumption. Write and read operations are performed using BL.

Sub-threshold leakage currents come into picture when VGS < VT. When “0” is stored in the cell, load and drive transistors are OFF. when load transistor is in OFF state, its VGS value is equal to 0V which results in flow of IOFF current. In this model since we have increased VSS value to 0.05V, when load transistor is in OFF state, its VGS value becomes -0.05V which reduces the sub threshold leakage current contribution of load transistor. Here, VDD value is decreased to 0.45V so when drive transistor in in OFF state its VSG value decreases, which reduces the sub threshold leakage current contribution of drive transistor. To reduce the leakage currents in sub-threshold region we have to reduce Vth which can be achieved by increasing the doping. In Microwind, since the doping concentrations cannot be changed, we consider body biasing. The change in body biasing affects Vth which is called “Body effect”.

**Body Effect equation is:**

V_{TN} = V_{TO} + \gamma ( \sqrt{ | {V_{SB} + 2\phi_{F} | } } - \sqrt{ | 2\phi_{F} | } )

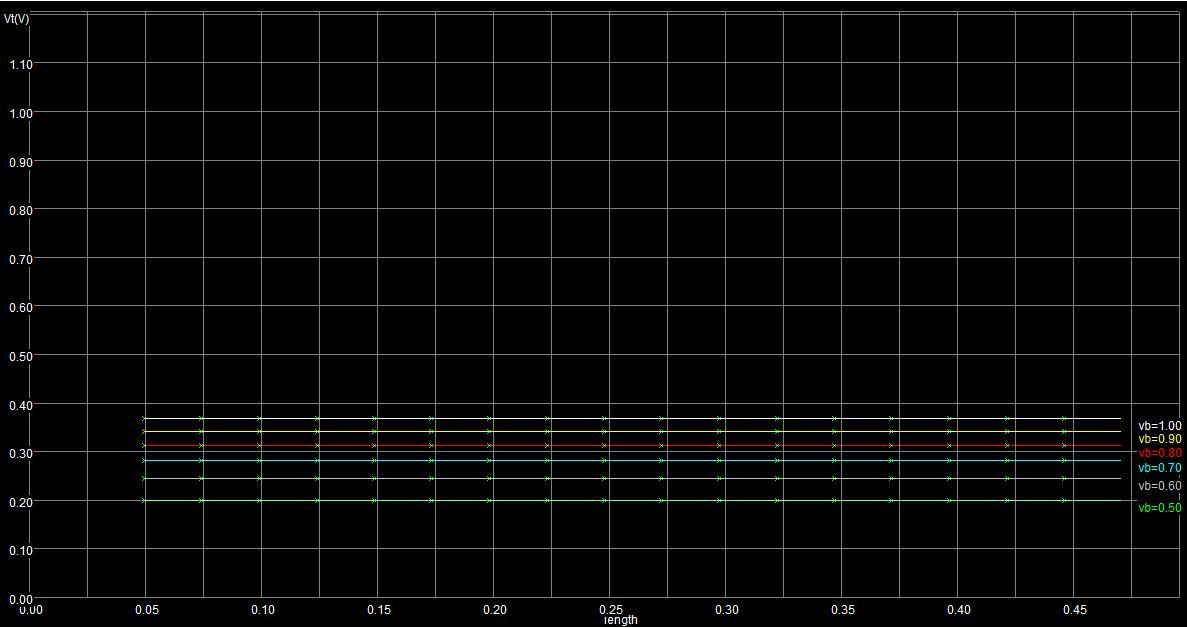
Where, VTN is the threshold voltage when substrate bias is present,VSB is the source-to-body substrate bias,ØF is the surface potential ,VTO is the threshold voltage for zero substrate bias.

\gamma = (t_{ox}/\epsilon_{ox})\sqrt{2q\epsilon_{si}N_A} is the body effect parameter,

t_{ox} is oxide thickness, \epsilon_{ox} is oxide permittivity,\epsilon_{si} is the permittivity of silicon, NA is the doping concentration and q is the charge of an electron.

**From the Changes in VTH with change in N-well biasing:**

From the graphs shown below we can observe the changes in VTH with change in Nwell biasing as follows.



**Load transistor body biasing to read “1” delay trade off:**

To reduce the leakage current of load transistor we have increased VSB in order to increase VTH. But when the threshold voltage is increased, current IDS in saturation region decreases.

Current equation in saturation region is given by

I­ds =

From the above equation we can observe then when Vt is increased current is decreased with power of two. If the bit line capacitance is large, more current is required to pull the word line to the value that is stored inside the cell. Since the current supplied by the load transistor is less more time is required to provide sufficient charge to overcome the capacitance.

But if the threshold voltage is decreased, it results in increase of leakage currents in turn increasing the delay in writing “0” and increasing the power consumed in write “0” and hold “0” condition.

So there exists a tradeoff between the body biasing voltage and delay during read“1”.

**Values used:** VDD= 0.45V VSS = 0.05V

VTN0 = 0.2V VTP0 = 0.2V

VTN = 0.16V VTP,ACCESS = 0.24V VTP,DRIVE = 0.28V

VNWELL, ACCESS = 0.6V

VNWELL, DRIVE = 0.7V

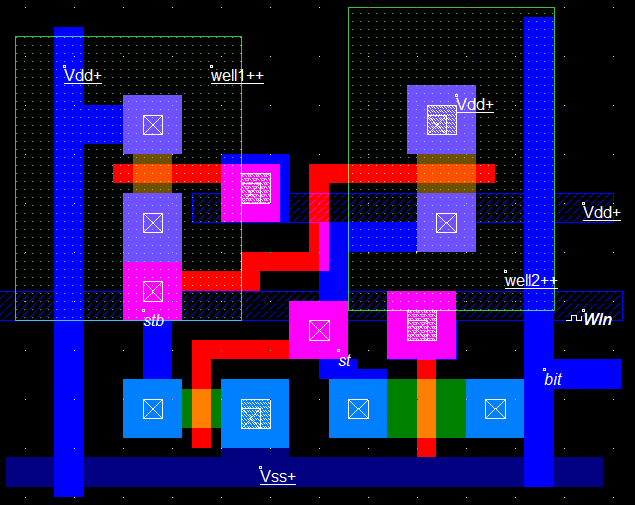
VWLN = 0.5 and 0V (ON and OFF)

VWLP = 0.3V

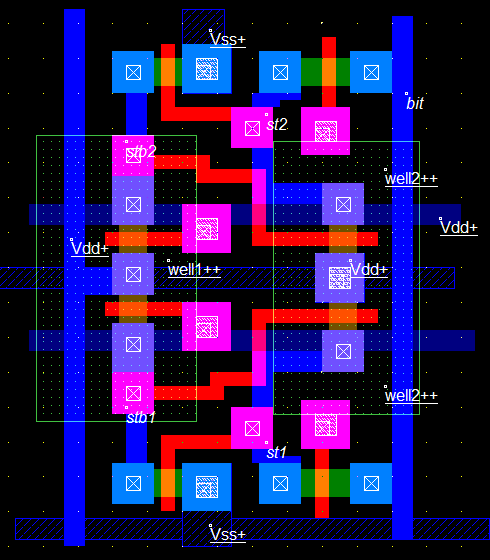
From the layout,

WNMOS,ACCESS= 6 lambda WLOAD = 6 lambda

WPMOS,ACCESS= 4 lambda WDRIVE = 4 lambda

****

Due to the presence of two N-wells and the distance between the N-wells must be greater than 11 lambda, the area increases than that of the conventional model. So to reduce the area of the cell in this model the drains of the PFET load and PFET driver are combined. This actually forms the structural unit of the array that will be described in chapter 6. If this type of combining two drains is done without using separate N-wells (changing Vt by some other means) the area of new cells formed will be approximately 50% of the 6tcell,which is a lot reduction in area.

****

# DIVIDED BITLINE TECHNIQUE

1. *Concept*

Dynamic power consumption in SRAM can be given by[]

where, is an external supply voltage, is the total current,is the effective active current is an internal supply voltage, is the total capacitance of the peripheral circuits, is the total static current, m is the number of columns and f is the operation frequency. To reduce the total power consumption, active current should be reduced as it dominates the total current. Active current is the current that flows during word line activation i.e., during charging or discharging of bit-line capacitance. This active current is directly proportional to bit-line capacitance. Divided word line approaches reduce the by reducing the value of m, i.e., the number of columns activated during a read or write operation.

The bit-line capacitance is mainly composed of the drain capacitance of the pass transistors of the SRAM cell and metal capacitance of bit-line. To reduce this capacitance, drain capacitance and metal capacitance should be reduced. Bit-line capacitance can be reduced by proposed divide bit line approach, where the number of transistors connected to the bit-line is reduced by combining two or more SRAM cells.

optimal number of grouping that must be done to get maximum reduction in delay can be calculated as follows. Let us take NFET drain capacitence= Cd, metal capacitance= Cm, metal resistance= Rm,drain resistance= Rd, Toatal no of cells=N

In the above observations the values taken are for one cell

Therfore In a single bitline approach total delay is due to all drain and metal connect.this can be written as

Td=N\*Cm\*Rm + N\*Cd\*Rd (3)

Now consider a divided bitline with N rows grouped with 8 cells each. Here delay is due to global bit line and one sub bitline.

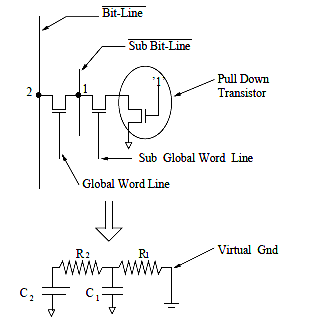


Fig (3) RC model

From the above figure total global bit-line delay can be written as sum of delays due to N metal pieces and N/m drain capacitance of pass transistors(N/m because N rows of m bits each will have N/m groups and n/m pass transistors).

T2=N\*Cm\*Rm + (N/m)\*Cd\*Rd (4)

ub bit line delay can be written as sum of delays of (m+2)drains (m drains in a m cell group plus one pass transistor drain plus one precharge drain ) and m metal lines which ca be written as

T1= (M\*Cm\*Rm) + (M+2)\*Rd\*Cd (5)

Now total delay can be considered as t1+t2

Td=M\*Cm\*Rm+(M+2)\*Rd\*Cd+N\*Cm\*Rm+(N/m)\*Rd\*Cd (6)

Now optimize the value of m for which the delay is minimum. To find the m for minimum value of delay we differentiate td with respect to m and equate it to zero.

Differentiating Eq (7.7) w.r.t to m

=

= + + +

=Rm\*Cm+ Rd\*Cd - (1/M2)N\*Rd\*Cd (7)

Now equating Eq (7.8) to zero, we get

Rm\*Cm + Rd\*Cd - (1/m2)N\*Rd\*Cd = 0

Let us take NFET drain capacitence = Cd=.07fF,

metal capacitance =Cm=.085fF

metal resistance=Rm=2.5 0hm

drain resistance=Rd=35 ohm

substituting above values

M=

For N=128

m=

Substituting m=10 and N=128 in (6)

# SIMULATION RESULTS

## Figures and Tables

Because the final formatting of your paper is limited in scale, you need to position figures and tables at the top and bottom of each column. Large figures and tables may span both columns. Place figure captions below the figures; place table titles above the tables. If your figure has two parts, include the labels “(a)” and “(b)” as part of the artwork. Please verify that the figures and tables you mention in the text actually exist. **Do not put borders around the outside of your figures.** Use the abbreviation “Fig.” even at the beginning of a sentence. Do not abbreviate “Table.” Tables are numbered with Roman numerals.

Include a note with your final paper indicating that you request color printing. **Do not use color unless it is necessary for the proper interpretation of your figures.** There is an additional charge for color printing.

Figure axis labels are often a source of confusion. Use words rather than symbols. As an example, write the quantity “Magnetization,” or “Magnetization *M*,” not just “*M*.” Put units in parentheses. Do not label axes only with units. As in Fig. 1, for example, write “Magnetization (A/m)” or “Magnetization (Am−1),” not just “A/m.” Do not label axes with a ratio of quantities and units. For example, write “Temperature (K),” not “Temperature/K.”

Multipliers can be especially confusing. Write “Magnetization (kA/m)” or “Magnetization (103 A/m).” Do not write “Magnetization (A/m) × 1000” because the reader would not know whether the top axis label in Fig. 1 meant 16000 A/m or 0.016 A/m. Figure labels should be legible, approximately 8 to 12 point type.

## References

Number citations consecutively in square brackets [1]. The sentence punctuation follows the brackets [2]. Multiple references [2], [3] are each numbered with separate brackets [1]–[3]. When citing a section in a book, please give the relevant page numbers [2]. In sentences, refer simply to the reference number, as in [3]. Do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] shows ... .” Number footnotes separately in superscripts (Insert | Footnote).[[2]](#footnote-3) Place the actual footnote at the bottom of the column in which it is cited; do not put footnotes in the reference list (endnotes). Use letters for table footnotes (see Table I).

Please note that the references at the end of this document are in the preferred referencing style. Give all authors’ names; do not use “*et al*.” unless there are six authors or more. Use a space after authors' initials. Papers that have not been published should be cited as “unpublished” [4]. Papers that have been submitted for publication should be cited as “submitted for publication” [5]. Papers that have been accepted for publication, but not yet specified for an issue should be cited as “to be published” [6]. Please give affiliations and addresses for private communications [7].

## Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have already been defined in the abstract. Abbreviations such as SI, ac, and dc do not have to be defined. Abbreviations that incorporate periods should not have spaces: write “C.N.R.S.,” not “C. N. R. S.” Do not use abbreviations in the title unless they are unavoidable (for example, “INTERNATIONAL JOURNAL OF ENGINEERING AND INNOVATIVE TECHNOLOGY” in the title of this article).

## Equations

Number equations consecutively with equation numbers in parentheses flush with the right margin, as in (1). First use the equation editor to create the equation. Then select the “Equation” markup style. Press the tab key and write the equation number in parentheses. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Use parentheses to avoid ambiguities in denominators. Punctuate equations when they are part of a sentence, as in

 (1)

Be sure that the symbols in your equation have been defined before the equation appears or immediately following. Italicize symbols (*T* might refer to temperature, but T is the unit tesla). Refer to “(1),” not “Eq. (1)” or “equation (1),” except at the beginning of a sentence: “Equation (1) is ... .”

## Other Recommendations

Use one space after periods and colons. Hyphenate complex modifiers: “zero-field-cooled magnetization.” Avoid dangling participles, such as, “Using (1), the potential was calculated.” [It is not clear who or what used (1).] Write instead, “The potential was calculated by using (1),” or “Using (1), we calculated the potential.”

Use a zero before decimal points: “0.25,” not “.25.” Use “cm3,” not “cc.” Indicate sample dimensions as “0.1 cm × 0.2 cm,” not “0.1 × 0.2 cm2.” The abbreviation for “seconds” is “s,” not “sec.” Do not mix complete spellings and abbreviations of units: use “Wb/m2” or “webers per square meter,” not “webers/m2.” When expressing a range of values, write “7 to 9” or “7-9,” not “7~9.”

A parenthetical statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.) In American English, periods and commas are within quotation marks, like “this period.” Other punctuation is “outside”! Avoid contractions; for example, write “do not” instead of “don’t.” The serial comma is preferred: “A, B, and C” instead of “A, B and C.”

If you wish, you may write in the first person singular or plural and use the active voice (“I observed that ...” or “We observed that ...” instead of “It was observed that ...”). Remember to check spelling. If your native language is not English, please get a native English-speaking colleague to proofread your paper.

# Some Common Mistakes

The word “data” is plural, not singular. The subscript for the permeability of vacuum µ0 is zero, not a lowercase letter “o.” The term for residual magnetization is “remanence”; the adjective is “remanent”; do not write “remnance” or “remnant.” Use the word “micrometer” instead of “micron.” A graph within a graph is an “inset,” not an “insert.” The word “alternatively” is preferred to the word “alternately” (unless you really mean something that alternates). Use the word “whereas” instead of “while” (unless you are referring to simultaneous events). Do not use the word “essentially” to mean “approximately” or “effectively.” Do not use the word “issue” as a euphemism for “problem.” When compositions are not specified, separate chemical symbols by en-dashes; for example, “NiMn” indicates the intermetallic compound Ni0.5Mn0.5 whereas “Ni–Mn” indicates an alloy of some composition NixMn1-x.

Be aware of the different meanings of the homophones “affect” (usually a verb) and “effect” (usually a noun), “complement” and “compliment,” “discreet” and “discrete,” “principal” (e.g., “principal investigator”) and “principle” (e.g., “principle of measurement”). Do not confuse “imply” and “infer.”

Prefixes such as “non,” “sub,” “micro,” “multi,” and “"ultra” are not independent words; they should be joined to the words they modify, usually without a hyphen. There is no period after the “et” in the Latin abbreviation “*et al.*” (it is also italicized). The abbreviation “i.e.,” means “that is,” and the abbreviation “e.g.,” means “for example” (these abbreviations are not italicized).

An excellent style manual and source of information for science writers is [9].

# Editorial Policy

The submitting author is responsible for obtaining agreement of all coauthors and any consent required from sponsors before submitting a paper. It is the obligation of the authors to cite relevant prior work.

Authors of rejected papers may revise and resubmit them to the journal again.

# Publication Principles

The contents of the journal are peer-reviewed and archival. The journal INTERNATIONAL JOURNAL OF ENGINEERING AND INNOVATIVE TECHNOLOGY (IJEIT) publishes scholarly articles of archival value as well as tutorial expositions and critical reviews of classical subjects and topics of current interest.

Authors should consider the following points:

1. Technical papers submitted for publication must advance the state of knowledge and must cite relevant prior work.
2. The length of a submitted paper should be commensurate with the importance, or appropriate to the complexity, of the work. For example, an obvious extension of previously published work might not be appropriate for publication or might be adequately treated in just a few pages.
3. Authors must convince both peer reviewers and the editors of the scientific and technical merit of a paper; the standards of proof are higher when extraordinary or unexpected results are reported.
4. Because replication is required for scientific progress, papers submitted for publication must provide sufficient information to allow readers to perform similar experiments or calculations and use the reported results. Although not everything need be disclosed, a paper must contain new, useable, and fully described information. For example, a specimen's chemical composition need not be reported if the main purpose of a paper is to introduce a new measurement technique. Authors should expect to be challenged by reviewers if the results are not supported by adequate data and critical details.

# Conclusion

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

singular heading even if you have many acknowledgments. Avoid expressions such as “One of us (S.B.A.) would like to thank ... .” Instead, write “F. A. Author thanks ... .” **Sponsor and financial support acknowledgments are placed in the unnumbered footnote on the first page**.

REFERENCES

1. S. Chen, B. Mulgrew, and P. M. Grant, “A clustering technique for digital communications channel equalization using radial basis function networks,” IEEE Trans. on Neural Networks, vol. 4, pp. 570-578, July 1993.
2. J. U. Duncombe, “Infrared navigation—Part I: An assessment of feasibility,” IEEE Trans. Electron Devices, vol. ED-11, pp. 34-39, Jan. 1959.
3. C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, “Rotation, scale, and translation resilient public watermarking for images,” IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.

1. [↑](#footnote-ref-2)
2. [↑](#footnote-ref-3)