IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-11 E-ISSN: 2347-2693

DSS Query Optimization and Effect of Input Output / Communication Cost Metrics

M. Sharma[1*], G. Singh[2], R. Singh[2]

[1*]Dept. of CSA, DAV University, Jalandhar, India
[2]Dept. of CS, Guru Nanak Dev University, Amritsar, India

Corresponding Author: manik_sharma25@yahoo.com

Available online at: www.ijcseonline.org

Received: 22/Aug/2017, Revised: 018/Sep/2017, Accepted: 29/Sep/2017, Published: 30/Nov/2017

Abstract—Decision Support System (DSS) query is an important type of distributed query. It plays an imperious role in decision making practise. However, it ingest loads of Input Output (I/O), processing and communication assets. Here, a 3-Join DSS query has been optimized using entropy and restricted chromosome based DSS query optimizer (ERC_QO). A study is carried out to inspect the consequences of varying the ratio of I/O and communication costs over Total Costs (total system resources). It is perceived that by plummeting the I/O to communication costs ratio, the communication costs can be more commendably optimized. For a 3-Join DSS query, the communication costs have been reduced by 90% approximately. Moreover, the Total Costs of 3-Join DSS query is abridged by 2%.

Keywords— DSS query, Query Optimization, I/O costs, Communication Costs etc.

I. INTRODUCTION

A DSS query is convoluted and data intensive query. The execution of DSS query demands momentous amount of I/O, processing and communication resources. The sum of I/O, processing and communication costs represent Total Costs of the query. Total Costs represents the amount of different resources required to execute the query. Due to convoluted nature and significant requirement of different resources, DSS query should be optimized before its execution [1][2]. Query optimization is a method of selecting the preeminent query execution plot as per optimization function.The query can be optimized by using diverse deterministic and stochastic techniques. Moreover, a query can be optimized by changing the order of sub operations, or by altering the location where sub operations would be executed. In addition to, one can optimize the query by executing it with different algorithmic approaches [3][4][5]. The distributed queries can be optimized by abating either the *Total Costs* or *Response Time* of a query. *Total Costs* are optimized for increasing the throughput of the system and *Response Time* is optimized to speed up the execution process of a query. In this research work, the focus is given on increasing the throughput of DSS query optimizer [6][7].

To optimize and analyse a 3-Join DSS query, a hybrid model called entropy and restricted chromosome based DSS query optimizer has been used [8][9][10]. The query optimizer has developed using the features of information theory and GA. The optimizer is designed to abate the resource ingestion. The use of GA assists in finding optimal results in minimum time. The growth of the chromosome is restricted to generate better generations. The major objective of this study is to determine the effect of varying I/O to communication costs over Total Costs and Communication Costs of a distributed DSS query.

Design and statistics of 3-join DSS query is represented in second section. Third and fourth section explained the experimental setup and design of entropy and restricted chromosome based DSS query optimizer (EGA_QO). Fifth and sixth section elucidated the assumption and the effect of varying input output costs respectively. Conclusion is framed in seventh section. Finally, references are mentioned in eight section of this manuscript.

II. DESIGN AND STATISTICS OF 3-JOIN DSS QUERY

Initially, a 3-Join DSS query has been considered for analysis. The statistics of the query are given below [8][9]:

Total Number of Operations	: 11
Total Number of Intermediate Fragments	: 14
Number of Selection Operations	: 04
Number of Projection Operations	: 04
Number of Joins Operations	: 03
Number of Base Relations	: 04
Number of Sites	: 04

Figure 1 represents the tree diagram for the 3-Join DSS query. Each node and edge represents the sub operation and the fragment of the query respectively. Sub operations, fragments and base relations are represented as *On, Fn and Bn* respectively. Here,

O1, O2, O3, O4: Selection operations

O5, O6, O7, O8 : projection operations

O9, O10, O11 : Join operations

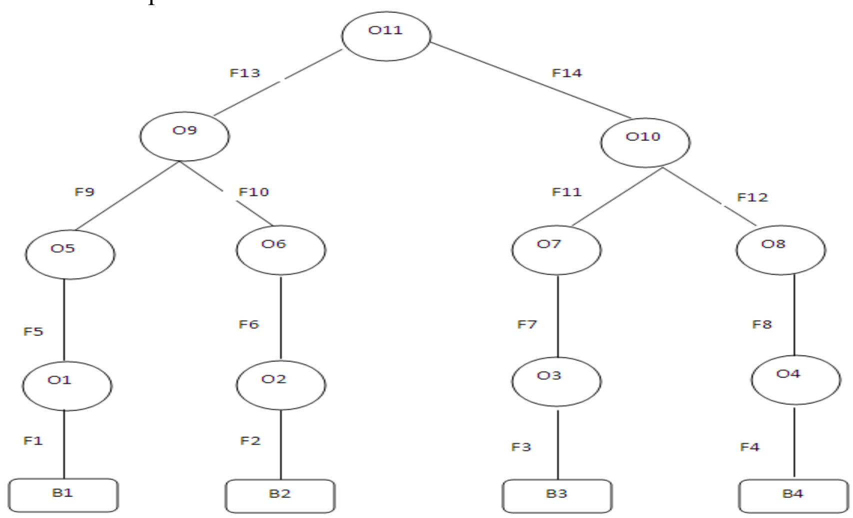

Figure 1: A 3-Join DSS Query

III. Environmental setup

ERC_QO is a hybrid DSS query optimizer and is developed by combining the features of both restricted chromosome and entropy. The working of ERC_QO is based upon different parameters like data allocation, costs coefficients (I/O, Processing and Communication), restriction in chromosome design, entropy, number of sites, operating site etc. The details of ERC_QO can be obtained from article mentioned in reference number [8][9].

Here, a 3-Join DSS query is supposed to be executed on a distributed system consisting of four different sites. The different costs coefficients as mentioned below are formulated using the cost model of Ozsu. Moreover, the ratio of 'Input Output' and 'Communication Costs' coefficients has also been setup as per Ozsu and Valduriez cost function specification[3]. The ratio was fixed as 1:1.6. As per distributed query 'Cost Model', the costs coefficients of *Input Output, Processing, Communication and Data allocation* are as given below:

Input Output Costs Coefficients 11 10 12 14	**Processing Costs Coefficients** 1.1 1 1.2 1.4
Communication Costs Coefficients (A Matrix of order 4 X 4) 0 16 19 22 16 0 19 22 19 19 0 22 22 22 22 0	**Data Allocation (Matrix of order 4 X 4)** 1 1 0 0 0 1 1 0 0 0 1 1 1 1 0 0

ERC_QO provides the design of chromosome with their corresponding I/O, processing, communication and total costs. The chromosome represented the query execution plan of a distributed DSS query. Each numeral value of a chromosome identified the location where a sub operation would be executed. First four digits represented 'Selection', followed by four 'Projection' and two 'Join' operations. The location of the final operation was fixed before execution of the program.

IV. Design of Entropy and Restricted Chromosome based DSS Query Optimizer (ERC_QO)

Here, the attention has been paid to examine the effect of varying the 'Input Output' to 'Communication Costs' coefficient ratio on the Total Costs of the distributed DSS query. In ERC_QO, the innovation lies in the restricted growth of chromosome and the use of Havrda and Charvat entropy. The individual elements of chromosome represent the location of the site where the concerned operation will be executed. Moreover, the chromosome design restricts the position of projection operations. The projection operation will be executed on same sites where the corresponding selection operations of the 3-Join query were executed [8][10]. In the following chromosome representation, a chromosome for operation site allocation problem is presented. The chromosome is made up of pairs. Each pair represents the operation and its location where it would be executed. In this case, ninth position is selected as crossover location. Therefore, swapping is performed on ninth, tenth and eleventh elements of the selected parents. Here, one point crossover procedure is used and is represented in Figure 2.

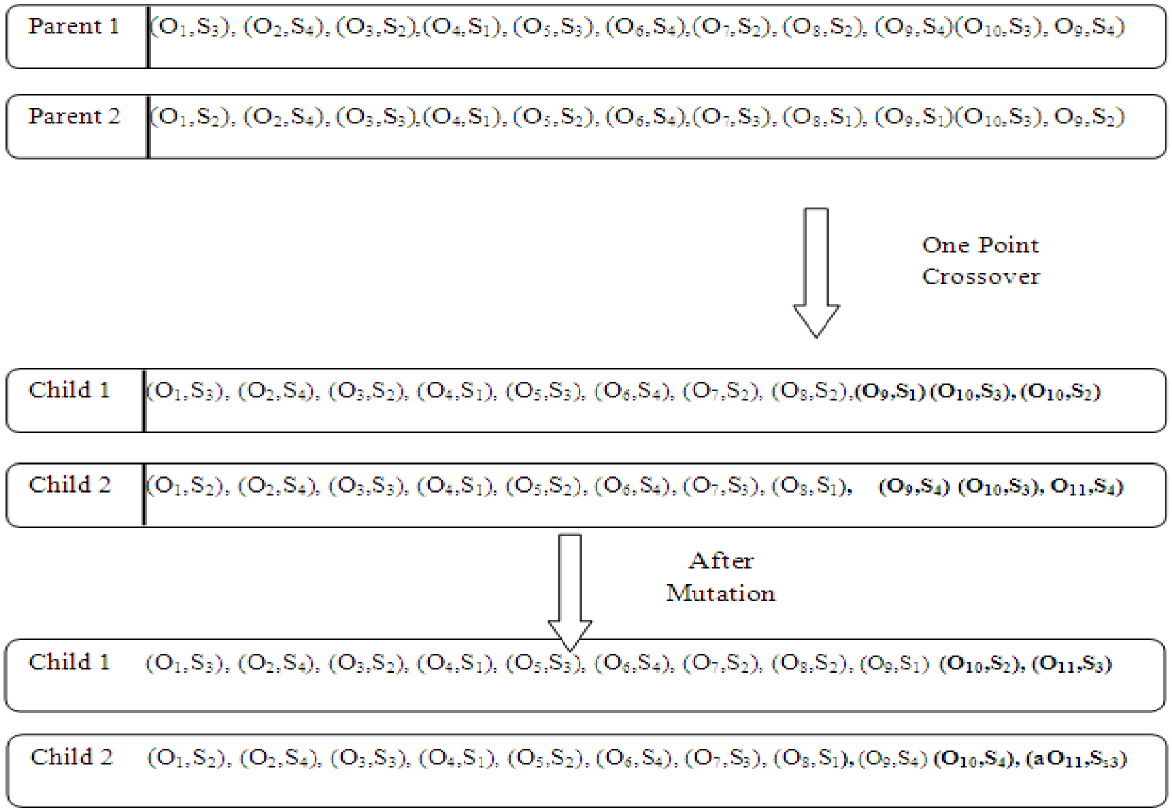

Figure 2: OnePoint Crossover

Mutation is a unary operator. It alters the selected chromosome. It normally shuffles or alters the bits of characters of the offsprings generated by crossover operator. Technically, it acts as an insurance policy to prevent any type of genetic loss of an individual chromosome (offspring)[11][12].

Additionally, the concept of entropy is used at two different levels. Firstly, the concept of entropy is employed for selection operation, so that every affiliate of current generation has uniform probability of selecting as a parent and to perform crossover and mutation operations. The entropy has also been incorporated in selecting a site for execution particular sub operation of DSS query. Here each permissible site has uniform probability of its selection. Furthermore, *Havrda &Charva tentropy* also assist to design low diversity population dilemma which on average transpires in the implementation of 'Genetic Algorithm'.

V. ASSUMPTION

In ERC_QO, the design of chromosome was restrained as execution of projection operation was limited to the sites where selection operation was performed. Further, the use of *Havrda and Charvat* entropy mitigated to sort out the low variation population problem.Moreover, maximum entropy was used to select the parent chromosome to generate offspring, and to allocate a site for the execution of sub operations. All the experiments were carried out based on the following assumptions.

- The computations were made based on the number of data blocks.
- Block size of a relation was assumed to be of 8Kbytes.
- The base relation was replicated randomly on any two different sites. Size of transitional fragments was premeditated based on the selectivity estimation techniques [Rho and March 1995].
- The default ratio of cost coefficients of 'I/O' and 'Communication' was assumed to be 1: 1.6.
- 'Selection' and 'Projection' operations were processed on the sites where the corresponding base relation was placed.
- 'Join' operations were allowed to be executed on any site of a distributed database network

VI. EFFECT OF VAYRING I/O COSTS

Table 1 represents some of the outcomes of the *ERC_QO*, when the '3-Join DSS' query was optimized.

Table 2: Outcome of ERC_QO

S.No.	Design of Chromosome	Input Output Costs	Processing Costs	Comm. Costs	Total Costs
1	2334134433	1466774	164555	21400	1535387
2	2344134433	1469424	164855	21400	1538125
3	2334233411	1392630	156240	24000	1461459
4	2344134431	1433743	160855	21400	1501298
9	1334134412	1361055	152705	21400	1426276
10	2344134421	1362380	152855	24600	1430845
11	2334234441	1501528	168450	22000	1571856
12	1331133144	1814100	179730	17600	2011430
13	1344234432	1398459	156900	21400	1464882
14	2244234431	1430166	160450	21400	1497602
15	2334234431	1430166	160450	21400	1497602

The ratio of 'Input Output Costs' and 'Communication Costs' were varied from 1:1.6 to 1:1. The analysis was carried out to examine the effect of using faster network communication media in the optimization process of the distributed *DSS* query. Five distinct cases were designed by varying the Input Output and Communication Costs coefficients ratio from 1:1.6 to 1:1 as given below:

Case 1 : I/O to Communication Costs Coefficients (1:1.6)

Input Output Costs Coefficients	Communication Costs Coefficients
11 10 12 14	0 16 19 22 16 0 19 22 19 19 0 22 22 22 22 0

Case II: I/O to Communication Costs Coefficients (1:1.5)

Input Output Costs Coefficients	Communication Costs Coefficients
11 10 12 14	0 15 18 21 15 0 18 21 18 18 0 21 21 21 21 0

Case III: I/O to Communication Costs Coefficients (1:1.4)

Input Output Costs Coefficients	Communication Costs Coefficients
11 10 12 14	0 14 17 20 14 0 17 20 17 17 0 20 20 20 20 0

Case IV: I/O to Communication Costs Coefficients (1:1.2)

Input Output Costs Coefficients	Communication Costs Coefficients
11 10 12 14	0 12 14 17 12 0 14 17 14 14 0 17 17 17 17 0

Case V: I/O to Communication Costs Coefficients (1:1)

Input Output Costs Coefficients	Communication Costs Coefficients
11 10 12 14	0 10 12 14 10 0 12 14 12 12 0 14 14 14 14 0

Table 2 represents the values of both 'Communication Costs' and *Total Costs* of a query when the 'Input Output' and 'Communication Costs' coefficients were varied from 1:1.6 to 1:1.

Table 2: Varying I/O to Comm. Costs Coefficients

S.No.	I/O Costs : Communication Costs	Total Cost	Communication Costs
1	1:1.6	1655630	17000
2	1:1.5	1639410	12000
3	1:1.4	1638810	11400
4	1:1.2	1637010	9600
5	1:1	1635410	8000

From Table 2, it is observed that the variation in the ratio of input output to communication costs coefficients brought a drastic change in the 'Communication Costs' of a query. The 'Communication Costs' is reduced to almost half of its value when the ratio was varied from 1:1.6 to 1:1. The effect of varying input output to communication costs coefficients ratio on the 'Communication Costs' of the '3-Join DSS' query is presented in Figure 3. When the ratio was varied from 1:1.6 to 1:1, the 'Communication Costs' of the query is reduced by 90%.

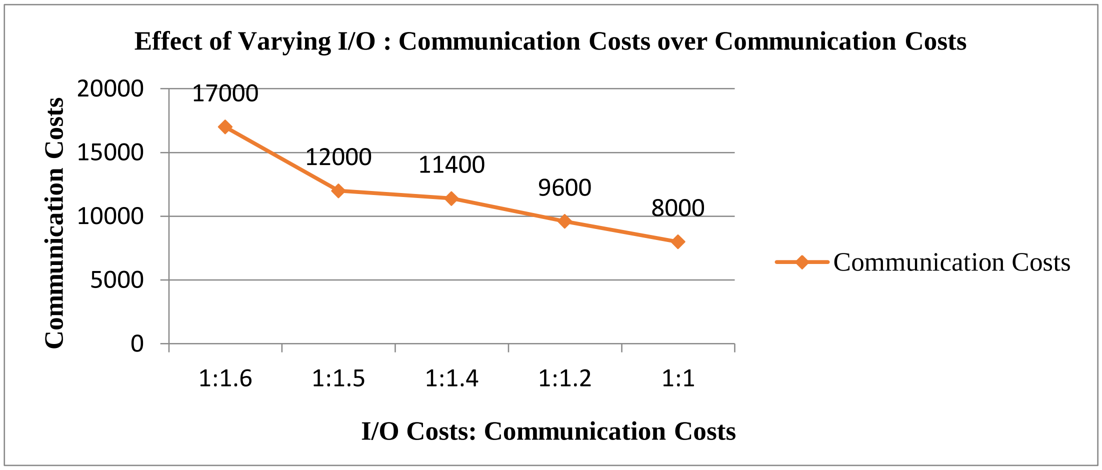

Figure 3: Analysis of Communication Costs

The following Figure 4 represents the values of the Total Costs of the '3-Join DSS' query when the ratio of input output to communication costs coefficients is varied from 1:1.6 to 1:1. Consequently, Total Costs of the query was reduced by 2%.

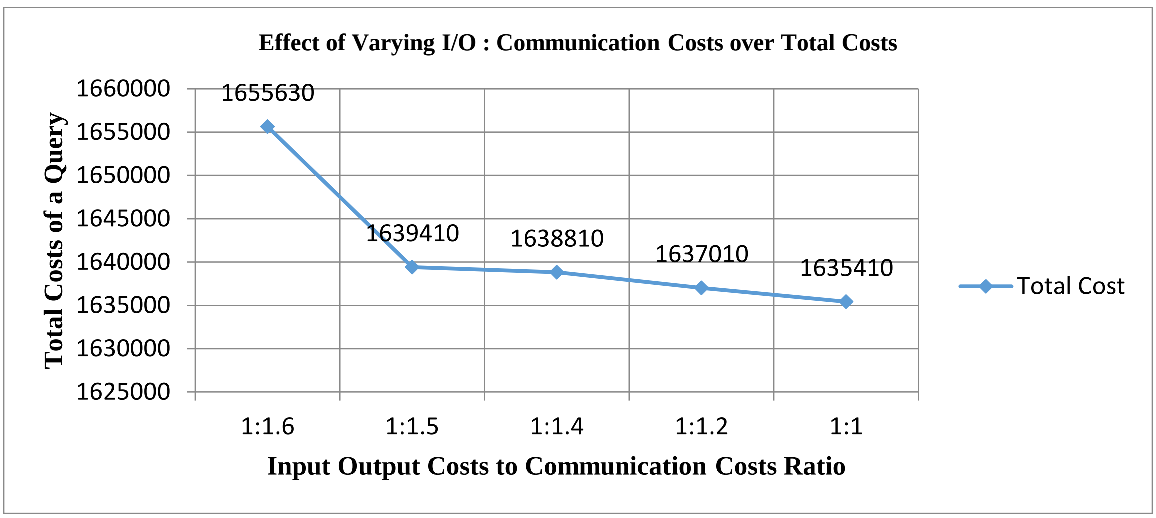

Figure 4: Analysis of Total Costs

The similar effect has been observed when the joins are increased from 1 to 10. By reducing the I/O to communication costs ratio, the total costs of DSS query having 1 to 10 join operations can be reduced by 2-4%. However, significant reduction in communication costs is observed.

VII. CONCLUSION

A query can be optimized using several techniques. Here, a 3-join DSS query has been optimized by using entropy based DSS query optimizer. The basis objective is to examine the effect of varying I/O to communication costs on the Total Costs of the query. Five different cases have been studied. It was experimentally observed that by reducing the I/O to communication costs from 1:1.6 to 1:1, one is able to drastically reduce the communication costs. In addition, for a 3-Join DSS query, the communication costs have been reduced from 17000 to 8000. In general, the total costs can be more effectively optimized by reducing the I/O to communication costs ratio.

REFERENCES

[1] C. D. French. *One Size Fits All- Database Arch. Don't Work for DSS*. ACM SIGMOD Newsletter1995:24-2:449-450.

[2] S. Elnaffar, P. Martin, *Is it DSS or OLTP: Automatically identifying DBMS Workload*, Journal of Intelligent Information System. 30(3) (2008) 249-271.

[3] M. T. Ozsu, V. Patrick, *Principles of Distributed Database System*, second ed., Pearson Education (chap. 1‑6).

[4] SB Yao, AR Hevener. *Query processing in DDS*. IEEE Trans. Soft. Eng. 1979;5(3):177‑87.

[5] Pelagatti G, S Ceri ,. *Allocation of operations in distributed database access.* IEEE Trans. Comp. 1982;31(2):119‑29.

[6] T. M. Martin, K.H. Lam, Judy I Russel, *An Evaluation of Site Selection Algorithms for Distributed Query Processing*, The Computer Journal. 33(1) (1990) 61-70.

[7] K. Donald, *The State of the Art in Distributed QP*, ACM Computing Surveys. 32(4) (2000) 422–469.

[8] M. Sharma, G. Singh, R. Singh. "*Design and Analysis of Stochastic DSS Query Optimizer in a Distributed Database System*". Egyptian Informatics Journal. doi:10.1016/j.eij.2015.10.003.

[9] M. Sharma, G. Singh, G. Singh, R. Singh, *Analysis of DSS Queries in DDS using Exhaustive and GA*, International Journal of Advanced Computing. 36(2) (2013) 1165-1174.

[10] M. Sharma, G. Singh, R. Singh and G Singh. 2015. "*Analysis of DSS Queries using Entropy based Restricted Genetic Algorithm*". Applied Mathematics and Information Science. Vol. 9, Issue 5.

[11] M. Sharma. 2013. "*Role and Working of GA in Computer Science*". International Journal of Computer Applications and Information Technology. 2013; 2(1): 27-32.

[12] S. Ender, C. Ahmat, *An Evolutionary GA for optimization of Distributed Database Queries*, The Computer Journal. 54(5) (2011) 717-725.

Authors Profile

Dr. M. Sharma is MCA, UGC-NET qualified. He has done his doctrate under the kind guidance of Dr. Gurvinder Singh, Dr. Rajinder Singh and Dr. Gurdev Singh. He has around 12 year of teaching expereince. Currently, he is working as an Assistant Professor in the department of Computer Science and Applicatons at DAV University Jalndhar. He has published 15 research papers in different reputed international journals including Scopus and Thomson Reuters and conferences including IEEE held at Bangkok and Calgry, Banff , Canada. His main research work focuses on query optimization, data mining, soft computing and machine learning.

Dr G Singh is working as a Professor and Dean, Faculty of Engineeering at Guru Nanak Dev University, Amritsar. He has more than 20 years of experience. He has published several books and research papers in different reputed international journals including Scopus and Thomson Reuters and conferences including IEEE and Springer. His main research work focuses on distributed computing, parallel processing, soft computing, database and machine learning.

Dr R Singh is working as a Professor and Head, department of Computer Science at Guru Nanak Dev University, Amritsar. He has more than 20 years of experience. He has published several books and research papers in different reputed international journals including Scopus and Thomson Reuters and conferences including IEEE and Springer. His main research work focuses on query optimization, soft computing and and machine learning.

