NECTEC Technical Journal

วารสารวิชาการเนคเทค ปีที่ 1 ฉบับที่ 2 เดือน พฤษภาคม – มิถุนายน 2542 Vol 1, No. 2, May – June 1999 ISSN 1513-2145

Preface

Full Paper

Short Paper

SchoolNet Thailand: An Attempt to Introduce Internet Technology in Thai Schools

Paisal Kiattananan
Dr. Thaweesak Koanantakool

The National Electronics and Computer Technology Center (NECTEC), May 1999

Abstract--SchoolNet Thailand was launched by the National Electronics and Computer Technology Center (NECTEC) in 1995, the year the government proclaimed to Thailand IT (Information Technology) Year . SchoolNet provides global Internet access to secondary schools (grades 7-12) throughout Thailand . By using technology to improve our education system, this project supports the human-resource-development emphasised in the 8th National Economic and Social Development Plan and National Information Technology Committee (NITC) IT-2000 Plan. The purpose of SchoolNet is to bring Thai students' education up to a higher standard and level of quality as well as to bridge the information gap between students in the urban and rural areas by using information technology. Starting with secondary schools, students benefit from nation-wide access to a free web-hosting service and Internet accounts.

The strength of a nation depends on the quality of its citizens. In order to remain economically competitive in the age of an increasingly globalized digital economy; Thailand needs a well-educated workforce adept at the use of Information Technology. Unfortunately, Thailand's existing educational system, like its economy, is badly in need of reform; it is therefore essential that something must be done to inject a new hope into the system and save it from decay.

Internet phenomenon is undoubtedly a major force that will change the world in a very significant way. We (NECTEC) regard it as an opportunity to implant a seed of change in our educational system. We believe that Internet will reinvent the concept of learning, it will also redefine teacher's role and introduce to us the idea of classroom without walls.

History of SchoolNet in Thailand

We launched the SchoolNet Thailand as a pilot project in late 1995, the year the government proclaimed to be Thailand IT (Information Technology) Year. The aim of SchoolNet is to provide global Internet access to secondary schools throughout Thailand. By using technology to improve our educational system, this project supports the human-resource-development emphasis in the 8th National Economic and Social Development Plan as well as the National IT 2000 Plan.

Initially, 50 public secondary schools were selected to participate. With cooperation from the private sector, some computer hardware and software were donated to the schools. We provided free Internet dial-up connection and training. However, there was a weakness in this early system as the network was based solely in Bangkok, schools outside the capital city had to shoulder the high cost of long-distance telephone calls in order to get connected. Even at the peak of Thailand's economic growth, government funding to support SchoolNet was unavailable.

Fortunately, a parallel development occurred in late 1996 - the Kanchanapisek (Golden Jubilee) year. NECTEC was assigned to coordinate and carry out a grand project to celebrate the 50th anniversary of His Majesty the King's accession to the throne. It was agreed that a digital library containing information about His Majesty the King's life and work would best preserve his legacy for later generations.

A nationwide "Intranet" was thus created that enables free local access to this digital library. Anybody with a PC and a

modem can dial a local number from any location in Thailand to access this electronic information free of charge. The website can also be accessed by anyone on the Internet. The success of this Kanchanapisek network lies not only in providing nationwide access to the Thai people but also in the promotion and creation of Thai-language content on the Internet.

Therefore, at the end of 1997, we had successfully completed three crucial components for Thailand's mass-education programs on the Internet, namely school awareness, Thai language content and a nationwide access network. The three components were then integrated in 1998 to form the present national SchoolNet network.

SchoolNet@1509: Striving for Universal Access for Our Children

The marriage between Kanchanapisek network and the original SchoolNet has given birth to a new national SchoolNet in February 1998. This new network is an initiative of Thailand's Princess Maha Chakri Sirindhorn and is also known as SchoolNet@1509. SchoolNet@1509 is designed to provide free Internet dial-up access to 1,500 schools nationwide. The network can be accessed by dialing the number 1509 from any location in the country. The cost for accessing SchoolNet@1509 is 3 Baht (which is a local telephone call) regardless of the location where the call originated, thus reducing the information gap between those living in the urban and the rural areas. We are proud that SchoolNet@1509 is the first serious attempt to provide universal access to the Thai people according to the Article 78 of the new Constitution.

SchoolNet@1509 has been achieved through the cooperation of four government agencies. A national SchoolNet committee was set up which consisted of NECTEC, the Telephone Organization of Thailand (TOT), the Communication Authority of Thailand (CAT) and the Ministry of Education (MOE). Each has an important role and contribution to SchoolNet@1509. TOT sponsors domestic Internet bandwidth while CAT donates international Internet bandwidth to SchoolNet@1509. The role of NECTEC is to design, maintain and operate the network and central computer systems in SchoolNet@1509. The MOE selects schools; it also coordinates, promotes and supports the use of Internet in these schools.

Due to the participation of several government agencies, SchoolNet policy concerning school eligibility for accounts in SchoolNet was changed. The new policy in SchoolNet@1509 allows both public and private schools in primary or secondary levels to participate; this is in contrast with the previous policy that allows only public schools in secondary level to be member of SchoolNet. At the time of this writing (April 1999), there are altogether 923 schools and 1757 users who are members of SchoolNet@1509.

Cooperation with MOE's Resource Center Project

Despite the fact that SchoolNet@1509 has laid a good foundation for Internet infrastructure in Thailand, the battle to bring schools online is far from over. One of the major obstacles that we faced is the lack of adequate computer and communication equipment in schools. In this respect, the Resource Center (RC) project of the Ministry of Education fits nicely into the picture. It complements SchoolNet@1509 by aiming at providing computer and communication systems for 420 schools nationwide.

In addition, the project will select 76 most promising schools (one school per province) out of 420 to become 'school nodes' of SchoolNet@1509. These school nodes will be linked to SchoolNet@1509 via high speed permanent connections (leased circuit in contrast with dial-up modem connections that most schools get). However, with this privilege come greater responsibilities, in other words these selected schools will have to become local training centers and provide support to other schools in their respective provinces.

Unlike SchoolNet@1509, MOE's Resource Center project has not yet been implemented, the government funding of this project has to be secured and a few details need to be worked out between MOE and NECTEC to ensure smooth cooperation. Barring any unexpected circumstances, the infrastructure both inside (RC) and outside (SchoolNet@1509) schools will be greatly enhanced by these projects but it only deals with the "hardware" side of the problem. When the infrastructure is ready for schools to connect to the Internet, two questions will immediately be asked "why do we need Internet in schools?" and "how do we use

this computer to connect to the Internet?". This raises two important issues concerning Internet content and human resource development.

Content Creation and Other Activities in SchoolNet Thailand

Internet is like a double-edged sword. While its potential to elevate the educational standard is huge, it is also a good medium for transferring dangerous or inappropriate materials to children. Students must be guided in such a way that they spend time wisely on the Net enriching their knowledge instead of idling away the time by playing games or engaging in meaningless online chat with their peers.

After spending the first few years developing the network infrastructure, we have gradually come to realize the importance of content – especially Thai-language content on the Internet. The first serious attempt to tackle this problem was undertaken in Kanchanapisek project where we worked with eleven organizations which had served the country through many successful projects due to His Majesty the King's initiative. Thousands of Thai language web pages were created that document vast array of information concerning the royal-initiated projects such as rural development, agriculture, irrigation and junior encyclopaedia project etc.

In a similar manner, SchoolNet@1509 needs good local content to attract teachers and students online. If left alone in the Cyberspace dominated by English content, language barrier will discourage most teachers and students from using the Internet. Therefore, it is essential that we have Thai language content with good educational value that is designed to help the children do better in school.

The SchoolNet Content Development project was therefore started in September 1998. We commissioned Kasetsart University to carry out the project in conjunction with the Institute for the Promotion of Teaching Science and Technology (IPST) and some selected schools. The objective is to create educational websites in Thai language for secondary school students featuring 7 major academic subjects, namely Computer Science, Mathematics, Physics, Chemistry, Biology, Engineering and Environment. These websites are scheduled to be launched by December 5, 1999 to celebrate the 72nd birthday of His Majesty the King of Thailand. Moreover, we hope that this effort will demonstrate and induce other schools to create their own educational websites and thus contribute to the overall content for school children in Thailand.

Apart from content creation, we feel that it is also important to have activities organized such that teachers and students learn how to get the most from the vast educational potential of the Internet. Held once a year since 1996, Seagate Technology (Thailand) has joined us in organizing Internet Training Camp for secondary school children. Students who signed up were trained in web development; they then competed against each other by building their own webpages centering around each year's theme, e.g. environment protection, the solar system etc.

Global Learning and Observations to Benefit the Environment (GLOBE) program is another activity that offers a good opportunity for teachers and students in SchoolNet@1509 to collaborate with their counterparts around the world to study and understand the global environment. GLOBE students make environmental observations at or near their schools and report their data through the Internet. Through our effort as the secretariat office of the National Information Technology Committee (NITC), Thailand is now one of the more than 80 countries that participate in the GLOBE program. A special committee has now been set up by the NITC to oversee Thailand's participation and the IPST was appointed as the country coordinator.

Internet Training for Schools in the Project

In 1996, the early phase of SchoolNet, we provided introductory training courses in Internet for participating schools at our own facility in NECTEC (Bangkok). However, as the project advances, the number of member schools increases rapidly all over the country. It is soon obvious to us that the early model of centralized training is no longer practical.

As SchoolNet@1509 made a debut in February 1998, we knew that we need some kind of distributed training model. Rajabhat Institutes soon emerged as our ideal partner. Although only 8 of their 36 campuses nationwide were able to offer Internet training courses, this was good enough since these 8 campuses were in different provinces and could serve as regional training centers for schools in that area. Having regional training centers, we helped schools save tremendously on travel expenses. We were also

able to train more schools with the help of these Rajabhat Institutes. Last October, 274 schools and 419 users were trained this way.

We firmly believe that this training model is the right way to go. We plan to expand it to every province in the country so that schools no longer have to travel outside their provinces to get the training and they will be able to obtain local support in the province. However, since Rajabhat has only 36 campuses, we therefore need to look for other partners. As mentioned earlier in this article, 76 schools will be selected from 420 in the Resource Center project to connect to SchoolNet@1509 as Internet nodes. These 76 schools (one per province) will be selected in such a way that they have necessary resources and the required capability to lead others in Internet development. We will work with these selected schools and train them to be trainers for other schools in their provinces.

This strategy of training the trainers will result in SchoolNet@1509 having 76 local training and support centers nationwide. If successful, it will accelerate the growth of SchoolNet@1509 significantly.

Linux School Internet Server (Linux-SIS)

It is often said that computer is too complicated to use. This is even more true for most teachers. Occasional training does help teachers become more familiar with the use of computer and the Internet but the best way is to make things simpler for them.

We developed Linux School Internet Server (Linux-SIS) to simplify the task of installing and managing Internet server in schools. As Linux operating system is a freeware, Linux-SIS also serves as a low cost alternative to other UNIX or Windows NT systems.

Linux-SIS is a software package consisting of Slackware Linux Distribution and other add-on freeware to ease installation and system management. The programs that we developed and bundled with Linux-SIS simplify installation process by presetting initial parameters for users. To customize the system, users just answer some simple questions and Linux-SIS will set it up accordingly. The tasks of system administrator are also simplified by a web-based tools that manage the server, thus eliminating the need for any technical background to perform them.

Schools that need Internet servers are finding Linux-SIS a great help to them. Linux-SIS can be used as an Internet gateway and let other PCs in school's local area networks share an Internet connection. It can also be used as an Intranet server for schools. We have been developing Linux-SIS since August 1996. It is now in its third version and is distributed with a book written by SchoolNet team.

Users Support: SchoolNet Volunteer Program

Normally after the Internet/PC training courses, quite a number of teachers often encounter technical problems when they go back to schools and really start getting online. These problems for new users get more complicated when the only place that they can get help is NECTEC. And a phone call to NECTEC's helpdesk might cost them 18 baht/min for some areas in the country. This is certainly not affordable for most schools.

One possible solution is to have the local community get involved. Since all of us should take greater responsibility in the education of our children, so why not sacrifice some of our time to help our schools get online? We believe that this spirit is what drives NetDay activities in the United States and the rest of the world. So we proposed a volunteer program for SchoolNet@1509. When SchoolNet Vounteer Program was made public, many people signed up and we selected 60 most qualified people from 23 provinces.

With the funding support of Kenan Institute Asia, we are now ready to kickstart the volunteer program in the next few weeks. Each volunteer will be responsible for a few schools in his/her area (province). The volunteers are expected to provide first-level support to the schools, this may include answering technical questions, giving advice or even go to schools to provide on-site support, in some cases. We hope that these volunteers will not only make lives easier for most teachers but they will also help convey accurate information about SchoolNet@1509 so that schools better understand our project.

Conclusion

Any improvement in the quality of our education must be achieved through the balance

of the three factors or 3T, namely Technology, Teaching materials (contents) and Teachers (their qualifications). As described in this article, we have put our development efforts in all of them. However, being an organization responsible for technology, so far our achievement in Technology (communication capability) is far exceeding the other two factors. Additional efforts are needed to convince related government agencies to act and improve upon Net contents as well as the teachers' training issue.

The telecommunication infrastructure is like the transportation infrastructure. Building a good road to a rural community without preparing them for the onslaught of urban culture and consumerism will only destroy their livelihoods. It is thus important that we strengthen the rural communities before exposing them to the economic risks of new road and electricity. By the same argument, Internet can be dangerous for an unprepared community. This is why we do regulate SchoolNet@1509 availability to only “known communities”. It will never be “unlimited usage” to “unlimited number of schools”.

It might be too early to judge whether SchoolNet@1509 will succeed in the long run. However, it is undeniable that this project has already made a significant impact on many schools in Thailand. And we have achieved this with very little dependence on the government budget.

Needless to say, NECTEC should be viewed as just an incubator of the early effort to introduce Internet in Thai schools. It is clear that no single organization in the country can accomplish this task alone. Thailand urgently needs a joint effort by different government ministries if it really wants to see that every student has a chance to get online.

Appendix: SchoolNet Milestones

Paisal Kiattananan

My resume is B.Sc. in Mathematics from Tsukuba University., Japan, M.Sc. in Computer Science from University. of Minnesota, USA.
I used to work in IBM Japan and IBM Thailand.
I joined NECTEC in Nov 1995. I am now in Network Technology Laboratory.

Thaweesak Koanantakool
Director of NECTEC

Dr.Thaweesak "Hugh" Koanantakool received his Bachelor and Ph.D. degrees in Electrical Engineering from Imperial College of Science and Technology, London University. He had a number of industrial contracts in the UK before he came back to Thailand to start his government service career in 1981. He taught in Electrical Engineering with the Faculty of Engineering, Prince of Songkla University. In 1985, he moved to Bangkok Thammasat University and was appointed Associate Director of the Information processing Institute for Education and Development. Since 1994, he became Deputy Director of NECTEC as well as leading the Network/Software Technology labs. Thaweesak introduced the Internet into Thailand and set up the largest academic and research network known as ThaiSarn under NECTEC. He later co-founded the first Internet Service Provider (ISP) owned by Thai government in 1995. The ISP, Internet Thailand Company Limited,at present is the largest ISP in Thailand, has 45% market share (by IP numbers managed). In 1996-1997, Thaweesak led Thailand's Information Superhighway test bed Project funded by NECTEC. The project was a major test bed in Thailand using ATM switches for both local area and wide-areas. In August 1998, Thaweesak was appointed the Director of NECTEC.

A Node-Driven Parse Pruning Technique for Probabilistic GLR Parsing

Virach Sornlertlamvanich
National Electronics and Computer Technology Center
22F Gypsum Metropolitan Tower, 539/2 Sriayudhya Rd.,
Rajthevi, Bangkok 10400, Thailand.
virach@links.nectec.or.th

ABSTRACT — This paper proposes a new technique, *a node-driven parse pruning technique*, in pruning the less probable parses for GLR parsing algorithm. Without decreasing the efficiency of GLR parsing, this technique estimates the number of parses in the GSS (graph-structured stack) based on the number of expanded nodes during the parse process. We show the evaluation results of various beam width settings for pruning, and compare the parse time and space consumption against full parsing results. Our node-driven parse pruning algorithm allows pruning in a left-to-right manner without modifying the GSS.

KEY WORDS — Probabilistic GLR parsing, parse pruning, GSS, beam search.

บทคัดย่อ — บทความนี้เสนอวิธีการในการขจัดผลการแจงส่วนที่มีความเป็นไปได้ต่ำออกไป, สำหรับการใช้งานในการแจงส่วนแบบ GLR. วิธีการนี้ได้อาศัยจำนวนขั้นของการเคลื่อนสถานะมาเป็นตัวประมาณจำนวนของแขนงของการแจงส่วนใน GSS, จึงเรียกวิธีนี้ว่า "การขจัดผลการแจงส่วนโดยอาศัยจำนวนโนด". เราได้ทำการทดลองเพื่อเปรียบเทียบการใช้เวลาและหน่วยความจำสำหรับการขจัดผลการแจงส่วนในเกณฑ์ต่างๆ กับการแจงส่วนปกติ. วิธีการนี้มีข้อดีอีกอย่างหนึ่งคือสามารถคำนวณได้ตามการแจงส่วนจากซ้ายไปขวาโดยไม่ได้เปลี่ยนแปลงโครงสร้างของ GSS ซึ่งมีประสิทธิภาพของการเก็บข้อมูลของการแจงส่วนที่ดีอยู่แล้ว.

คำสำคัญ — การแจกส่วน GLR ด้วยความน่าจะเป็น, การขจัดแขนงของการแจงส่วน, GSS, การสืบค้นแบบบีม.

1 Introduction

Pruning is an essential paradigm to reduce the search space in parsing. The idea of pruning is to exclude hypotheses from further investigation if the parses turn out to be unlikely, based on evaluation of partial data. The efficiency of pruning technique depends mostly on the parsing paradigm which ranks the parse candidates according to their likelihood. The preciseness in distributing the parse probabilities into each partial parse tree is essentially considered for applying the pruning technique. This pruning technique is generally proposed for any probabilistic parsing models applied to the GLR parsing algorithm. However, the probabilistic GLR parsing model [9, 11] is introduced to eveluate

the pruning technique because of its efficiency over all other models on GLR parsing as reported in [10]

Extracting all possible parses from a packed parse forest is crucially constrained by both time and memory space concerns. Carroll and Briscoe [2] have proposed a method for extracting the n-best parses from a complete packed parse forest. However, their method still requires the parser to parse to completion, and then identifies the n-best parses from the resultant packed parse forest. Their method can only save time in the actual extraction of the n-best parses, and does not concern itself with time and memory space consumption during the parse process. Sentences, in general, are ambiguous because of the wide-coverage nature of context-free grammars, but most of the possible parses for a sentence are pretty senseless. In practice, it is not reasonable to parse exhaustively to obtain some of the most probable parses. We would obtain the results more quickly if it were possible to prune off the less probable parses as early as possible.

The compaction of the graph-structured stack (GSS) prevents us from applying the Viterbi algorithm [13] directly to GLR parsing. By way of the GSS, the parse stack is dynamically changed and does not keep trace of the various parses. Both of the Viterbi beam-search methods proposed in [14, 15] and [16] need additional storage to keep trace of the parses, which overrides the benefits of using GSS in GLR parsing.

Lavie and Tomita [7] introduced a beam search heuristic for GLR* parsing. GLR* parsing is a noise-skipping parsing algorithm which allows shift operations to be performed from inactive state nodes of the GSS. This amounts to skipping words at any previous state in the GSS. The purpose of introducing the beam search algorithm is to limit the number of inactive state nodes for performing shift operations. The algorithm simply considers performing shift operations from the nearest state nodes until the number of state nodes reaches the limit. In fact, any state nodes may be merged and be common to several different sub-parses. Therefore, any undesirable less-probable parses that end up with the same state nodes are included within the beam, which leads to an inefficient beam search. It is also possible that the most probable parses be overlooked because of inaccuracy in scoring and parse estimation.

Pruning with a beam search technique is widely discussed in the speech processing community [12, 5, 8, 16]. Steinbiss et al. [12] give a good summary of previous research on beam search methods and propose some improvements to beam searching, as *histogram pruning*. This method introduces an additional pre-specified upper limit on the number of active points per frame (or active nodes per time frame) to limit the expansion of hypotheses. The result of their experiments show that the search space is efficiently reduced by observing the distribution of the number of states over the parse.

In this paper, we propose a new method for pruning parses that have a lower probability than parses within a predetermined beam width using a histogram pruning-like algorithm called the *node-driven parse pruning algorithm*. The number of expanded nodes is effectively used to estimate the number of parses in the GSS. We also show the evaluation results of various beam width settings, and compare the parse time and space consumption against full parsing results. Our node-driven parse pruning algorithm allows pruning in a left-to-right manner without modifying the GSS.

2 Probabilistic GLR Parsing

The probabilistic GLR language model (PGLR) has previously been proven to be better than existing models, in particular the model proposed by Briscoe and Carroll [1] and the baseline model using a probabilistic context-free grammar (PCFG), in parsing strings of parts-of-speech (non-word-based parsing) [10]. Parsing a sentence from the morphological level makes the task much

more complex because of the increase of parse ambiguity stemming from word segmentation ambiguities and multiple corresponding sequences of parts-of-speech. In this paper, we empirically evaluate the preciseness of a probabilistic model for PGLR against that for Briscoe and Carroll's model (B&C), which is based on the same GLR parsing framework. We also examine the benefits of context-sensitivity in GLR parsing, of the PGLR model against the "two-level PCFG" model [4] or "pseudo context-sensitive grammar" model (PCSG)—recently presented in [3]—which has been shown to capture greater context-sensitivity than the original PCFG model, by empirical results and qualitative analysis.

Like the B&C model, PGLR inherits the benefits of context-sensitivity in generalized LR parsing (GLR). Its LR parsing table ("LR table" for short) is generated from a context-free grammar (CFG) by decomposing a parse into a sequence of actions. Every action in the LR table is determined by the pairing of a state and input symbol, so that it is valid to regard the state/input symbol pair as the context for determining an action. As a result, PGLR inherently captures two levels of context, i.e. global context over structures from the source CFG, and local n-gram context from adjoining pre-terminal constraints. Inui et al. [6] showed that B&C has some defects in distributing parse probabilities over the actions of an LR table. One is that, in B&C, no distinction is made between actions when normalizing action probabilities over the states in an LR table, while PGLR distinguishes the action probability normalization of states reached immediately after applying a shift action, from states reached immediately after applying a reduce action. B&C repeatedly counts the next input symbol when computing the probabilities (though the next input symbol is deterministic), if parsing is at the state reached immediately after applying a reduce action. Redundantly including the probabilities of the preceding input symbols in this case significantly distorts the overall parse probabilities. Moreover, subdividing reduce action probabilities according to the states reached after applying reduce actions is also redundant because resultant stack-top states after popping for reduce actions are always deterministic. B&C thus estimates parse probabilities lower than they should be.

Considering a parse derivation as a sequence of transitions between LR parse stacks (T) and assuming that the current stack σ_i contains all the information of its preceding parse derivation, PGLR defines the probability of a complete stack transition as:

$$P(T) = \prod_{i=1}^{n} P(l_i, a_i, \sigma_i | \sigma_{i-1}) \qquad (1)$$

where l_i is an input symbol and a_i is an action.

The PGLR model distributes the probability of a complete stack transition into each transition by assuming that the stack-top state (s_i) represents the stack information beneath it, then:

$$P(l_i, a_i, \sigma_i | \sigma_{i-1}) \approx \begin{cases} P(l_i|s_{i-1}) \cdot P(a_i|s_{i-1}, l_i) & \\ \quad = P(l_i, a_i|s_{i-1}) & (s_{i-1} \in \boldsymbol{S_s}) \\ P(a_i|s_{i-1}, l_i) & (s_{i-1} \in \boldsymbol{S_r}) \end{cases}$$

such that:

$$\sum_{l \in La(s)} \sum_{a \in Act(s,l)} p(a) = 1 \quad (\text{for } s \in \boldsymbol{S_s})$$

$$\sum_{a \in Act(s,l)} p(a) = 1 \quad (\text{for } s \in \boldsymbol{S_r})$$

where $p(a)$ is the probability of an action a, $\boldsymbol{S_s}$ is the class of states reached after applying a shift action, including the initial state, and $\boldsymbol{S_r}$ is the class of states reached after applying a reduce action.

Therefore, the parse probability upto the current state can be calculated and stored in the current stack-top state.

3 The node-driven parse pruning algorithm

It is inefficient to compute parse probabilities for all parses from the initial state successively up to the current state of parsing, because we have to keep trace of all possible parses individually. This also degrades the benefits of local ambiguity packing in the graph-structured stack (GSS). To counter this inefficiency, we observe the number of state nodes at each stage of parsing time and compute all possible parses only when the number is more than a threshold. Since the number of state nodes in a GSS can be viewed as an indicator of the degree of ambiguity, we indirectly estimate the number of parses by observing the number of state nodes in the GSS, and apply this as a threshold for activating the parse pruning process as shown in **Algorithm-1**.

The threshold T_t at time t is computed by:

$$T_t = G_t \cdot n_t \tag{2}$$

where T_t is the estimated number of parses at time t, G_t is the gain based on the number of state nodes and the length of the input string up to time t, and n_t is the number of state nodes at time t.

The gain G_t can be computed by:

$$G_t = \frac{\sum_{i=1}^{L} n_{t-i} T_{t-i}}{\sum_{i=1}^{L} n_{t-i}^2} \tag{3}$$

where L is the number of past observations (a good setting for L is 5, as reported in [5]). The gain is used in adaptive pruning [5] by regarding a pruning process as a non-linear time-variant dynamical system. In our implementation, we simply set the gain G_t as a linear time-variant to reduce the computational overhead. Since our beam width is fixed, the estimated number of partial parses at each parsing time is used to activate the parse pruning algorithm.

Algorithm-1. Node-driven parse pruning process.

1. *If the number of parses estimated from the number of state nodes in the GSS is over the threshold T_t, compute the number of partial parses, else return.*
2. *If the number of parses is greater than the predetermined beam width, then compute the probabilities for each partial parse and individually store the sequences of state transitions with corresponding probability at each active state node (top node of each stack), else return.*
3. *Sort the sequences of state transitions according to the probabilities and determine the minimum probability of the parse within the beam width.*
4. *Mark sequences of state transitions that have probabilities less than the minimum as 'pruned'.*
5. *Apply the next actions to the active state nodes only if there is at least one possible parse (unmarked sequence of state transitions). For reduce actions, check reduceable parses with the sequences of state transitions at the active state nodes.*

Figure 1 exemplifies the GSS in the process of parse pruning. Suppose that the beam width is equal to one, and the first parse (state sequence (0,3,6,13,9)) at node 9 is the only parse within the beam width. Here, only the action [re1,14] is executed, with the result as shown in Figure 1 (b). Note that we do not extend the stack at active state node 13 because all parses up to state node 13 are marked to be pruned off. At active state node 17, after trying the action [re1,17], the state sequence (0,4,11,3,9) which is marked to be pruned off, is activated. Therefore, the parse after this action is also disregarded.

As a result, we can parse with smaller memory space and lower computational time overhead than that required in full parsing. However, parsing with this pruning technique gives appropriate results if and only if the exploited probabilistic model provides precise probabilistic estimates for the partial parses.

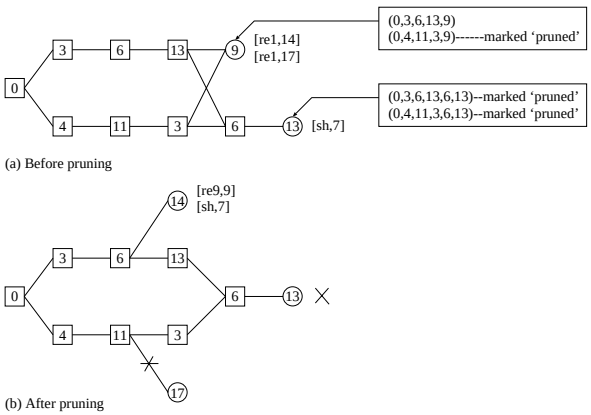


Figure 1: Parse pruning within a graph-structured stack. Circled state numbers are active nodes. All possible parses (sequences of state transitions) at each active node are shown in the box pointing to that node. The action/state pairs after applying the actions are shown in the square brackets next to the active nodes.

This is because the beam search is an approximate heuristic method that does not guarantee that the interpretation of a sentence is the best possible interpretation.

4 Efficiency of parse pruning in PGLR

We calculated the efficiency of parse pruning using the PGLR model for a varying beam width. Parsing can be sped up by reducing the beam width, excepting that the correct parses can potentially be pruned off if the beam width is too small. Figure 2 shows that PGLR provides quite a precise probabilistic estimate for partial parses, in that the parsing accuracy increases steeply with a small increase in the beam width. The parser performs equally well with a beam width of around 30 as with full parsing. Time consumption in parsing using our pruning technique is linear in sentence length, while it is exponential for full parsing. For example, our pruning technique requires only $\frac{1}{10000}$ of the parsing time required for full parsing, for 25 character long sentences (there are more than 200,000 parses if completely parsed).

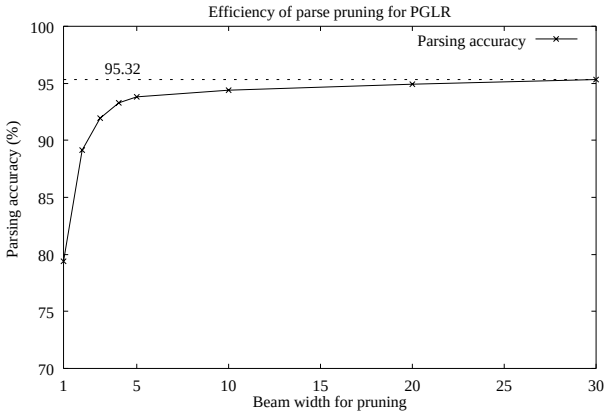


Figure 2: Parsing accuracy under a varying beam width for parse pruning.

We evaluated our pruning algorithm by observing time and space consumption between full parsing and parsing with the pruning algorithm at the beam width of 30. Distributions of state nodes against input symbols in single sentence, are shown in Figures 3 and 4. It is obvious that our parse pruning algorithm can drastically reduce the number of nodes used in parsing both sentences considered. Consequently, the parsing time for both sentences is also visibly reduced because of the reduction in search space. Parsing time is reduced from 1709.62 seconds to 1.0 seconds and 649.49 seconds to 5.52 seconds, in parsing the 33 and 36 character length sentences, respectively.

Table 1 shows the average efficiency of time and space consumption when parsing with the node-driven parse pruning algorithm at a beam width of 30, as compared to full parsing.

The setting for the beam width is a trade-off between parsing accuracy and parsing time. In practice, a beam width of around 20 is likely to be sufficient to produce satisfactory parsing results for the ATR corpus.

Table 1: Average time and space consumption when parsing with the node-driven parse pruning algorithm, as compared to full parsing.

	Average number of state nodes per sentence	**Average parse time, seconds per sentence**
Full parsing	9,146	163
Beam width = 30	630	0.243

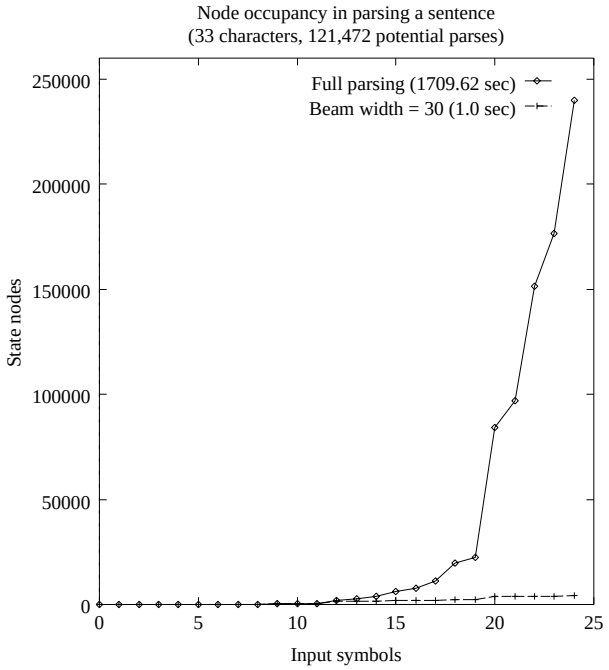


Figure 3: Distribution of state nodes over input symbols in parsing a sentence of 33 characters with 121,472 potential parses.

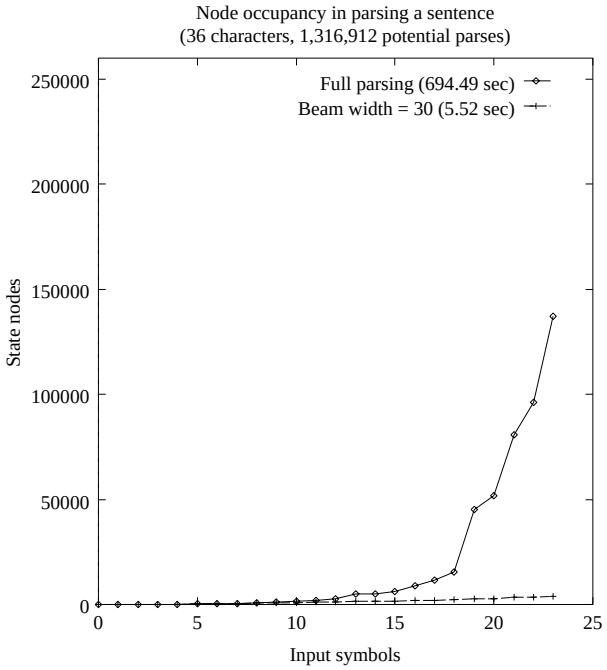


Figure 4: Distribution of state nodes over input symbols in parsing a sentence of 36 characters with 1,316,912 potential parses.

5 Conclusion

Beam searching is an approximate-based heuristic method that does not guarantee that the final interpretation of a sentence is the best possible interpretation. Nevertheless, by carefully managing the number of parses in the GSS using the *node-driven parse pruning technique* we make significant efficiency gains in both time and space consumption. Moreover, when coupled with a precise model for partial parse probability estimation, our pruning technique yields the same results as full parsing at a beam width of about 30, using about 0.07% of the relative parsing space and 0.0015% of relative parsing time over full parsing.

Our *node-driven parse pruning algorithm* allows parse pruning in the GSS in a strict left to right fashion, with the benefit of being able to disregard less-probable parses at as early a stage as possible.

Our pruning algorithm also retains the advantages of the GSS, with neglible cost in estimating the number of parses in the GSS.

References

[1] T. Briscoe and J. Carroll. Generalized Probabilistic LR Parsing of Natural Language (Corpora) with Unification-Based Grammars. *Computational Linguistics*, 19(1):25–59, 1993.

[2] J. Carroll and T. Briscoe. Probabilistic Normalisation and Unpacking of Packed Parse Forests for Unification-Based Grammars. In *Proceedings of AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pages 33–38, 1992.

[3] E. Charniak and G. Carroll. Context-Sensitive Statistics for Improved Grammatical Language Models. In *Proceedings of AAAI-94*, pages 728–733, 1994.

[4] M. Chitrao and R. Grishman. Statistical Parsing of Messages. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 263–266, 1990.

[5] H. V. Hamme and F. V. Aelten. An Adaptive-Beam Pruning Technique for Continuous Speech Recognition. In *Proceedings of International Conference on Spoken Language Processing*, pages 2083–2086, 1996.

[6] K. Inui, V. Sornlertlamvanich, H. Tanaka, and T. Tokunaga. A New Formalization of Probabilistic GLR Parsing. In *Proceedings of the 5th International Workshop on Parsing Technologies*, 1997.

[7] A. Lavie and M. Tomita. GLR* - An Efficient Noise-Skipping Parsing Algorithm for Context-Free Grammars. In *Recent Advances in Parsing Technology*, chapter 10, pages 183–200. Kluwer Academic Publishers, 1996.

[8] S. Ortmanns, H. Ney, and A. Eiden. Language-Model Look-Ahead for Large Vocabulary Speech Recognition. In *Proceedings of International Conference on Spoken Language Processing*, pages 2095–2098, 1996.

[9] V. Sornlertlamvanich. *Probabilistic Language Modeling for Generalized LR Parsing*. Doctoral dissertation, Tokyo Institute of Technology, Tokyo, Japan, March 1998.

[10] V. Sornlertlamvanich, K. Inui, K. Shirai, H. Tanaka, T. Tokunaga, and T. Takezawa. Empirical Evaluation of Probabilistic GLR Parsing. In *Proceedings of the Natural Language Processing Pacific Rim Symposium*, pages 169–174, 1997.

[11] V. Sornlertlamvanich, K. Inui, H. Tanaka, T. Tokunaga, and T. Takezawa. Empirical Support for New Probabilistic Generalized LR Parsing. *Journal of Natural Language Processing*, 6(3):3–22, 1999.

[12] V. Steinbiss, B.-H. Tran, and H. Ney. Improvements in Beam Search. In *Proceedings of International Conference on Spoken Language Processing*, pages 2143–2146, 1994.

[13] A. J. Viterbi. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. In *Proceedings of IEEE Transactions on Information Theory*, pages 260–269, 1967.

[14] J. H. Wright and E. N. Wrigley. GLR Parsing with Probability. In *Generalized LR Parsing*, pages 113–128. Kluwer Academic Publishers, 1991.

[15] J. H. Wright, E. N. Wrigley, and R. Sharman. Adaptive Probabilistic Generalized LR Parsing. In *Proceedings of the 2nd International Workshop on Parsing Technologies*, pages 100–109, 1991.

[16] T. Yamada and S. Sagayama. LR-Parser-Driven Viterbi Search with Hypotheses Merging Mechanism Context-Dependent Phone Models. In *Proceedings of International Conference on Spoken Language Processing*, pages 2103–2106, 1996.

Virach SORNLERTLAMVANICH:

is a researcher of the National Electronics and Computer Technology (NECTEC) of Thailand since 1992. He received the B.Eng. and M.Eng. degrees from Kyoto University, in 1984 and 1986, respectively. From 1988 to 1992, he joined NEC Corporation and involved in the Multi-lingual Machine Translation Project supported by MITI. He received the D.Eng. degree from Tokyo Institute of Technology in 1998. His research interests are natural language processing, lexical acquisition and information retrieval.

Economic Dispatch with Linear Decreasing and Staircase Incremental Cost Functions by Micro Genetic Algorithms

W. Ongsakul and J. Tippayachai
Department of Electrical Engineering
Sirindhorn International Institute of Technology
Thammasat University, Pathumthani 12121, Thailand

ABSTRACT- In this paper, a Micro Genetic Algorithm (MGA) to solving the economic dispatch (ED) problem with linear decreasing and staircase incremental cost (IC) functions for combined cycle (CC) units is proposed. To demonstrate its advantage, the proposed MGA methods with two different encoding schemes are tested and compared to the unconstrained Brute Force (BF), Simple Genetic algorithm (SGA), Merit Order Loading (MOL), and equal-lambda based Newton methods on five combined cycle units. As the ramp rate constraints are taken in account, the proposed MGA solutions are ensured to be feasible. Test results indicate that the solutions are found to be close to the optimal solution of the unconstrained BF method and its total fuel costs are lower than those of the SGA, MOL, and Newton methods.

KEYWORDS: Micro Genetic Algorithm (MGA), Economic Dispatch (ED), Brute Force (BF), Simple Genetic Algorithm (SGA), Merit Order Loading (MOL), Newton Method.

บทคัดย่อ บทความนี้นำเสนอไมโครจินิติก อัลกอลิทึ่มในการแก้ปัญหาการ Dispatch หน่วยผลิตไฟฟ้าอย่างประหยัดสำหรับหน่วยผลิตพลังความร้อนร่วมที่มีสมการ Incremental Cost ที่มีความชันเป็นลบ และ ที่เป็นขั้นบันได เราทดสอบ ไมโครจินิติกอัลกอลิทึ่ม และ เปรียบเทียบผลกับวิธี Brute Force ที่ไม่คิดขีดจำกัดอัตราการเพิ่มกำลังการผลิต, วิธีจินิติกอัลกอลิทึ่มอย่างง่าย, วิธี Merit Order Loading (MOL), และ วิธีนิวตั้น บนระบบไฟฟ้ากำลังที่มีหน่วยผลิตพลังความร้อนร่วมจำนวน 5 หน่วยเพื่อที่จะแสดงถึงประโยชน์ของอัลกอลิทึ่ม ผลที่ได้จากการทดสอบมีความเป็นไปได้ในการ Dispatch เนื่องจากเราพิจารณาอัตราการเพิ่มกำลังการผลิตในการคำนวณ ผลการทดสอบให้ต้นทุนการผลิตรวมใกล้เคียงกับต้นทุนการผลิตรวมต่ำสุด และ ต้นทุนการผลิตรวมที่ได้ต่ำกว่า ผลที่ได้จากวิธีจินิติกแบบธรรมดา, วิธี MOL, และ วิธีนิวตั้น

คำสำคัญ : ไมโครจินิติกอัลกอลิทึ่ม, การ dispatch หน่วยผลิตไฟฟ้าแบบประหยัด, Brute Force, จินิติกอัลกอลิทึ่มแบบธรรมดา, Merit Order Loading, วิธีนิวตั้น

1. Introduction

Economic dispatch (ED) is used to determine the optimal schedule of on-line generating outputs so as to meet the load demand at the minimum operating cost. The existing ED program, a standard function of the Energy Management System (EMS), National Control Center (NCC), Electricity Generating Authority of Thailand (EGAT), is applicable only for monotonically increasing incremental cost (IC) functions which is limited in flexibility. Neither linear decreasing IC functions nor staircase IC functions can be handled by the program [1].

At present, combined cycle (CC) units of EGAT system are always scheduled to serve the base load. However, during light load periods when there are no other generating units that can further reduce their outputs, there is a need to vary the CC units in an economical and smooth manner instead of a zero-one discrete basis. Therefore, if the linear decreasing IC and decreasing staircase IC functions (or non-convex and non-smooth input-output functions) of the CC units are included in the database, the conventional ED program based on the equal lambda methodology cannot determine the optimal solution [2].

Ongsakul [1,2] proposed the Merit Order Loading (MOL) method based on the unit lambda values at the highest operating outputs to solving the ED problem with linear decreasing IC and decreasing staircase IC of CC units. It was shown that the proposed MOL solution was close to the optimal solution and the real time implementation was valid on the existing EMS of EGAT without violating the ramp rate constraints. However, the monthly fuel costs of CC units of EGAT are in the order of several billion Thai baht, a few percent improvement of the existing ED program for CC units can lead to substantial fuel cost savings. Accordingly, this paper will investigate a Micro Genetic algorithm (MGA) to solving the ED with linear decreasing IC and decreasing staircase IC which are non-monotonically increasing functions.

Bakirtzis et al [3] proposed the Genetic algorithm to solving ED problem without convexity restrictions on the generator cost functions with valve point loading. The proposed method outperformed the dynamic programming in terms of computing times for the generating units ranging from 9 to 72. But the success rate to the optimal solution was still only 40% for the 72 generating unit system.

Wong et al [4] proposed the Genetic/Simulated-annealing approaches to ED problem. The algorithm was developed based on the combination of the incremental genetic-algorithm approach and the simulated-annealing technique. It was shown to be computationally faster than the earlier simulated annealing based method on a 13 generator practical system.

Sheble et al [5,6] proposed the Genetic algorithm (GA) and refined Genetic algorithm (RGA) to solving ED with valve point loading. Several techniques were explored to enhance the efficiency and accuracy such as mutation prediction, elitism, interval approximation and penalty factors. Test results however were shown on a very small size three-bus system.

Chen et al [7] proposed the large-scale ED by genetic algorithm with a normalized lambda encoding method. The network losses, ramp rate constraints, and prohibited zone were also taken into account. The proposed GA was tested on the system sizes ranging from 5 to 40 units. Even though this encoding scheme is practical for large scale implementation, its application is restricted to monotonically increasing IC functions only.

In this paper, a Micro Genetic algorithm (MGA) to solving the ED with linear decreasing IC and decreasing staircase IC functions of CC units is proposed. The ramp rate constraints are also taken into account to insure the feasibility of solutions. Two different encoding methods including concatenated and embedded methods are investigated. The MGA is tested and compared to Simple Genetic algorithm (SGA), MOL, Newton and unconstrained BF on a five CC unit system.

Two types of cost functions used for CC generating units at EGAT can be summarized as follows [1].
Second order polynomial cost function for a CC unit: One combined cycle unit consists of a series of single-cycle gas turbines in conjunction with a heat-recovery steam generator (HRSG). In a closed cycle operation mode, a cost function is obtained by curve-fitting among at least three test points using the least squares method. In fact, the second order polynomial cost function for a CC unit is estimated by $a + bP + cP^2$, where the coefficient $c < 0$, and IC is a linear decreasing function ($b+2cP$). This phenomenon is explained by the characteristics of a CC unit in such a way that the higher the output is, the better the efficiency will be.

Piecewise linear cost functions for a CC unit: To obtain extreme accuracy, instead of using only one quadratic equation to represent the cost function of a CC unit, several piecewise linear functions are employed to represent cost functions for each mode of the closed cycle operation. The piecewise linear cost functions and their decreasing staircase IC functions of one of the closed cycle modes are as follows.

$$C_i(P_i) = a_{i1} + b_{i1} P_i,\ P_{i,min} \le P_i < P_{i,int}, \quad (1a)$$
$$= a_{i2} + b_{i2} P_i,\ P_{i,int} \le P_i \le P_{i,max}, \quad (1b)$$
$$IC_i(P_i) = b_{i1},\quad P_{i,min} \le P_i < P_{i,int}, \quad (2a)$$
$$= b_{i2},\quad P_{i,int} \le P_i \le P_{i,max}, \quad (2b)$$

where $b_{i1} > b_{i2} > 0$.

The organization of the paper is as follows. The ED problem formulation is introduced in Section 2. Section 3 describes the SGA and MGA. The experimental results on a five unit system are given in Section 4. Conclusion is given in the last section.

2. Economic Dispatch Formulation

2.1 Basic Economic Dispatch Formulation

The conventional ED problem is to minimize the total cost function. The problem is formulated as:

$$\text{Minimize } C_T = \sum_{i=1}^{N} C_i(P_i), \quad (3)$$

subject to a power balance constraint:

$$\sum_{i=1}^{N} P_i = P_D + P_L(P_1,\ldots,P_N), \quad (4)$$

and operating limit constraints:

$$P_{i,min} \le P_i \le P_{i,max}, \qquad i = 1, \ldots, N, \quad (5)$$

where,

C_T = total fuel cost (baht/hr),

$C_i(P_i)$ = cost of the ith generating unit (baht/hr),

P_i = real power output of the ith generating unit (MW),

P_D = total load demand (MW),

$P_L(P_1,\ldots,P_N)$ = total transmission line loss (MW),

N = total number of on-line units to be dispatched,

$P_{i,min}$ = minimum power output of the ith unit (MW),

$P_{i,max}$ = maximum power output of the ith unit (MW).

2.2 Economic Dispatch Formulation with Ramp Rate Limits

Ramp rate of generating units are due to the fact that CC generating outputs can be not adjusted instantaneously. Therefore, to reflect the actual operating process, ED problem should include the ramp rate limits to ensure the feasibility of the solutions.

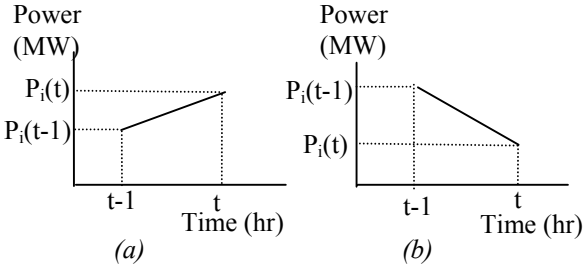


Figure 1. Two Possible Situations of On-line ith Unit

As shown in Figure 1, the inequality constraints of ramp rate limits are given as:

1. if the i*th* generation unit output increases (see Figure 1a)

$$P_i(t) - P_i(t\text{-}1) \le UR_i \quad (6)$$

2. if the i*th* generation unit output decreases (see Figure 1b)

$$P_i(t\text{-}1) - P_i(t) \le DR_i \quad (7)$$

Where UR_i and DR_i are up ramp and down ramp limits of the i*th* unit which are in the units of MW/hr. In practice, DR_i is greater than UR_i. Combining equations (3), (4), (5), (6) and (7), the constrained ED problem formulation becomes:

$$\begin{aligned} &\text{Minimize } C_T = \sum_{i=1}^{N} C_i(P_i(t)), \\ &\text{Subject to: } \sum_{i=1}^{N} P_i(t) = P_D(t) + P_L(P_1(t),\ldots,P_N(t)), \\ &P_{i,low}(t) \le P_i(t) \le P_{i,high}(t),\ i = 1, \ldots, N. \end{aligned} \quad (8)$$

Where,

$P_{i,low}(t)$ = the possible lowest power output of the ith unit at time t ($Max(P_{i,min}, P_i(t\text{-}1) - DR_i)$),

$P_{i,high}(t)$ = the possible highest power output of the i*th* unit at time t ($Min(P_{i,max}, P_i(t\text{-}1) + UR_i)$),

In this paper, the total power loss is neglected, thus $P_L(P_1(t)\ldots,P_N(t)) = 0$.

3. Genetic Algorithms

Genetic algorithm (GA) is essentially a searching method based on the concept of natural selection and natural genetics. GA searches on encoded bit strings (usually binary representations) called individuals rather than the real data points in solution space. GA has the ability to solve non-smooth, non-continuous, and non-differentiable cost functions which is not possible to obtain the optimal solution by a classical Lagrange method. GA uses the objective function to evaluate the performance, not its derivatives or auxiliary equations; and it has ability to exploit prior knowledge from previous solution guess to increase the performance of future solutions. Furthermore, GA exploits probability transition rules rather than deterministic rules [8].

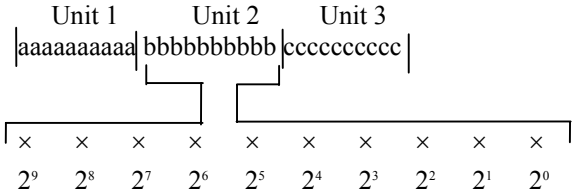
3.1 Simple Genetic Algorithm

In this paper, we shall use the Simple Genetic algorithm (SGA) developed by Goldberg [8] to solve the ED problem for comparison. The outputs of the N generating units have to satisfy the power balance constraint, operating limit constraints, and ramp rate constraints. For arbitrary free unit outputs P_i, $P_{i,low}(t) \le P_i(t) \le P_{i,high}(t)$, $i = 1,\ldots,N\text{-}1$, it is assumed that the N*th* reference unit power output is constrained by the power balance equation as:

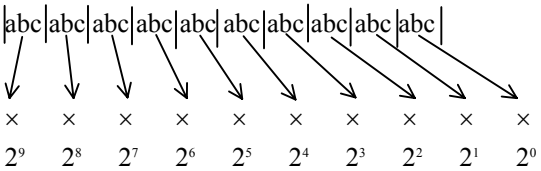
$$P_N(t) = P_D(t) - \sum_{i=1}^{N-1} P_i(t) \quad (9)$$

3.1.1 Encoding and Decoding

In this paper, concatenated and embedded encoding methods are explored.



(a) Concatenated Encoding



(b) Embedded Encoding

Figure 2. Two Different Encoding Schemes

As shown in Figure 2, each unit output of *N-1* free units is encoded in a binary based string normalized over its operating range. The concatenated encoding method stacks each unit's normalized string in series with each other to constitute the string individual. Each unit's string structure is assigned by the same number of *n* bits. On the other hand, the embedded encoding method employs the same binary system and encoding as the concatenated one. The only difference is that the assigned bit structures of each unit string are embedded within each other throughout the individual. Each individual consists of a series of smaller string structures (3 instead of 10). For both methods, a string individual has *n(N-1)* bits.

For a concatenated encoding example, three generating unit power outputs are encoded in a 30 bit string individual as:

0111111101 1011011011 0100011011

To obtain the actual generating power output of each unit for fitness function evaluation to be discussed, we need to decode each of 10 bit string to the decimal value by,

$$P_i(t) = P_{i,low}(t) + [B_i * (P_{i,high}(t) - P_{i,low}(t))/(2^n - 1)] \quad (10)$$

Where,

B_i = decimal integer value of converted binary string of the i*th* unit,

n = number of bits representing each unit output.

For example, the first generating unit power output, encoded by the first ten bits of the binary string individual, has UR_1 and DR_1 as 50 and 80 MW/hr respectively and suppose $P_1(t\text{-}1)$ is 230 MW, $P_{1,min}$ =100, and $P_{1,max}$ = 250. Then

$$P_{1,low}(t) = Max(100, 230-80) = 150 \text{ MW},$$
$$P_{1,high}(t) = Min(250, 230+50) = 250 \text{ MW},$$
$$B_1 = (2^9*0) + (2^8*1) + (2^7*1) + (2^6*1) + (2^5*1) + (2^4*1) + (2^3*1) + (2^2*1) + (2^1*0) + (2^0*1) = 509.$$

Therefore, $P_1(t) = 150 + [509*(250 - 150)/(2^{10} - 1)] = 199.76$ MW. In this paper, each free unit is represented by 15 bits. The more the number of bits, the higher the resolution.

SGA randomly generates initial *NP* (a specified population size) string individuals. For each generation (or iteration), SGA performs four basic operations: fitness function evaluation, reproduction, crossover, and mutation.

3.1.2 Fitness Function Evaluation

The performance in finding optimum solution of SGA is mainly related to the highest single fitness value. The fitness function including the power balance constraint of the *j*th individual is:

$$f_j = 0.5/Cst_j + 0.5/Pow_j, \; j=1,\ldots,NP \quad (11)$$

Where,

Pow_j $= 1 + (\Sigma_{i=1,N} P_i^j(t) - P_D(t))^2/P_D(t)$,

Cst_j $= 1 + k*(C_T(P_1^j(t),..,P_N^j(t)) - Cost_{min})^2$,

$P_i^j(t)$ = real power output of the i*th* unit of the *j*th individual at time *t*,

$Cost_{min}$ = minimum total fuel cost at $P_{i,min}$, $i=1,..,N$ (baht),

k = a constant value.

The fitness function evaluates a power balance difference and a total fuel cost difference. By experiment, k = 0.001 is used to scale down the total fuel cost difference squared component. Otherwise, the cost difference squared component will dominate the power balance difference squared over the total load demand component. After decoding the *j*th individual to $[P_1^j(t),\ldots,P_{N-1}^j(t)]$ and substituting in Eq. (9),

if $P_N^j(t) < P_{N,low}(t)$, $P_N^j(t) = P_{N,low}(t)$, and

if $P_N^j(t) > P_{N,high}(t)$, $P_N^j(t) = P_{N,high}(t)$.

We use $[P_1^j(t),..,P_N^j(t)]$ to evaluate the fitness value in Eq. (11) even though the power balance may not be satisfied. By using Eq. (11), the fitness value is ensured to be in the range of 0 to 1.

3.1.3 Reproduction

This reproduction method is based on the biased roulette wheel or "survival of the fittest" aspect [8]. The bigger the string individual fitness value, the higher the probability to have copies of them in the mating pool. For a population size *NP*, a reproduction probability and cumulative reproduction probability of the i*th* string individual with fitness value f_i, are

$$p_{rep,j} = \frac{f_j}{\sum_{j=1}^{NP} f_j}; \; Cumulative \; p_{rep,j} = \sum_{k=1}^{j} p_{rep,k}. \quad (12)$$

Where $j = 1,..,NP$. The selected parents' string individuals to be copied to the mating pool are those having cumulative $p_{rep,j}$ just above the real number randomized between 0 and 1.

3.1.4 Crossover

Crossover is a process of exchanging bits between two string individuals. In particular, two individuals from the mating pool are randomly selected as parent individuals based on the biased roulette wheel. Then arbitrary positions on both individuals are chosen for crossing locations, where the exchanges of bits take place. A crossing mask is employed to determine the crossing locations. Two parent individuals will exchange their bits at every location where the corresponding position in the mask is one. As an example, it is assumed that two 30 bit string parent individuals are selected for two-point and uniform crossover as follows:

```
Position:   1234567890 1234567890 1234567890
Parent1:    0111111101 1011011011 0100011011
Parent2:    1011110010 1111001100 1010011001
```

For the two-point crossover, the mask comprises one set of 1's bits surrounded by two sets of 0's bits [9]. Two crossing locations are arbitrarily selected from bit positions 2 to 30. If the crossing locations are 5 and 23, we will have

```
Position:   1234567890 1234567890 1234567890
Mask:       0000111111 1111111111 1110000000
Offspring1: 0111110010 1111001100 1010011011
Offspring2: 1011111101 1011011011 0100011001
```

For the uniform crossover, the mask comprises a string of 1's and 0's bit randomly distributed along its length [10]. The number of exchanged bits depends on the number of 1's bits in the mask varying from 0% to 50% of the total number of bits in a mask.

```
Position:  1234567890 1234567890 1234567890
Mask:      0100110111 0011101000 1001010010
Offspring1:0011111010 1011011011 1100011001
Offspring2:1111110101 1111001100 0010011011
```

The uniform crossover generates new offspring individuals to participate in genetic process. Without crossover, the fittest individual is obtained from the initial random population. However, were every individual to cross with another one after reproduction (crossover probability = 1.0), then we might lose many superior individuals. It has been shown in [10] that the convergence rate of uniform crossover is faster than the two-point crossover. Hence, the uniform crossover is used in this paper.

3.1.5 Mutation

Mutation is a process of flipping bits in a randomly chosen offspring string individual at random positions after performing crossover. In other words, it is a toggle from 0 to 1 or vice versa in a binary based system. Mutation is designed to give the offspring individual characteristics which do not exist in parent individuals. The rate of mutation is much lower than that of crossover since it is considered to be a secondary role. It is about 0.001 to 0.01 depending upon the types of applications. For instance, if we have offspring individual bits mutated at positions 2,15, and 21, then

```
Position:  1234567890 1234567890 1234567890
Mask:      0100000000 0000100000 1000000000
Offspring1:0111110010 1111001100 1010011011
                ↓          ↓          ↓
Offspring1:0011110010 1111101100 0010011011
```

Note the mutation should be employed with caution since the high mutation rate will deteriorate the search performance.

3.2 Micro Genetic Algorithm

Micro GA, one of the variation of the conventional SGA, was originally proposed by Krishnakumar [11]. In addition to performing four basic operations, MGA uses the elitism principle and checks the convergence of population at the end of each generation. Since any of the four basic operations do not guarantee that the new population string individuals are always better. Elitism guarantees that the best string individual survives. If the best individual in the current generation is worse

Table 1. Input-Output Characteristic of Rayong CC Units

RY unit no.	P_{min} (MW)	P_{max} (MW)	Fuel cost (baht/Gcal)	Input-output coefficient (GCal/hr)			Ramp rate (MW/hr)	
				A_i	B_i	C_i	UR_i	DR_i
1	100	300	273.8	-123.26930	3.1770742	-0.00275183	50	80
2	100	300	273.8	-127.79030	3.2410401	-0.00310412	40	70
3	100	150	273.8	-30.27881	2.2650570	-0.00092385	40	60
4	100	300	273.8	-127.79030	3.2410401	-0.00310412	40	70

Table 2. Input-Output Characteristic of Khanom CC Units

KN unit no.	P_{min} (MW)	P_{int} (MW)	P_{max} (MW)	Fuel cost (baht/Mbtu)	Input-output coefficient (Mbtu/hr)			Ramp rate (MW/hr)	
					A_i	B_i	C_i	UR_i	DR_i
1.1	376	495	-	79.413	450.5580	7.3659	-	40	50
1.2	-	495	678	79.413	820.2499	6.6173	-	40	50

than the previous generation, the best individual of previous generation will be randomly replaced to any individual of the current generation that would be parent individuals of the next generation. Consequently, MGA with elitism guarantees that at least the best individual exists until the last generation.

MGA checks the convergence after applying elitism. If the best string individual at each generation, which has the highest fitness value, has the total number of bit difference from the other individuals less than 5% of the total number of bits in the population size ($n\times(N\text{-}1)\times NP$), that population is converged. Then MGA needs to copy the best individual to the next generation whereas the other individuals are all re-initialized [11].

In general, MGA performs well on a very small size of population. By experiment, the mutation and uniform crossover probability are 0.00 and 0.50 respectively. The attractive aspect of MGA is that it requires a relatively smaller population size than the SGA which results in less computation time. Moreover, once the population is converged, MGA tries to explore another set of solutions which would result in a higher chance to bail the solutions out of the local optimal solutions.

4. Experimental Results

For cost functions data, four units of Rayong (RY) CC and one unit of Khanom (KN) CC power plants [2] are used as benchmark data because their IC functions are linear decreasing and decreasing staircase functions, respectively. The low and high operating limits, coefficients of input-output functions, ramp rates, as well as the gas fuel costs of RY and KN units are shown in Tables 1 and 2, respectively. It is noted that an input-output function is given as $A_i + B_iP_i + C_iP_i^2$ and the cost function is calculated by multiplying the input-output function by the fuel cost.

Table 3. Parameter Selection of SGA and MGA

Method	Pop. size	Crossover prob.	Mutation prob.	Maximum gen. limit
SGA with concatenated (SGAC)	90	0.5	0.02	1500
SGA with embedded (SGAE)	55	0.45	0.01	1500
MGA with concatenated (MGAC)	5	0.5	0	500
MGA with embedded (MGAE)	5	0.5	0	500

According to our experiments, the parameters selection which are suitable for our ED problem are shown in Table 3. These parameters include the population size, crossover probability, mutation probability, and maximum generation limits. SGA and MGA with concatenated and embedded encoding methods are investigated and compared. Note SGA retains the best individual for each generation and its solution is selected from the best individuals of all generations.
Computer programs for the SGA and MGA were developed in Pascal programming language. The experimental results obtained for increasing load demands from 776 to 1728 MW are compared to the unconstrained BF, MOL, and equal-lambda based

Newton results. The load demand step is 10 MW. The ramp rate constraints in MW/hr are taken into account for each step.

Table 4. Comparisons of Total Fuel Costs

Method	Total fuel costs (baht)	Total cost difference from BF (baht)	% total cost difference
BF	68,364,198	0	0
MGAC	68,484,219	120,021	0.1756
MGAE	68,489,662	125,464	0.1835
SGAC	68,521,126	156,928	0.2295
SGAE	68,521,520	157,322	0.2301
MOL Pmin	68,665,609	301,411	0.4409
MOL Pmax	68,750,475	386,277	0.5650
Newton	70,366,638	2,002,440	2.9291

As shown in Table 4, the percentage of total fuel cost difference of all load demand steps of MGA with concatenated and embedded encoding methods are 0.1756% and 0.1835% higher than the unconstrained BF whereas SGA with concatenated and embedded encoding methods are 0.2295% and 0.2301% higher. MGA with concatenated encoding method has a lower fuel cost than the MOL based on the unit lambda values at P_{min} and P_{max} results in [2] by 0.2653% and 0.3894%, respectively. This improvement of ED program would therefore result in substantial fuel cost savings.

We measure the performance of MGAC and MGAE by the total cost difference from the unconstrained BF. Therefore, it is difficult to determine the success rate if it is defined by how many times that we achieve the optimal solution from a certain number of runs. However, from 20 runs of MGAC and MGAE without ramp rate limits, the percentage average total cost difference from BF is 0.0993%, where its minimum and maximum total cost difference are 0.03% and 0.1631% respectively.

MGA average computing time is 2.42 seconds compared to 40 seconds of SGA on a Pentium 166 PC platform. This is because MGA requires a smaller size of population (5) than SGA (55 and 90). Moreover, the number of maximum generation limit of MGA (500) is smaller than SGA (1500) as the iteration processes of both methods continue until they reach their maximum generation limits. It should be observed that the unconstrained BF does not take the ramp rate constraints into account, thereby its solution is not guaranteed to be feasible. But we use its results for benchmarking. Note the equal-lambda based Newton has the highest fuel cost since the equal lambda principle is not applicable for linear decreasing and staircase IC functions [2].

5. Conclusions

A Micro Genetic algorithm (MGA) method to solving the ramp rate constrained economic dispatch problem with linear decreasing IC and staircase IC functions is proposed. The MGA outperforms SGA in terms of lower total fuel costs and faster computing times since it has a higher chance to obtain the optimal solution with a relatively smaller size of population. It is shown that the total fuel cost of MGA is less than MOL, leading to substantial fuel cost savings. Real-time implementation of ED for EGAT CC units with linear decreasing and staircase IC functions by MGA is potentially viable.

Acknowledgments

The authors would like to thank the National Electronics and Computer Technology Center (NECTEC) for their partial support under contract NT-38-06-CST-13-01. In addition, the authors would like to thank the Electricity Generating Authority of Thailand (EGAT) for the test data. Finally, the authors would like to thank Mr. Thakolsri for their assistance in programming work.

References

[1] W. Ongsakul. "Economic Dispatch with Linear Decreasing and Staircase Incremental Cost Units and Its Applications in Thailand, *Proc. of the 32nd Univ. Power Eng. Conf.*, Manchester, UK, Vol. 1, 1997, pp. 375-378.

[2] W. Ongsakul, "Real-Time Economic Dispatch Using Merit Order Loading for Linear Decreasing and Staircase Incremental Cost Functions", *Electric Power System Research*, 1999. (in press)

[3] A. Bakirtzis, V. Petridis, S. Kazarlis, "Genetic Algorithm Solution to the Economic Dispatch Problem", *IEE Proc.-Gen. Tran. Dist.*, Vol. 141, No. 4, July 1994, pp. 377-382.

[4] K. P. Wong, Y. W. Wong, "Genetic and Genetic/Simulated-Annealing Approaches to Economic Dispatch", *IEE Proc.-Gen. Trans. Dist.*, Vol. 141, No. 5, Sept. 1994, pp. 507-153.

[5] D. C. Walters, G. B. Sheble, "Genetic Algorithm Solution of Economic Dispatch with Valve Point Loading", *IEEE Trans. on Power Sys.*, Vol. 8, No. 3, Aug. 1993, pp. 1325-1332.

[6] G. B. Sheble, K. Brittig, "Refined Genetic Algorithm - Economic Dispatch Example", *IEEE Trans. on Power Sys.*, Vol. 10, No. 1, Feb. 1995, pp. 117-124.

[7] P. H. Chen, H. C. Chang, "Large-Scale Economic Dispatch by Genetic Algorithm," *IEEE Trans on Power Sys*, Vol. 10, No. 4, Nov. 1995, pp.1919-1926.

[8] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, 1989.

[9] L. Booker, "Improving Search in Genetic Algorithms," in 'Genetic Algorithms and Simulated Annealing' (Pitman, London, 1987), pp. 61-73.

[10] G. Syswerda, "Uniform Crossover in Genetic Algorithms," *Proceeding of the 3rd International Conferences on Genetic Algorithms*, June 1989, pp. 2-9.

[11] K. Krishakumar, "Micro-Genetic Algorithms for Stationary and Non-Stationary Function Optimization," *SPIE; Intelligent Control and Adaptive System*, Vol. 1196, Philadelphia, PA, 1989

Biographies

Dr. Weerakorn ONGSAKUL received his B.Eng. degree in Electrical Engineering from Chulalongkorn University, Bangkok, Thailand in 1988. He received his M.S. and Ph.D. degrees in Electrical Engineering from Texas A&M University in 1991 and 1994 respectively. He joined the department of Electrical Engineering, Sirindhorn International Institute of Technology, Thammasat University, Thailand in 1995. He is currently an assistant professor there. His current interests is in power system operation & control, computer applications to power systems, parallel processing applications, applications of genetic algorithms to power system optimization, and transmission pricing.

Mr. Jarurote TIPPAYACHAI received his B.Eng. degree in Electrical Engineering from Sirindhorn International Institute of Technology,Thammasat University, Thailand in 1999. He has been working as a research assistant for the department of Electrical Engineering since April 1999. His current interest is in applications of genetic algorithms to power generation economic dispatch.

NECTEC Technical Journal
70 Vol 1, No. 2, March 1999

A Statistical Grammar Acquisition Method Based on Clustering Analysis using a Bracketed Corpus

Thanaruk Theeramunkong
Department of Information Technology
Sirindhorn International Institute of Technology
Thammasat University, Pathumthani 12121, Thailand
ping@siit.tu.ac.th

ABSTRACT – This paper proposes a new method for learning a context-sensitive conditional probability context-free grammar from an unlabeled bracketed corpus based on clustering analysis and describes a natural language parsing model which uses a probability-based scoring function of the grammar to rank parses of a sentence. By grouping brackets in a corpus into a number of similar bracket groups based on their local contextual information, the corpus is automatically labeled with some nonterminal labels, and consequently a grammar with conditional probabilities is acquired. The statistical parsing model provides a framework for finding the most likely parse of a sentence based on these conditional probabilities. Experiments using Wall Street Journal data show that our approach achieves a relatively high accuracy: 88 % recall, 72 % precision and 0.7 crossing brackets per sentence for sentences shorter than 10 words, and 71 % recall, 51 % precision and 3.4 crossing brackets for sentences between 10-19 words. This result supports the assumption that local contextual statistics obtained from an unlabeled bracketed corpus are effective for learning a useful grammar and parsing.

KEY WORDS – Statistical Parsing, Grammar Acquisition, Clustering Analysis, Local Contextual Information

บทคัดย่อ – บทความนี้นำเสนอวิธีการเรียนกฎไวยากรณ์ที่มีค่าความน่าจะเป็นของการใช้กฎตามเนื้อความ (context) ที่อยู่รอบ การเรียนรู้นี้ใช้ชุดประโยค(corpus)ที่มีโครงสร้างของประโยคแต่ไม่มีการใส่ Label บอกว่าโครงสร้างนั้นคืออะไร นอกจากนี้บทความนี้ยังเสนอโมเดลการวิเคราะห์ประโยคเชิงโครงสร้าง (parsing) ที่ใช้กฎไวยากรณ์ที่คำนวณได้จากข้างต้น เพื่อทำการจัดลำดับความเป็นไปได้ของผลลัพธ์แต่ละอันที่ได้จากการวิเคราะห์โดยใช้ชุดประโยคภาษาอังกฤษขนาดใหญ่ของ Wall Street Journal มาใช้ในการทดลอง เราพบว่าวิธีการที่เสนอสามารถวิเคราะห์ประโยคได้อย่างถูกต้องสูงโดยมีระดับความถูกต้องอยู่ที่ 88 % recall 72 % precision และ 0.7 crossing brackets ต่อประโยค ในกรณีประโยคที่สั้นกว่า 10 คำ ส่วนสำหรับประโยคที่ยาว 10-19 คำ จะมีความถูกต้องอยู่ที่ระดับ 71 % recall, 51 % precision และ 3.4 crossing brackets ต่อประโยค ผลที่ได้นี้สนับสนุนสมมุติฐานที่ว่าเนื้อความใกล้ที่อยู่รอบ (local context) มีประโยชน์และประสิทธิผลในการช่วยให้สามารถเรียนรู้กฎไวยากรณ์อันจะเป็นประโยชน์ในการวิเคราะห์ประโยคด้วย

คำสำคัญ – การวิเคราะห์ประโยชน์ตามสถิติ, การเรียนรู้กฎไวยากรณ์, การวิเคราะห์โดยแบ่งกลุ่ม, ข้อมูลเนื้อความใกล้ที่อยู่รอบ

1. Introduction

Most natural language processing systems utilize grammars for parsing sentences in order to recognize their structure and finally to understand their meaning. Due to the difficulty and complexity of constructing a grammar by hand, there were several approaches developed for automatically training grammars from a large corpus with some probabilistic models. These methods can be characterized by properties of the corpus they used, such as whether it includes information of brackets, lexical labels, nonterminal labels and so on.

Recently several parsed corpora which include full bracketing, tagging and nonterminal labels have been available for researchers to use for constructing a probabilistic grammar [5][7][13] [14]. Most researches on these grammars calculate statistics of a grammar from a fully-parsed corpus with nonterminal labels and apply them to rank the possible parses of a sentence. While these researches report some promising results, it seems a hard task for a corpus builder to determine nonterminal labels for a corpus in comparison with annotating brackets and lexical labels, and the way to assign a nonterminal label to each constituent in the parsed sentence is usually ad hoc and arbitrary. From this point, it seems worth inferring a grammar from corpora without nonterminal labels. Moreover, compared with corpora including nonterminal labels, there are more existing corpora which include bracketings without nonterminal labels such as EDR corpus [8] and ATIS spoken language corpus [10]. The well-known standard method to infer a probabilistic context-free grammar from a bracketed/unbracketed corpus without nonterminal labels is so-called inside-outside algorithm which was originally proposed by Baker [2] and was implemented as applications for speech and language in [12], [16] and [18]. Although encouraging results were shown in these works, the derived grammars were restricted to Chomsky normal-form CFGs and there were problems of the small size of acceptable training corpora and the relatively high computation time required for training the grammars.

Towards the problems, this paper proposes a new method which can learn a standard CFG with less computational cost by adopting techniques of clustering analysis to construct a context-sensitive probabilistic grammar from a bracketed corpus where nonterminal labels are not annotated. Another claim of this paper is that statistics from a large bracketed corpus without nonterminal labels combined with clustering techniques can help us construct a probabilistic grammar which produces an accurate natural language statistical parser. In this method, nonterminal labels for brackets in a bracketed corpus can be automatically assigned by making use of local contextual information which is defined as a set of category pairs of left and right words of a constituent in the phrase structure of a sentence. In this research, based on the assumption that not all contexts are useful in every case, effectiveness of contexts is also investigated. By using only effective contexts, it is possible for us to improve training speed and memory space without a sacrifice of accuracy. Finally, a statistical parsing model based on the acquired grammar is provided and the performance is shown through some experiments using the WSJ corpus.

2. Grammar Acquisition as Clustering Process

In the past, Theeramunkong [20] proposed a method of grouping brackets in a bracketed corpus (with lexical tags but no nonterminal labels), according to their local contextual information, as a first step towards the automatic acquisition of a context-free grammar. The basic idea is to apply clustering analysis to find out a number of groups of similar brackets in the corpus and then to assign each group with a same nonterminal label. Clustering analysis is a generic name of a variety of mathematical methods that can be used to find out which objects in a set are similar. Its applications on natural language processing are varied such as in areas of word classification, text categorization and so on [11][17]. However, there is still few researches which apply clustering analysis for grammar inference and parsing [15]. This section gives an explanation of grammar acquisition based on clustering analysis. In the first place, let us consider the following example of the parse structures of two sentences in the corpus in Figure 1.

```
Sentence (1)    : A big man slipped on the ice.
Parse Tree (1) : (((DT,"a")((JJ,"big")(NN,"man")))
                  ((VB,"slipped")((IN,"on")
                  ((DT,"the")(NN,"ice")))))
Sentence (2)    : The boy dropped his wallet
                   somewhere.
Parse Tree (2) : (((DT,"the")(NN,"boy"))
                  (((VB,"dropped")((PRP\$,"his")
                    (NN,"wallet")))
                    (RB,"somewhere")))
```

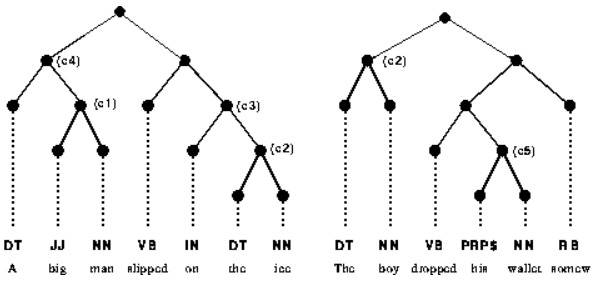


Figure 1. The graphical representation of the parse structures of "A big man slipped on the ice" and "the boy dropped his wallet somewhere"

In the parse structures, leaf nodes are given tags while there is no label for intermediate

nodes. Note that each node corresponds to a bracket in the corpus. With this corpus, the grammar learning task corresponds to a process to determine the label for each intermediate node. In other words, this task is concerned with the way to cluster the brackets into some certain groups based on their similarity and give each group a label. For instance, in figure 1, it is reasonable to classify the brackets (c2),(c4) and (c5) into a same group and give them a same label (e.g., NP(noun phrase)). As the result, we obtain three grammar rules: *NP → (DT)(NN), NP → (PRP$)(NN) and NP → (DT)(c1)*. To do this, the grammar acquisition algorithm operates in five steps as follows.

1. Assign a unique label to each node of which lower nodes are assigned labels. At the initial step, such node is one whose lower nodes are lexical categories. For example, in figure 1, there are three unique labels derived: *c1 → (JJ)(NN), c2 → (DT)(NN)* and *c5 → (PRP$)(NN)*. This process is performed throughout all parse trees in the corpus.
2. Calculate the similarity of every pair of the derived labels.
3. Merge the most similar pair to a single new label(i.e., a label group) and recalculate the similarity of this new label with other labels.
4. Repeat (3) until a termination condition is detected. Finally, a certain set of label groups is derived.
5. Replace labels in each label group with a new label in the corpus. For example, if *(DT)(NN)* and *(PRP$)(NN)* are in the same label group, we replace them with a new label (such as *NP*) in the whole corpus.
6. Repeat (1)-(5) until all nodes in the corpus are assigned labels.

To compute the similarity of labels, the concept of local contextual information is applied. In this work, the local contextual information is defined as categories of the words immediately before and after a label. This information is shown to be powerful for acquiring phrase structures in a sentence in [6]. In our preliminary experiments, we also found out that the information are potential for characterizing constituents in a sentence.

2.1 Distributional Similarity

While there are a number of measures which can be used for representing the similarity of labels in the step 2, measures which make use of relative entropy (Kullback-Leibler distance) are practical interest and scientific. One of these measures is divergence which has a symmetrical property. Its application on natural language processing was firstly proposed by Harris [9] and was shown successfully for detecting phrase structures in [6] [17]. Basically, divergence, as well as relative entropy, is not exactly similarity measure instead it indicates distributional dissimilarity. That means the large value it gets, the less similarity it means. The detail of divergence is illustrated below.

Let P_{c1} and P_{c2} be two probability distributions of labels *c1* and *c2* over contexts, *CT*. The relative entropy between P_{c1} and P_{c2} is:

$$D(P_{c1}||P_{c2}) = \sum_{e \in CT} p(e|c1) \times \log (p(e|c1)/p(e|c2))$$

Relative entropy $D(P_{c1}||P_{c2})$ is a measure of the amount of extra information beyond P_{c1} needed to describe P_{c2}. The divergence between P_{c1} and P_{c2} is defined as $D(P_{c1}||P_{c2})+D(P_{c2}||P_{c1})$, and is a measure of how difficult it is to distinguish between the two distributions. The context is defined as a pair of words immediately before and after a label(bracket). Any two labels are considered to be identical when they are distributionally similar, i.e., the divergence is low. From the practical point view, this measure addresses a problem of sparseness in limited data. Particularly, when $p(e|c2)$ is zero, we cannot calculate the divergence of two probability distributions because the denominator becomes zero. To cope with this problem, the original probability can be modified by a popular technique into the following formula.

$$p(e|ci) = \lambda(N(ci,e)/N(c\_i)) + (1-\lambda)(1/|CT|)$$

where, *N(ci)* and *N(ci,e)* are the occurrence frequency of *ci* and *(ci,e),* respectively. $|CT|$ is the number of possible contexts and λ is an interpolation coefficient. As defining contexts by the left and right lexical categories, $|CT|$ is the square of the number of existing lexical categories. In the formula, the first term means the original estimated probability and the second term expresses a uniform distribution, where the probability of all events is estimated to a fixed uniform number. λ is applied as a balancing weight between the observed distribution and the uniform distribution. In our experimental results, λ is assigned with a value of 0.6 which seems to make a good estimate.

2.2 Termination Condition

During iteratively merging the most similar labels, all labels will finally be gathered to a single group. Due to this, a criterion is needed for determining whether this merging process should be continued

or terminated. In this section, we describe a criterion named *differential entropy* which is a measure of entropy (perplexity) fluctuation before and after merging a pair of labels. Let *c1* and *c2* be the most similar pair of labels. Also let *c3* be the result label. *p(e|c1), p(e|c2)* and *p(e|c3)* are probability distributions over contexts *e* of *c1, c2* and *c3*, respectively. *p(c1), p(c2)* and *p(c3)* are estimated probabilities of *c1, c2* and *c3*, respectively. The differential entropy (*DE*) is defined as follows.

$$\begin{aligned} DE &= \text{Consequence Entropy - Previous Entropy} \\ &= -p(c3) \times \sum_e p(e|c3) \log p(e|c3) \\ &\quad + p(c1) \times \sum_e p(e|c1) \log p(e|c1) \\ &\quad + p(c2) \times \sum_e p(e|c2) \log p(e|c2) \end{aligned}$$

where $\sum_e p(e|ci) \log p(e|ci)$ is the total entropy over various contexts of label *ci*. The larger *DE* is, the larger the information fluctuation before and after merging becomes. In general, a small fluctuation is preferred to a larger one because when *DE* is large, the current merging process introduces a large amount of information fluctuation and its reliability becomes low.

3. Local Context Effective

As the similarity of any two labels is estimated based on local contextual information which is defined by a set of category pairs of left and right words, there is an interesting question of which contexts are useful for calculation of similarity. In the past, effectiveness of contexts is indicated in some previous researches [3]. One of suitable measures for representing effectiveness of a context is dispersion of the context on labels. This measure expresses that the number of useful contexts should be diverse for different labels. From this, the effectiveness (*E*) of a context (*c*) can be defined using variance as follow:

$$\begin{aligned} E(c) &= \sum_{a \in A} ((N(a,c) - N^*(c))^2 / |A| \\ N^*(c) &= \sum_{a \in A} N(a,c) / |A| \end{aligned}$$

where *A* is a set of all labels and *a* is one of its individual member. *N(a,c)* is the number of times a label *a* and a context *c* are cooccurred. *N*(c)* is an averaged value of *N(a,c)* on a label *a*. In order to take large advantage of context in clustering, it is preferable to choose a context c with a high value of E(c) because this context trends to have a high discrimination for characterizing labels. Ranking the contexts by the effectiveness value *E*, some rank higher contexts are selected for clustering the labels instead of all contexts. This enables us to decrease computation time and space without sacrificing the accuracy of the clustering results and sometimes also helps us to remove some noises due to useless contexts. Some experiments were done to support this assumption and their results are shown in the next section.

4. Statistical Parsing Model

This section describes a statistical parsing model which takes a sentence as input and produce a phrase-structure tree as output. In this problem, there are two components taken into account: a statistical model and parsing process. The model assigns a probability to every candidate parse tree for a sentence. Formally, given a sentence *S* and a tree *T*, the model estimates the conditional probability *P(T|S)*. The most likely parse under the model is $argmax_T P(T|S)$ and the parsing process is a method to find this parse. While a model of a simple probabilistic CFG applies the probability of a parse which is defined as the multiplication of the probability of all applied rules, however, for the purposes of our model where left and right contexts of a constituent are taken into account, the model can be defined as follow.

$$P(T|S) = \Pi_{(ri,ci) \in T} \; P(ri,ci)$$

where *ri* is an application rule in the tree and *ci* is the left and right contexts at the place the rule is applied. Similar to most probabilistic models and our clustering process, there is a problem of low-frequency events in this model. Although some statistical NL applications apply backing-off estimation techniques to handle low-frequency events, our model uses a simple interpolation estimation by adding a uniform probability to every events. Moreover, we make use of the geometric mean of the probability instead of the original probability in order to eliminate the effect of the number of rule applications as done in [13]. The modified model is:

$$P(T|S) = (\Pi_{(ri,ci) \in T} (\alpha * P(ri,ci) + (1-\alpha) * (1/(NrNc)))^{1/|T|}$$

Here, α is a balancing weight between the observed distribution and the uniform distribution and it is assigned with 0.95 in our experiments. The applied parsing algorithm is a simple bottom-up chart parser whose scoring function is based on this model. The grammar used is one trained by the algorithm described in section 2. A dynamic programming algorithm is used: if there are two proposed constituents which span the same set of words and have the same label, then the lower probability constituent can be safely discarded.

5. Experimental Evaluation

To give some support to our suggested grammar acquisition method and statistical parsing model, three following evaluation experiments are made. The experiments use texts from the Wall Street Journal (WSJ) Corpus and its bracketed version provided by the Penn Treebank. Out of nearly 49,000 sentences (1,222,065 words), we extracted 48,000 sentences (1,172,710 words) as possible material source for training a grammar and 2000 sentences (49,355 words) as source for testing.

The first experiment involves an evaluation of performance of our proposed grammar learning method shown in the section 2. In this preliminary experiment, only rules which have lexical categories as their right hand side are considered and the acquired nonterminal labels are compared with those assigned in the WSJ corpus. The second experiment stands for investigating effectiveness of contexts described in section 3. The purpose is to find out useful contexts and use them instead of all contexts based on the assumption that not all contexts are useful for clustering brackets in grammar acquisition. Reducing the number of contexts will help us to improve the computation time and space. The last experiment is carried out for evaluating the whole grammar which is learned based on local contextual information and indicating the performance of our statistical parsing model using the acquired grammar. The measures used for this evaluation are bracketing recall, precision and crossing.

5.1 Evaluation of Clustering in Grammar Acquisition

This subsection shows some results of our preliminary experiments to confirm effectiveness of the proposed grammar acquisition techniques. The grammar is learned from the WSJ bracketed corpus where all nonterminals are omitted. In this experiment, we focus on only the rules with lexical categories as their right hand side. For instance, *c1 →(JJ)(NN), c2→(DT)(NN)* and *c5→(PRP$)(NN)* in figure 1. Due to the reason of computation time and space, we use the rule tokens which appear more than 500 times in the corpus. The number of initial rules is 51. From these rules, the most similar pair is calculated and merged to a new label. The merging process is carried out in iterative way. In each iterative step of the merging process, differential entropies are calculated. During the merging process, there are some sharp peaks indicating the rapid fluctuation of entropy. These sharp peaks can be used as a step to terminate the merging process. In the experiments, a peak with $DE > 0.12$ is applied. As the result, the process is halted up at the 45th step and 6 groups are obtained.

This result is evaluated by comparing the system's result with nonterminal symbols given in the WSJ corpus. The evaluation method utilizes a contingency table model which is introduced in [19] and widely used in Information Retrieval and Psychology [1][11]. The following measures are considered.

- Positive Recall (PR) : $a/(a+c)$
- Positive Precision (PP) : $a/(a+b)$
- Negative Recall (NR) : $d/(b+d)$
- Negative Precision (NP) : $d/(c+d)$
- F-measure(FM) : $((\beta^2+1)\times PP\times PR)/(\beta^2\times PP+PR)$

where a is the number of the label pairs which the WSJ corpus assigns in the same group and so does the system, b is the number of the pairs which the WSJ corpus does not assign in the same group but the system does, c is the number of the pairs which the WSJ assigned but the system does not, and d is the number of the pairs which both the WSJ and the system does not assign in the same group. The F-measure is used as a combined measure of recall and precision, where β is the weight of recall relative to precision. Here, we use $\beta = 1.0$, equal weight.

The result shows 0.93 % PR, 0.93 % PP, 0.92 % NR, 0.92 % NP and 0.93 % FM, which are all relatively good values. Especially, PP shows that almost all same labels in the WSJ are assigned in same groups. In order to investigate whether the application of differential entropy to cut off the merging process is appropriate, we plot values of these measures at all merging steps as shown in figure 2. From the graphs, we found out that the best solution is located at around 44th-45th merging step. This is consistent with the grouping result of our approach. Moreover, the precision equals 100 % from 1st-38nd steps, indicating that the merging process is suitable.

5.2 Checking Context Effectiveness

As another experiment, we examine effectiveness of contexts in the clustering process in order to reduce the computation time and space. Variance is used for expressing effective of a context. The assumption is that a context with has the highest variance is the most effective. The experiment is done by selecting the top N of contexts and use it instead of all contexts in the clustering process.

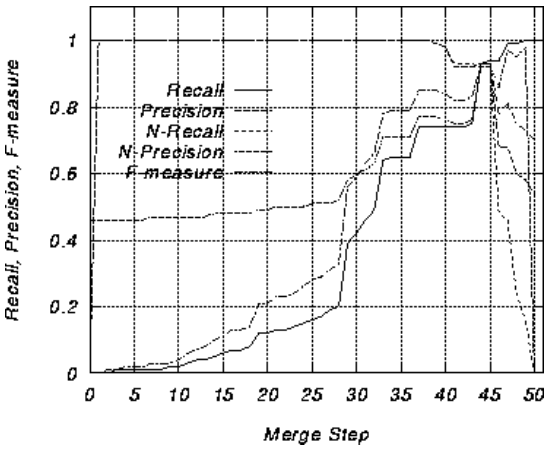


Figure 2. The transition of PR, PP, NR, NP and FM during the merging process

Besides cases of N = 10, 50, 200, 400 and all (2401), a case that 200 contexts are randomly chosen from all contexts, is taken into account in order to examine the assumption that variance is efficient. In this case, 3 trials are made and the average value is employed. Due to the limit of paper space, we show only F-measure in Figure 3. The graphs tell us that the case of top 200 seems superior to the case of 200 random contexts in all merging step. This means that variance seems to be a good measure for selecting a set of effective contexts in the clustering process.

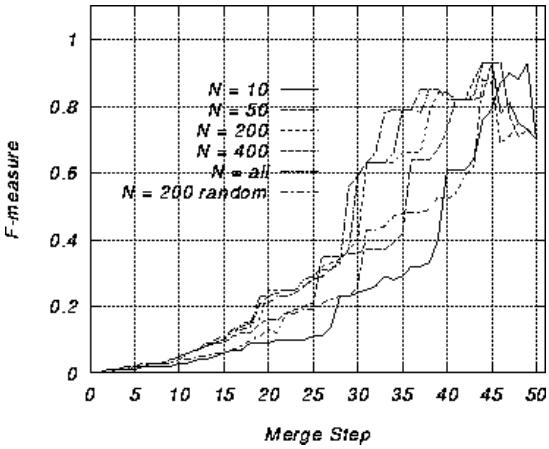


Figure 3. The transition of PR, PP, NR, NP and FM during the merging process

Furthermore, we can observe that a high accuracy can be achieved even if not all contexts are taken into account. From this result, the best F-measures are all 0.93 and the number of groups are 2, 5, 5 and 6 for each case, i.e., 10, 50, 200 and 400. Excepts the case of 10, all cases shows a good result compared with all contexts (0.93, 6 groups). This result tells us that it is reasonable to select contexts with large values of variance to ones with small variance and a relatively large number of contexts are enough for the clustering process. By preliminary experiments, we found out that the following criterion is sufficient for determining the number of contexts. Contexts are selected in the order of their variance and a context will be accepted when its variance is more than 10 % of the average variance of the previous contexts.

5.3 Performance of Statistical Parsing Model

Utilizing top N contexts, we learn the whole grammar based on the algorithm given in section 2. Brackets (rules) which are occurred more than 40 times in the corpus are considered and the number of contexts used is determined by the criterion described in the previous subsection. As the result of the grammar acquisition process, 1396 rules are acquired. These rules are attached with the conditional probability based on contexts (the left and right categories of the rules). The chart parser tries to find the best parse of the sentence. 48,000 sentences are used for training a grammar and 2000 sentences are for a test set. To evaluate the performance, the PARSEVAL measures as defined in [4] are used:

$$\text{Precision} = \frac{\text{number of correct brackets in proposed parses}}{\text{number of brackets in proposed parses}}$$

$$\text{Recall} = \frac{\text{number of correct brackets in proposed parses}}{\text{number of brackets in treebank parses}}$$

The parser generates the most likely parse based on context-sensitive condition probability of the grammar. Among 2000 test sentences, only 1874 sentences can be parsed owing to two following reasons: (1) our algorithm considers rules which occur more than 40 times in the corpus, (2) test sentences have different characteristics from training sentences. Table 1 displays the detail results of our statistical parser evaluated against the WSJ corpus.

93 % of sentences can be parsed with 71 % recall, 52 % precision and 4.5 crossings per sentence. For short sentences (3-9 words), the parser achieves up to 88 % recall and 71 % precision with only 0.71 crossings. For moderately long sentences (10-19 and 20-30 words), it works with 60-71 % recall and 41-51 % precision. From this result, the proposed parsing model is shown to succeed with high bracketing recalls to some degree. Although our parser cannot achieve good precision, it is not so a serious problem because our

parser tries to give more detail bracketing for a sentence than that given in the WSJ corpus. In the next section, the comparison with other researches will be discussed.

Table 1. Parsing accuracy using the WSJ Corpus

Sent. Length	3-9	3-15	10-19	20-30	3-40
Comparisons	393	988	875	484	1862
Avg. Sent. Len.	7.0	10.3	14.0	24.0	16.33
TBank Parses	4.81	6.90	9.37	15.93	10.85
System's Parse	10.86	16.58	23.14	40.73	27.18
Crossings/Sent.	0.72	1.89	3.36	7.92	4.52
Sent. Cross=0 (%)	56.7	33.1	13.6	2.5	19.0
Sent. Cross≤1 (%)	79.4	50.4	25.4	6.0	30.3
Sent. Cross≤2 (%)	93.4	67.0	41.5	9.5	41.8
Recall	88.2	79.3	71.2	59.7	70.8
Precision	71.9	60.6	51.3	41.2	52.1

6. Related Works and Discussion

In this section, our approach is compared with some previous interesting methods. These methods can be classified into non-grammar-based and grammar-based approaches. For non-grammar-based approaches, the most successful probabilistic parser named SPATTER is proposed by Magerman [14]. The parser is constructed by using decision-tree learning techniques and can succeed up to 86-90 % of bracketing accuracy(both recall and precision) when training with the WSJ corpus, a fully-parsed corpus with nonterminal labels. Later Collins [7] introduced a statistical parser which is based on probabilities of bigram dependencies between head-words in a parse tree. At least the same accuracy as SPATTER was acquired for this parser. These two methods utilized a corpus which includes both lexical categories and nonterminal categories. However, it seems a hard task to assign nonterminal labels for a corpus and the way to assign a nonterminal label to each constituent in the parsed sentence is arduous and arbitrary. It follows that it is worth trying to infer a grammar from corpora without nonterminal labels.

One of the most promising results of grammar inference based on grammar-based approaches is the inside-outside algorithm proposed by [12] to construct the grammar from unbracketed corpus. This algorithm is an extension of forward-backward algorithm which infers the parameters of a stochastic context-free grammar. In this research the acquired grammar is evaluated based on its entropy or perplexity where the accuracy of parsing is not taken into account. As another research, Pereira and Schabes [16][18] proposed a modified method to infer a stochastic grammar from a partially parsed corpus and evaluated the results with a bracketed corpus. This approach gained up to around 90 % bracketing recall for short sentences(0-15 words) but it suffered with a large amount ambiguity for long ones(20-30) where 70 % recall is gained. The acquired grammar is normally in Chomsky normal-form which is a special case of grammar although he claimed that all of CFGs can be in this form. This type of the grammar makes all output parses of this method be in the form of binary-branching trees and then the bracketing precision cannot be taken into account because correct parses in the corpus need not be in this form. On the other hand, our proposed approach can learn a standard CFG with 88 % recall for short sentences and 60 % recall for long ones. This result shows that our method gets the same level of accuracy as the inside-outside algorithm does. However, our approach can learn a grammar which is not restricted to Chomsky normal-form and performs with less computational cost compared with the approaches applying the inside-outside algorithm.

7. Conclusion

In this paper, we proposed a method of applying clustering analysis to learn a context-sensitive probabilistic grammar from an unlabeled bracketed corpus. Supported by some experiments, local contextual information which is left and right categories of a constituent was shown to be useful for acquiring a context-sensitive conditional probability context-free grammar from a corpus. A probabilistic parsing model using the acquired grammar was described and its potential was examined. Through experiments, our parser can achieve high parsing accuracy to some extent compared with other previous approaches with less computational cost. As our further work, there are still many possibilities for improvement which are encouraging. For instance, it is possible to use lexical information and head information in clustering and constructing a probabilistic grammar.

Reference

[1] Agarwal, Rajeev, "Evaluation of semantic clusters", *Proceedings of 33rd Annual Meeting of the ACL* , 1995, pp. 284-286

[2] Baker, J.K., "Trainable grammars for speech recognition", *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America (Klatt, D.H. and Wolf, J.J. eds.)*, 1979, pp. 547-550

[3] Bartell, B.T., Cottrell, G.W. and Belew, R.K., "Representing documents using an explicit model of their similarities", *Journal of the American Society for Information Science*, Vol. 46, No. 4, 1995, pp. 254-271
[4] Black, E. and al., et, "A procedure for quantitatively comparing the syntactic coverage of English grammars", *Proceedings of 1991 DARPA Speech and Natural Language Workshop*, 1991, pp. 306-311
[5] Black, E., Jelinek, F., Lafferty, J., Magerman, D.M., Mercer, R. and Roukos, S., "Towards history-based grammars: Using richer models for probabilistic parsing", *Proceedings of 1992 DARPA Speech and Natural Language Workshop*, 1992, pp. 134-139
[6] Brill, E., "Automatically acquiring phrase structure using distributional analysis", *Proceedings of Speech and Natural Language Workshop*, 1992, pp. 155-159
[7] Collins, M.J., "A new statistical parser based on bigram lexical dependencies", *Proceedings of 34th Annual Meeting of the ACL* , 1996, pp. 184-191
[8] EDR, "EDR Electronic Dictionary User's Manual" (in Japanese), *Japan Electronic Dictionary Research Institute*, 1994, Ed. 2.1
[9] Harris, Z., "Structural Linguistics", Chicago, University of Chicago Press, 1951
[10] Hemphill, C.T., Godfrey, J.J., Doddington, G.R., "The ATIS spoken language systems pilot corpus", *Proceedings of 1990 DARPA Speech and Natural Language Workshop*, 1990
[11] Iwayama, M. and Tokunaga, T., "Hierarchical bayesian clustering for automatic text classification", *IJCAI*, 1995, pp. 1322-1327
[12] Lari, K. and Young, S.J., "The estimation of stochastic context-free grammars using the inside-outside algorithm", *Computer speech and recognition*, Vol. 4, 1990, pp. 35-56
[13] Magerman, D.M. and Marcus, M.P., "Pearl: a probabilistic chart parser", *Proceedings of the European ACL Conference*, 1991
[14] Magerman, D.M., "Statistical decision-tree models for parsing", *Proceedings of 33rd Annual Meeting of the ACL* , 1995, pp. 276-283
[15] Mori, S. and Nagao M., "Parsing without grammar", *Proceedings of the 4th International Workshop on Parsing Technologies*, 1995, pp. 174-185
[16] Pereira, F. and Schabes, Y., "Inside-outside reestimation from partially bracketed corpora", *Proceedings of 30th Annual Meeting of the ACL* , 1992, pp. 128-135
[17] Pereira, F., Tishby, N. and Lee, L., "Distributional clustering of English words", *Proceedings of 31st Annual Meeting of the ACL* , 1993, pp. 183-190
[18] Schabes, Y., Roth M. and Osborne, R., "Parsing the Wall Street Journal with the inside-outside algorithm", *Proceedings of 6th European Chapter of ACL* , 1993, pp. 341-347
[19] Swets, J.A., "Effectiveness of information retrieval methods", *American Documentation*, Vol. 20, 1969, pp. 72-89
[20] Theeramunkong, T. and Okumura, M., "Towards automatic grammar acquisition from a bracketed corpus", *Proceedings of 4th International Workshop on Very Large Corpora*, 1996, pp. 168-177

Thanaruk THEERAMUNKONG
Born in 1967. He received a bachelor degree in Electric and Electronics and master and doctor degrees in Computer Science from Tokyo Institute of Technology in 1990, 1992, and 1995 respectively. From 1995 to 1997, he worked as research associates at Graduate School of Information Science, Japan Advanced Institute of Science and Technology. At present, he is working as assistant professor at Sirindhorn International Institute of Technology. His current research interests include natural language processing, machine learning, corpus-based processing, information retrieval and database technology.

Development of the Public Internet Exchange (PIE) in Thailand

Chalermpol Charnsripinyo, Angkana Angkalukkana,
Thaweesak Koanantakool, Pornthep Narula

Abstract --An Internet exchange is defined as a high-speed network or switch serving as an interconnection point for traffic exchange among Internet Service Providers (ISPs). Many countries have tried to set up their local Internet exchanges to reduce the operating cost of ISPs due to the expensive international links.

This paper describes development of the Public Internet Exchange (PIE) which is one of the two local Internet exchanges in Thailand. All aspects of PIE including history, current status, architecture, sample traffic statistics, problems, and policy are mentioned in details.

keywords: Internet Exchange, Thailand, PIE, Public Internet Exchange, Internet, Internet Service Provider, NECTEC, CAT

บทคัดย่อ-- ศูนย์แลกเปลี่ยนข้อมูลอินเทอร์เน็ต หมายถึงเครือข่ายหรืออุปกรณ์ที่ทำหน้าที่เป็นจุดเชื่อมต่อระหว่างผู้ให้บริการอินเทอร์เน็ตต่างๆ เพื่อให้สามารถทำการรับ-ส่งข้อมูลระหว่างกันได้ ซึ่งหลายประเทศได้พยายามจัดตั้งศูนย์แลกเปลี่ยนข้อมูลอินเทอร์เน็ตภายในประเทศขึ้นเพื่อช่วยลดต้นทุนในการดำเนินการของผู้ให้บริการอินเทอร์เน็ตในด้านการใช้งานวงจรสื่อสารระหว่างประเทศ

รายงานฉบับนี้เป็นการอธิบายถึงการพัฒนา ศูนย์แลกเปลี่ยนข้อมูลอินเทอร์เน็ตสาธารณะซึ่งเป็นหนึ่งในสองศูนย์แลกเปลี่ยนข้อมูลอินเทอร์เน็ตของประเทศไทย โดยจะได้กล่าวถึงความเป็นมา สถานะปัจจุบัน สถาปัตยกรรมของระบบ ตัวอย่างสถิติการใช้งานอุปสรรคและวิธีการป้องกันและแก้ไข รวมทั้งนโยบายการพัฒนาและบริการโดยละเอียด

คำสำคัญ: ศูนย์แลกเปลี่ยนข้อมูลอินเทอร์เน็ต, ประเทศไทย, PIE, ศูนย์แลกเปลี่ยนอินเทอร์เน็ตสาธารณะ, อินเทอร์เน็ต, ผู้ให้บริการอินเทอร์เน็ต, ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ, การสื่อสารแห่งประเทศไทย

Introduction

The Internet is known as a worldwide computer network. It consists of networks of networks connected together with various media and protocols. Almost every country that has Internet connections has dedicated circuits to different ISPs in the United States. Therefore, data traffic exchange among many networks happens in the US. This is the burden of ISPs due to the cost of international links. To save operating costs, many countries are currently running their own exchange points for local traffic exchange.

The situation of the Internet development in Thailand was quite similar to several other developing countries. Most Thai ISPs have their own overseas links to the US. The intra-country traffic among networks in Thailand had to go through exchange points in the US that increased the operating cost of Thai ISPs. This resulted in the set up of local Internet exchange points in Thailand. The success story of Internet development and the setting up of the Public Internet Exchange in Thailand will be given in this paper.

The history of Internet development in Thailand

Development of the Internet in Thailand began with a group of university professors who would like to keep contact with friends and researchers in other countries. In 1987, the Prince of Songkhla University (PSU) set up the first email connectivity to the University of Melbourne, Australia.

At the same time, the Asian Institute of Technology (AIT) tested UUCP connections to the University of Melbourne and the University of Tokyo via the Communication Authority of Thailand (CAT) X.25 service. All initial

connections were based on dial-up links to those universities. In early 1988, the Thai Computer Science Network (TCSNet) was founded with support from the Australian government, through the Australian International Development Plan (IDP). Three universities that are PSU, AIT and Chulalongkorn University (CU), initially joined the TCSNet. Later on In 1991, Thammasat Unversity (TU) installed MHSNet software and a 14.4 kbps modem with support from the Australian Academic and Research Network (AARNet) and became a new gateway for Thailand's academic and research institutions to communicate with University of Melbourne. The electronic mail service via MHSNet and UUCP became an important tool for Thai academics at the time.

Apart from the aforementioned universities, the National Electronics and Computer Technology Center (NECTEC) was running another Inter-University Network over X.25. The Inter-University Network project was identified since 1987 as an academic and research network focusing on the key issues of telecommunication infrastructures and databases. In January 1992, an electronic mail committee called NEWgroup (NECTEC's Email Working Group) was set up to represent demanding users to interact with NECTEC. This resulted in the permanent set up of the Thai Social/Scientific, Academic and Research Network (ThaiSarn) which was the merger of TCSNet and the Inter-University Network.

ThaiSarn was initially funded by the national budget via NECTEC and was technically supported by NECTEC's in-house Network Technology Laboratory (NTL), founded in April 1992 to look after the services and technical aspects of this network. Since the world-wide Internet were mainly based on the Internet Protocol (IP) technology and had very high growth rate at the time, NECTEC decided that ThaiSarn should also migrate to full Internet Protocol. Therefore, the early-day MHSNet and UUCP dial-up links as well as X.25 were gradually replaced by full fledged TCP/IP links on leased lines, while new nodes were added. The first 9.6kbps international gateway was the link between CU and UUNET in 1992. Shortly after that, another 64 kbps international line was launched between NECTEC, the hub of ThaiSarn, and UUNET.

All institutions linked to ThaiSarn have to abide by ThaiSarn Acceptable Use Policy (ThaiSarn AUP). Not only academic institutions but also many government agencies used ThaiSarn services. In 1994, some schools began to test Internet connections. Prototype nodes for schools started connecting to ThaiSarn. This resulted in the set up of the SchoolNet project by NECTEC in 1995, Thailand's IT Year.

Internet on the nonacademic side also began in 1995. After six months of feasibility study by a joint working group, the Communications Authority of Thailand (CAT) and the Telephone Organization of Thailand (TOT) approved a joint venture proposal from NECTEC to commercialize of the Internet in Thailand. Internet Thailand Company Limited, jointly invested by NECTEC, CAT, and TOT, received the first operating licence from CAT. The Internet Thailand Company launched its full-scale services with its first 512 kbps international gateway through UUNET, VA, USA in March 1995. Almost at the same time, KSC Comnet also received an operating licence from CAT and became Thailand's second commercial Internet Service Provider (ISP). A few months later, CAT approved proposals from Loxley Information and other two companies for operating license. The Internet usage and connectivity had continued to grow since then. A number of ISPs was later established to provide Internet services nationwide for both corporate and individual subscribers.

Current status of Internet in Thailand

As of October 1998, ThaiSarn, the biggest academic and research network in Thailand, has been expanded to cover almost all state universities nationwide. There are 66 nodes connected to ThaiSarn. Major universities are connected with 2 Mbps (E1) links and many institutions are connecting with 64 - 512 kbps links. The total domestic bandwidth of ThaiSarn is currently above 17 Mbps. ThaiSarn has a 2.5 Mbps gateway to the global Internet via Internet Thailand, and another 2 Mbps link to the National Center for Scientific Information System (NACSIS), Japan, for communication between ThaiSarn and SINET, Japanese academic network. Apart from ThaiSarn, some universities also have direct domestic and international links to ISPs.

The SchoolNet@1509 project that provides free dial-up Internet access to schools has been expanded to cover schools nationalwide. With support from TOT and CAT, NECTEC finished the initial phase of SchoolNet project with 20 POPs in March 1998. There are more than 80 schools using the service. With minimum of personnel and resource, the project started with the philosophy of "equal" minimum accessibility for all. The network will later be upgraded to meet the real demand. It is expected to reach 2500 schools by the end of 1999.

On the non-academic side, Thailand currently has fifteen ISPs operating nationwide. Most of them have their own international links to their upstream ISPs in the US. In July 1996, CAT started providing its own wholesale Internet service, called the International Internet Gateway (IIG), which resells its international Internet bandwidths to local ISPs who cannot afford their own international links. Some small ISPs use CAT's IIG service to save the cost of operations. Nevertheless, most ISPs still prefer to have their own international lines due to reliability and competition. The combined international link capacity is above 30 Mbps (see http://www.nectec.or.th/inet-map/).

Setting up of Internet Exchanges in Thailand

Before the existence of local Internet exchange points in Thailand, all ISPs and their customers have to communicate with each other via their overseas links. Almost all international lines are linked to different ISPs in the United States. With limited bandwidths of international links that is expensive and relatively slow, the burden from intra-country traffic via overseas links was apparent.

Thailand National Internet Exchange (NIX) was first set up by CAT for exchanging domestic traffic among Thai ISPs after CAT started operating the Internet IIG) service in July 1996. IIG and NIX were linked locally by a 10 Mbps Ethernet cable and later upgraded to a 100 Mbps link. Major ISPs have at least a 512 kbps link to NIX and pay CAT a monthly maintenance fee for the service.

The ThaiSarn Public Internet Exchange (PIE) was set up by NECTEC in November 1997. It was set up as a peering point for Thai ISPs to access public information on the ThaiSarn Public access Network (PubNet), a network of national servers such as FTP, Web, Gopher, Cache, News, Media servers, etc. In addition, it plays a significant role as the alternative exchange point with CAT's NIX to alleviate the single point of failure. Since all ISPs are required to use CAT leased-lines to connect to NIX while PIE allows ISPs to use any other circuit providers' services, such as TOT, TelecomAsia (TA), etc., a failure in one of the telecom infrastructure will not paralyze the others.

PIE is operated by NTL and located at NECTEC's network hub that permits broadband interconnection to PubNet, ThaiSarn, SchoolNet@1509, Kanchanapisek Network, and government networks. In the first year, NECTEC fully subsidizes the operating cost including electricity and a rack space at PIE. NTL assures that all PIE participants are given equal opportunities to access all information servers and networks connected to NECTEC hub.

At present, NIX and PIE are linked with a 2 Mbps leased circuit. There are eight and ten ISPs connecting to NIX and PIE, respectively. Five of these are connected to both exchange points. However, there are three ISPs that have not yet connected to any local exchange points.

The following table shows the date and link speed of each ISP connected to NIX and PIE.

Internet Service Providers and NIX	NIX		PIE	
	Date Connected	Current link speed	Date Connected	Current link speed
A-Net Co., Ltd.	Feb 4, 1998	512k	Jan 26, 1998	512k
Asia Access Internet Service	-	-	Dec 24, 1997	512k
Asia Infonet Co.,Ltd.	-	-	Dec 30, 1997	512k
C.S. Communications Co.,Ltd	Oct 21, 1997	512k	May 25, 1998	256k
Chomanan Group Co., Ltd.	-	-	-	-
Data Line Thai Co., Ltd.	Aug 10, 1997	64k	-	-
Far East Internet Co.,Ltd.	-	-	Jun 16, 1998	128k
Infonews Co., Ltd.	-	-	-	-
Internet Thailand Service Center	Nov 11, 1996	512k	Nov 15, 1997	10,000k
KSC Commercial Internet Co., Ltd.	Aug 22, 1997	2,048k + 512k	Jun 24, 1998	1,024k
Loxley Information Service Co., Ltd.	Feb 5, 1998	2,048k	Apr 17, 1998	512k
Samart CyberNet Co., Ltd.	-	-	Dec 3, 1997	1,024k

Siam Global Access Co.,Ltd.	N/A	512k	Oct 26, 1998	128k
WorldNET & Services Co.,Ltd.	Sep 27, 1997	512k	-	-
NIX	N/A	N/A	Jun 11, 1998	2,048k

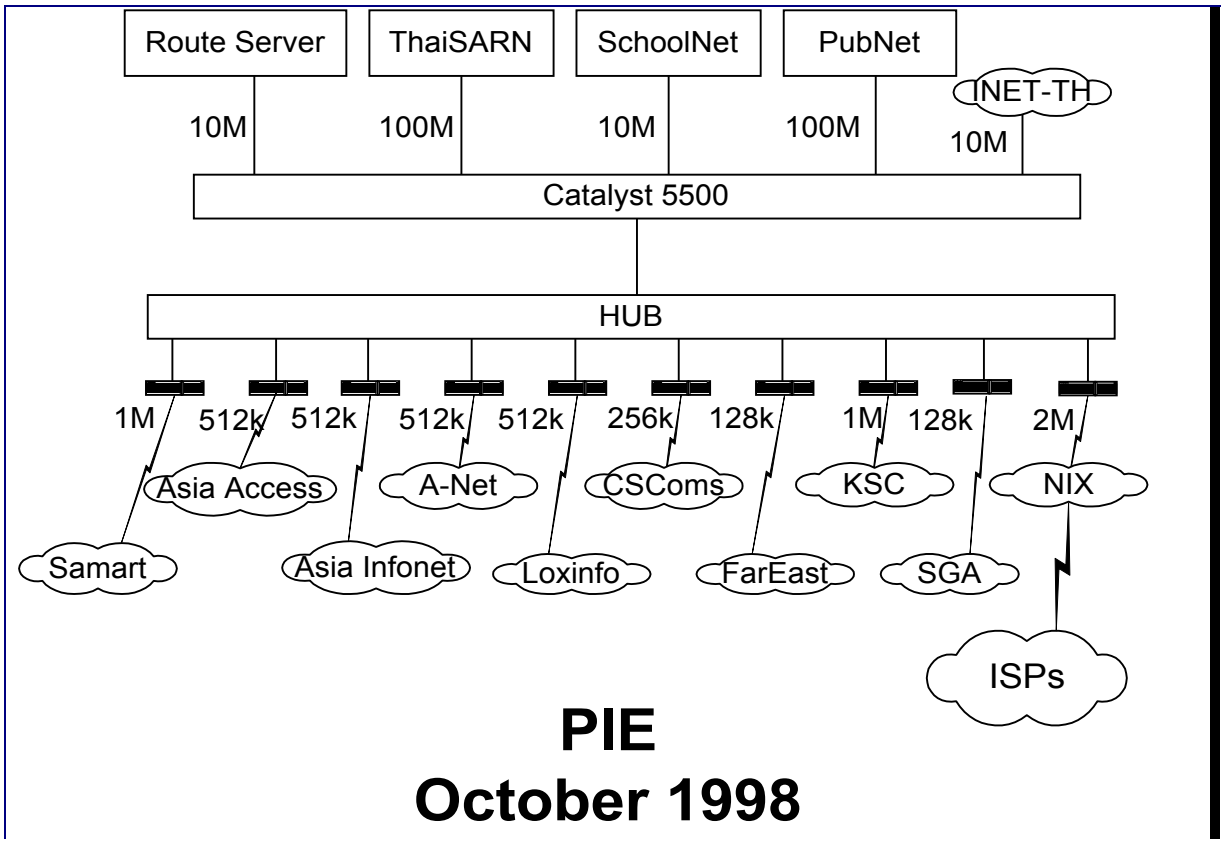
Technical aspects of PIE

PIE's backbone is currently a simple network connecting participants' routers together. Each participant rents its own dedicated circuit at the desired speed of a minimum 128 kbps to PIE and places a router there. All participant's routers are linked together with Ethernet interface at the speeds of 10 or 100 Mbps. Initially, NTL/NECTEC is supporting one 100Base-TX and twelve 10Base-T ports for founding members of PIE. A dedicated router is serving as the route server of PIE. A Cisco Catalyst 5500 LAN switch is used to link PIE to PubNet, ThaiSarn, and SchoolNet at the speed of 100 Mbps. (See diagram)

The route server is the center of routing updates processing for all participants's routers to eliminate the load of multi-sessions route-exchange peering. Routing information is exchanged by means of the BGP-4 routing protocol. Each participant must announce all routes of its entire autonomous systems and its domestic downstreams. The route server will re-broadcast the entire routing entries and updates to all other routers in PIE. The current method to prevent incorrect routing information from propagating through PIE is the use of network access lists on the route server.

PIE Policies

PIE Pilot Project is part of the National Information Infrastructure action plan to support a strong, unified and most economical means for running the Internet in Thailand. It is planned to be a joint mission between Government, Academic/Research Institutions and the private sector.



PIE Project will ensure that all participants will be given equal opportunities to access the third-generation academic/research network "ThaiSarn-III", the Knowledge Distribution Network of the Kanchanapisek Network, and the public portion of the Government Information Network (GINet). PIE is initially funded by the Royal Thai Government, and is operated as a not-for-profit task of NECTEC.

Traffic to and from PubNet and ThaiSarn is provided for free to all PIE Participants (PIEPs) provided that it strictly abides by ThaiSarn Acceptable Use Policy (ThaiSarn AUP). Inter-ISP traffics may or may not follow ThaiSarn AUP; NTL does not concern or impose any restriction over commercial traffic as long as it is not destined for or transit via ThaiSarn and/or PubNet. Violation to ThaiSarn AUP simply means the PIEP is terminated from PIE.

NTL, the Network Operating Center of PIE, will measure traffics of each PIEP. Should an incoming or outgoing traffic of a PIEP circuit exceed 60% average utilization for seven (7) consecutive days, the PIEP must upgrade the under-capacity link to meet PIE standard QOS.

Connection to PIE Pilot Project is free of charge for one year. Eligible PIEP must have a valid Thai ISP licence granted by the Communications Authority of Thailand. Costing for further funding of PIE Project is subject to future government budget availability and the actual cost of previous year's support to each PIEP. However, the long-term service subject to future government budget availability and the actual cost of previous year's support to each PIEP. However, the long-term service charges (if any) will be based on costs and not-for-profit scheme.

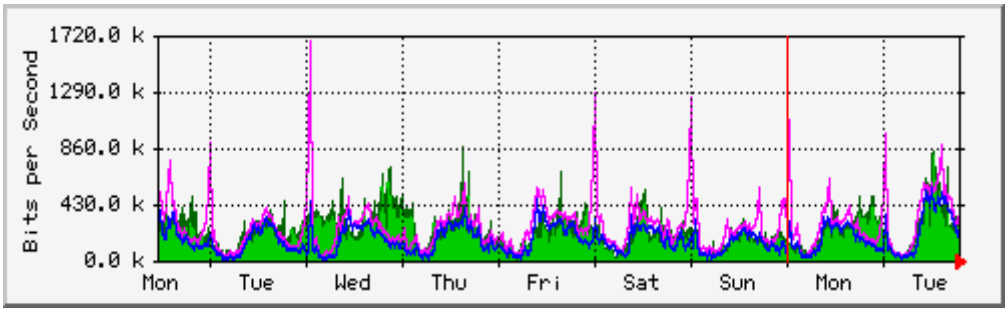
Traffic statistics and analysis

PIE is now operating in a fourteen-tiered fashion with ten ISPs, ThaiSarn, PubNet, SchoolNet@1509, and NIX. To monitor the traffics load on various network links, the Multi Router Traffic Grapher (MRTG) is installed on a unix machine and used to generate graphs representing the traffics on each monitored link into webpages.

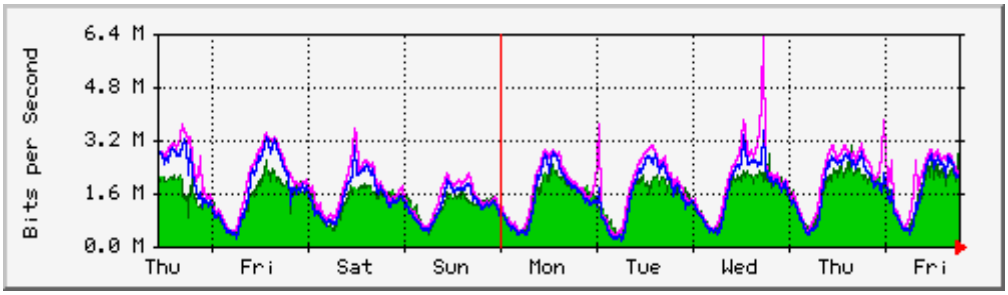
The Graphs shown in the following figures are Weekly Traffic Graph (averaged every 30 minute) of PubNet, ThaiSarn and NIX connections to PIE, respectively. The green color (filled area) in the graph is the In-traffic to PIE and the blue color (Dark line) is the Out-traffic from PIE

.

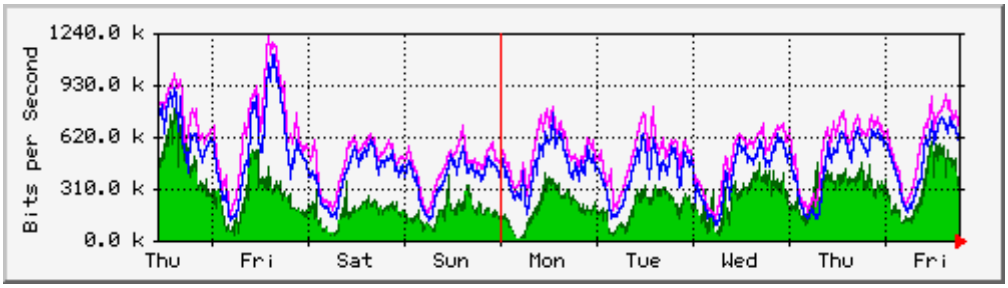
PubNet



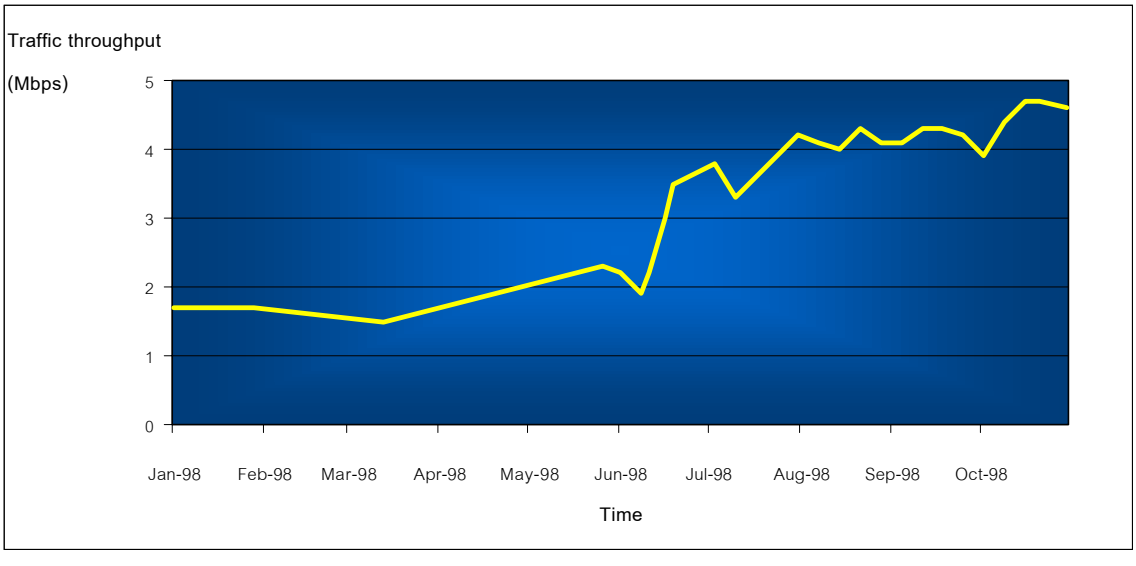
ThaiSarn



NIX



Traffic Growth



The following table shows PIE traffic Statistics in 8-day Average.

ISPs	traffic injected into PIE		Traffic taken from PIE		Line Utilization	
	kbps	Share	kbps	Share	To PIE	from PIE
InternetThailand	1174.7	26.0%	629.2	14.0%	11.7%	6.3%
Samart	346.4	7.7%	397.7	8.9%	33.8%	38.8%
AsiaAccess	42.4	0.9%	129.8	2.9%	8.3%	25.3%
AsiaInfonet	68.2	1.5%	170.9	3.8%	13.3%	33.4%
ANET	213.7	4.7%	337.8	7.5%	41.7%	66.0%
LoxInfo	0.3	0.0%	75.2	1.7%	0.1%	14.7%
CSComs	25.1	0.6%	104.2	2.3%	9.8%	40.7%

FarEast	9.7	0.2%	21.0	0.5%	7.6%	16.4%
KSC	465.0	10.3%	367.9	8.2%	45.4%	35.9%
SGA	22.5	0.5%	7.6	0.2%	17.6%	5.9%
NIX	475.3	10.5%	555.4	12.4%	23.2%	27.1%
ThaiSarn	1422.1	31.5%	1491.4	33.3%	1.4%	1.5%
PubNet	187.8	4.2%	130.9	2.9%	0.2%	0.1%
SchoolNet	62.8	1.4%	61.3	1.4%	0.6%	0.6%
TOTAL	**4516.1**		**4480.3**			

As can be seen from the graphs and table, the current traffic report as of Wednesday, 28 October 1998 at 17:30 can be summarized as follows:

TOTAL AVERAGE TRAFFIC CIRCULATION
4.5 Mb/s (full-duplex)
DAILY TRAFFIC VOLUME
46.3 GB
WEEKLY TRAFFIC VOLUME
324.3 GB
MONTHLY TRAFFIC VOLUME
1389.9 GB

marginal error of measurement (unbalanced sum) 0.1%

sampling interval 5 minutes through a 8-day window

Total traffic at the sustained transfer rate of 4.5 Mb/s is 46.3 GB per day.

From the traffic figures above, if we simply convert the average traffic volume 4.5 Mb/s to be bandwidths of international links based on QoS of 60% average utilization, we may need approximately a total of 7.5 Mbps for the purpose of traffic exchange among participants. That means PIE is helping saving the cost of 8 Mbps international bandwidth and providing faster domestic-traffic exchange for Thai ISPs.

The following table is percentage of time that each link connected to PIE failed (PIE Outage) during 21 April 1998 - 14 September 1998.

Internet Service Providers	**Outage (%)**						
	Apr	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Average**
Internet Thailand Service Center	0.00	0.03	0.00	0.00	0.00	0.00	0.005
Samart CyberNet Co., Ltd.	0.27	0.57	1.58	0.00	0.00	0.04	0.411
Asia Access Internet Service	1.37	0.17	0.71	0.18	0.22	0.14	0.463
Asia Infonet Co., Ltd.	0.20	1.92	0.00	0.00	0.00	0.02	0.356
A-Net Co., Ltd.	0.00	5.95	0.42	0.69	5.15	0.04	2.041
Loxley Information Service Co., Ltd.	0.00	0.09	0.00	1.06	0.19	0.04	0.231
C.S. Communications Co., Ltd	-	0.01	0.47	0.05	0.11	0.12	0.152
Far East Internet Co.,Ltd.	-	-	0.04	0.00	0.00	0.04	0.005
KSC Commercial Internet Co., Ltd.	-	-	0.00	0.02	0.29	0.05	0.090
NIX	-	-	-	0.16	1.32	0.00	0.493

Problems and Challenges

Like several Internet exchanges in other countries, PIE is not without problems. The idea of setting up PIE is by chance synchronous with several ISPs' desires to have an alternative Internet exchange to prevent a single point of failure from NIX. PIE is currently a pilot project supported by NTL/NECTEC. CAT also connects NIX to PIE with a 2 Mbps link for the purpose of traffic exchange between two Internet exchange points. It

is the challenge to prove that should Thailand have two internet exchange points.

On technical side, we have policy that PIE participants must announce full routes to PIE but we currently do not have a mechanism in place to verify this. When an ISP mistakenly announces problematic routes, say a subnet owned by another ISP, it creates wrong routes on PIE. The administrators of PIE won't know the problem until there is complaint from the affected parties. To solve the problem, we have to know which

networks belong to whom. The Routing Arbiter Database (RADB) and related tools are being studied to manage routing policy on PIE.

Future Plan of PIE

To date, PIE is still in its infant stage. The backbone of PIE is a simple network and the service is free of charge. However, NTL/NECTEC plans to improve PIE to be a mature exchange point of Thailand after the one-year pilot project. Three main issues are under discussion for future development of PIE.

First, on the network issue, PIE will continue as a simple network with the replacement of 10 Mbps hubs by 10/100 Mbps Ethernet switch. The link between NIX and PIE will be upgraded to meet PIE QoS and support heavy traffic between two domestic exchange points. New technology such as ATM technology will be studied and tested by NTL to prepare for the high-speed network service in the future.

Next, on the operation and service issue, new softwares and tools are being studied and will be implemented to improve PIE operations and services. The Routing Arbiter Database (RADB), RAToolSet and Route Server (RS) are being studied to see if they are suitable for PIE. New programs for traffic statistics and analysis are to be developed to provide advance services.

Last but not least, on the policy issue, funding models to support PIE after the end of the one-year pilot are being explored. As from November 1, 1998, it will be run by "cost-sharing-model" based on the total expenses of the previous year. It was planned to be a totally non-profit service. However, it may not cost anything to the taxpayer either. The format of consortium would possibly be established to ensure the survival of PIE . NTL/NECTEC is still the neutral party for PIE and will continue to support PIE operations. Nevertheless, the aforementioned issues for the future of PIE are to be presented among PIE participants for further discussions.

Conclusion

The Internet is too big and important to be controlled or governed by any single organization. When the Internet growth in a country reaches a certain state, a way must be found to organize the networks so that it can survive and keep growing. Setting up local Internet exchanges in each country is one way to reduce the unnecessary load on international links and the complexity of network topology.

For the situation in Thailand, the Communications Authority of Thailand (CAT), who regulates Internet in Thailand, first set up Thailand National Internet Exchange (NIX) in July 1996. In November 1997, Network Technology Laboratory (NTL) of the National Electronics and Computer Technology Center (NECTEC), which took part in the development of the Internet in Thailand from the beginning, set up Public Internet Exchange (PIE) as the alternative domestic exchange.

PIE is considered a successful project from its very beginning. The reasons may be ascribed as follows:

- It is the access point to PubNet that provide many free services to the public. ISPs have gained a great benefit from PubNet services.
- It is located at the networks hub of Thailand. ThaiSarn, SchoolNet, PubNet, and PIE are linked together at 100Mbps. Therefore, it provides faster services to PIE.
- It is operated by a neutral party with experienced staffs. NTL/NECTEC is perceived as an appropriate party because NTL is a research laboratory. it does not provide commercial Internet services. In addition, NTL/NECTEC has set up and managed major networks in Thailand.
- It is a low-cost service. Besides small investment in a small router and monthly fee for the leased circuit, participants do not have to pay any service fee to PIE in the first year.
- It satisfies ISPs' requirement as the alternative exchange point with good management and service. PIE provides traffic statistics, mail alerts, and QoS control for all participants.
- All participants have gained benefits from PIE broadband interconnection to national networks.

To date, NIX and PIE are the two local Internet exchanges in Thailand. All parties, both ISPs and Internet users, have gained significant benefits from the local exchanges. However, problems on NIX and PIE must be solved. NIX and PIE will continue to evolve in order to survive and support Internet growth in the future.

References and Future Updated Information

- The Internet Connectivity Map of Thailand: see "http://www.nectec.or.th/inet-map/".
- The SchoolNet Thailand project : see "http://www.school.net.th/".
- The Thai Social/Scientific, Academic and Research Network (ThaiSarn): see "http://ntl.nectec.or.th/thaisarn/".
- ThaiSarn Public access Network (PubNet): see "http://ntl.nectec.or.th/pubnet/".
- Public Internet Exchange (PIE): see "http://ntl.nectec.or.th/pie/".

Acronyms

BGP-4	Border Gateway Protocol version 4
CAT	The Communications Authority of Thailand
ISP	Internet Service Provider
NACSIS	The National Center for Scientific Information System
NECTEC	The National Electronics and Computer Technology Center
NTL	Network Technology Laboratory
PIE	ThaiSarn Public Internet Exchange
PIEP	Public Internet Exchange Participant
PoP	Point of Presence
QOS	Quality of Service
RADB	Routing Arbiter Database
RS	Route Server
ThaiSarn	The Thai Social/Scientific, Academic and Research Network
NIX	National Internet Exchange
TOT	The Telephone Organization of Thailand

Thaweesak Koanantakool
Director of NECTEC

Dr.Thaweesak "Hugh" Koanantakool received his Bachelor and Ph.D. degrees in Electrical Engineering from Imperial College of Science and Technology, London University. He had a number of industrial contracts in the UK before he came back to Thailand to start his government service career in 1981. He taught in Electrical Engineering with the Faculty of Engineering, Prince of Songkla University. In 1985, he moved to Bangkok Thammasat University and was appointed Associate Director of the Information processing Institute for Education and Development. Since 1994, he became Deputy Director of NECTEC as well as leading the Network/Software Technology labs. Thaweesak introduced the Internet into Thailand and set up the largest academic and research network known as ThaiSarn under NECTEC. He later co-founded the first Internet Service Provider (ISP) owned by Thai government in 1995. The ISP, Internet Thailand Company Limited, at present is the largest ISP in Thailand, has 45% market share (by IP numbers managed).

In 1996-1997, Thaweesak led Thailand's Information Superhighway test bed Project funded by NECTEC. The project was a major test bed in Thailand using ATM switches for both local area and wide-areas. In August 1998, Thaweesak was appointed the Director of NECTEC.

Chalermpol Charnsripinyo
Researcher

Chalermpol Charnsripinyo received his BS in Computer Science from Thammasat University, Bangkok, in March 1992. He started his work at Network Technology Laboratory (NTL), NECTEC, as a Research Assistant in May 1992. His main responsibility was the development of ThaiSarn network that was being set up at the time. He had worked at NTL for almost four years before he got the Royal Thai Government Scholarship to continue graduate study in Jan 1996. He is currently pursuing his Ph.D. in Telecommunication program in the US. His areas of interests concentrate on high-speed networking, ATM networking, Internet technologies, Network Performance and Analysis, Network Management, Parallel and Distributed System, and Real-time systems.

Angkana Angkalukkana
Research Assistant

Angkana Angkalukkana was born in Bangkok, Thailand on 21 April 1976. She received her BS in Computer Science from Thammasat University in 1997. She joined the Network Technology Laboratory, NECTEC since May 1997. She was initially responsible for network engineering in ThaiSarn and PubNet. Her current responsibilities include maintaining, designing and developing network for ThaiSarn, SchoolNet, PIE, and PubNet. Her interests are

Internet Routing Architectures, ATM Technology, Network Security and Internet Technologies.

Pornthep Narula
Research Assistant

Pornthep Narula, a.k.a. Tep, joined NTL right after receiving his B.Eng. (Computer) from the KMITL, Bangkok, back in May 1994. His roles at NTL range from system engineering to system and network design and capacity planning. Between August 1997 and January 1998, he was sent to be a Visiting Researcher at the Telecommunication Researcher Laboratories (TRLabs), Winnipeg, Canada, to conduct research in the area of Distributed Computing. Presently, Tep has been assigned to look after the Public Internet Exchange (PIE) and technical aspects of the Kanchanapisek Network (KPNet). He is also leading the R&D efforts of GITS, providing advisory to the SchoolNet@1509 and PubNet projects, and taking part in NTL's Public-Key Infrastructure Task-Force (PKITF). Tep's personal interest nowadays centers around open distributed computing infrastructure, information security and privacy, and Internet as the universal information infrastructure. Specific topics of interest include smart card, embedded systems, and PKI.

NECTEC Technical Journal
90 Vol 1,No. 2, May-June 1999

Amorphous silicon oxide and its application to metal/n-i-p/ ITO type a-Si solar cells[1]

P. Sichanugrist[2], T. Sasaki, A. Asano, Y. Ichikawa, H. Sakai
Fuji Electric Corporate Research and Development, Ltd. 2-2-1 Nagasaka, Yokosuka-shi, Kanagawa-ken 240-01, Japan

Abstract -- P-type and n-type amorphous silicon oxide (a-SiO) films with a microcrystalline Si phasewere deposited by plasma CVD using a gas mixture of SiH_4-CO_2-H_2-$B2H_6$ and SiH_4-CO_2-H_2-PH_6 respectively. These films had lower absorption coefficients than conventional a-SiO due to larger oxygen contents and microcrystalline phase, but their conductivities were still high since they contained some microcrystalline Si phase. Furthermore, it was found that it is easier to make microcrystalline a-SiO than a-SiC films at low substrate temperature. By applying these films to the p-layer of metal/nip/ITO type cells, higher performance was obtained, compared to the cells with conventional microcrystalline Si p-layer deposited at lower temperature and conventional a-SiO p-layer. From these results, we consider that these novel a-SiO films with microcrystalline Si phase are a promising material for the window layer of a-Si solar cells.

Keywords : solar cells, amorphous silicon oxide, microcrystalline phase

บทคัดย่อ -- ฟิล์มชนิดใหม่ที่มีส่วนผสมระหว่างอะมอร์ฟัสซิลิกอนออกไซด์และผลึกซิลิกอนขนาดเล็ก แบบพีและเอ็นสามารถสร้างด้วยวิธีพลาสม่า CVD จากก๊าซผสมของ SiH_4-CO_2-H_2-$B2H_6$ และ SiH_4-CO_2-H_2-PH_6 ได้ ฟิล์มชนิดนี้มีสัมประสิทธิ์ของการดูดกลืนแสงที่ต่ำกว่าฟิล์มชนิดอะมอร์ฟัสซิลิกอนออกไซด์แบบเดิม เนื่องจากมีส่วนผสมของออกซิเจนมากขึ้น และมีส่วนที่เป็นผลึกอยู่ด้วย แต่ยังคงมีคุณสมบัติที่นำไฟฟ้าได้ดีอีกด้วย เนื่องจากมีส่วนที่เป็นผลึกอยู่ เมื่อนำไปประยุกต์ใช้เป็นชั้นพี ของเซลล์แสงอาทิตย์แบบ Metal / nip / ITO พบว่า เซลล์แสงอาทิตย์ที่ได้มีประสิทธิภาพสูงขึ้น ทำให้สามารถสรุปได้ว่า ฟิล์มชนิดใหม่ที่ค้นพบนี้ เป็นฟิล์มที่เหมาะที่จะใช้ในชั้นรับแสงของเซลล์แสงอาทิตย์

คำสำคัญ : เซลล์แสงอาทิตย์, อะมอร์ฟัสซิลิกอนออกไซด์, ผลึกขนาดเล็ก

1. Introduction

Recently, we have demonstrated that hydrogenated amorphous silicon oxide(a-SiO:H) films deposited by plasma CVD using a gas mixture of SiH_4-CO_2-H_2 have a better film quality than conventional a-SiC films. Applying this a-SiO layer to the p-layer of glass/SnOJpin/metal, we have obtained 12.5% for 1 cm^2 area cell [1]. To achieve higher photovoltaic performance, we have

[1] This article is a reprint of the article appeared in the Solar Energy Materials and Solar Cells 34 (1994) pp. 415-422.

[2] He is currently working with the National Electronics and Computer Technology Center (NECTEC).

concentrated on developing wider bandgap materials having the same electrical properties. Further more, we also considered a different device structure which is (inverted pin) type structure since I layer can be deposited at higher substrate temperature in this cell structure.

For wider bandgap materials, a-SiC with microcrystalline Si phase (,ac-SiC) has been developed by Y. Hattori et al. using ECR plasma CVD [2] and by K. Hanakiet al. [3] or S. Guha et al. using RF plasma CVD [4]. It was reported that higher performance were obtained with these p-type ,ac-SiC films [2,5]. However, up to date there are no reports on a-SiO with microcrystalline Si phase (,u c-SiO) and its application to the cell.

In this paper, we have studied for the first time p-type and n-type ,ac-SiO films deposited by RF plasma CVD using a gas mixture of SiH_4-CO_2-H_2-B_2H_6 and SiH_4-CO_2-H_2-PH_3 , respectively. Their film structure were evaluated and their optical and electrical properties were also measured and compared with those of conventional p(a-SiO) films. To compare the properties of these ,ac-SiO films, we have also tried to deposit p(yc-SiC) film from the gas mixture of SiH_4-C_2H_2-H_2-B_2H_6 under the same deposition condition. Furthermore, we will report the results of applying these p(,ac-SiO) films to the p-layer of metal/nip/ITO type cell.

2. Experimental

In this experiment, to deposit ,ac-SiO films, CO_2 was introduced to the gas mixture under normal deposition conditions for,uc-Si deposition. Table 1 shows the details of our typical deposition condition for ,ac-SiO films. Films having thicknesses of 200-300 nm were deposited on the Corning glass and c-Si substrate. The microcrystallization and the film structure were evaluated by X-ray diffraction, Raman, and FT-IR spectroscopy. The oxygen and carbon contents in the films were measured by ESCA while the hydrogen content was evaluated from FT-IR spectra. To evaluate their electrical and optical properties, dark conductivity and photoconductivity measured under AM 1.5, 100 mW/cm^2 , optical bandgap , and absorption coefficient were used.For cell fabrication, n-, i- and p-layers were deposited in different chambers in order to suppress the dopant contamination into the i-layer.

Table 1 Typical deposition condition used for the deposition of p(yc-SiO) and n(yc-SiO)

CO_2 /(SiH_4 + $C0_2$)	0-0.6
H_2 /SiH_4	160-320
B_2H_6(PH_3)/SiH_4	0.6-1.0%
Substrate temperature	100-250°C
Pressure	1.0 Torr
RF power density	50 mW/cm^2

3. Results and discussion

3.1. Film structure

3.1.1. Contents of silicon, oxygen, carbon and hydrogen

Table 2 shows the contents of silicon, oxygen, carbon, and hydrogen in p(,ac-SiO) films, compared with those in conventional p(a-SiO) films. As can be seen from this table, p (,ac-SiO) film has a higher oxygen content in the film than the a-SiO film which results to its higher bandgap. On the other hand, the carbon content in,uc-SiO films is also less than detection limit (0.5%). The hydrogen content in,ac-SiO films is about half of conventional p (a-SiO) film which is normally observed in ,uc-Si films.

3.1.2. Microcrystallinity

It is confirmed from the Raman spectra that there is a distinct narrow peak at around 520 cm-' corresponding to the TO phonon mode of crystalline Si in the above-mentioned p(yc-SiO) films, meaning that these films contain some microcrystal-line phase. However, the peak intensity becomes smaller with an increase of CO_2 gas ratio , showing that the microcrystal-lization becomes more difficult with more oxygen. It is found that besides a large mount of oxygen, this crystalline Si phase also makes the absorption of ,ac-SiO films lower.

Furthermore, we have evaluated the crystallite size and volume fraction of ,µc-si in p(,ac-SiO) films from the measured X-ray diffraction pattern. Their data are shown in Table 3. As expected, the crystallite size and volume fraction become smaller with an Increase of oxygen

Table 2 Contents of silicon, oxygen, carbon and hydrogen in p(yc-SiO) films

Type	CO_2 /(CO_2 + SiH_4)	Si (%)	O (%)	C (%)	H atom /cm^3
p(a-SiO)	0.67	78.4	21.6	N.D.	$1.37e^{22}$
p(yc-siO)	0	100	N.D.	N.D.	$1.68e^{21}$
	0.37	73.9	26.0	N.D.	$8.71e^{21}$
	0.54	65.0	35.0	N.D.	$7.30e^{21}$

Table 3 Crystallite size and volume fraction of ,uc-Si in p(,8c-SiO) films

Type	CO_2 /(CO_2 + SiH_4)	Crystallite Size(°A)	Volume fraction (%)
P(μc-SiO)	0	298	100(assumed)
	0.37	127	58
	0.54	50	39

3.1.3. IR spectra

From IR spectra, it is observed that p(, (lC-SiO) films also contain oxygen atoms in a form of Si-O-Si, not O-H, which is similar to a-SiO films [6].

3.2. Optical and electrical properties

Fig. 1 shows the relation between the gas ratio of CO_2 to SiH_4[CO_2(SiH_4 +CO_2)] and the absorption coefficient for the p - type , ac-SiO films . As can be
Seen From this figure, the absorption coefficient decreases with an increase of the gas ratio due to the incorporation of oxygen into the film. The same result is also observed for the n-type films.

Fig. 2 shows the effect the gas ratio, CO_2/(SiH_4+CO_2) on the optical bandgap (Eopt) dark conductivity (tJd) and photoconductivity ((Tph) f p-type, ac-SiO films. The conductivity decreases with increasing the gas ratio while the bandgap increases due to the incorporation of oxygen into the film. It should be noted here that the conductivities of n-type yc-SiO films are higher than those of p-type ,uc-SiO films.

Fig. 3 shows the dark conductivity and photoconductivity as a function of Eopt for conventional p-type a-SiO and ,ac-SiO films. As can be seen from this figure,the conductivities of p(,u c-SiO) films are about four orders magnitude higher than those of conventional p (a-SiO) films in the optical gap range of 2.15-2.25 eV, showing that the ,ac-SiO film is the promising material for the window layer of a-Si solar cells.

3.3. The deposition of ,ac-SiC film under the same deposition conditions

To compare with p(,ac-SiO) films, we have fabricated some p-type ,uc-SiC films from the gas mixture of SiH_4-C_2H_2-H_2-B_2H_6. We have found that it is difficult to fabricate a-SiC with microcrystalline Si under the same conditions at the substrate temperature of 150° C. This may come from the fact that the substrate temperature was not high enough for the deposition of ,ac-SiC.

3.4. Application to metal/nip/ito type solar cell.

The above-mentioned p(,ac-SiO) films were applied to the p-layers of metal/ni-p/ITO type solar cells. Inthis type of solar cells, the microcrystalline Si p layers should be normally

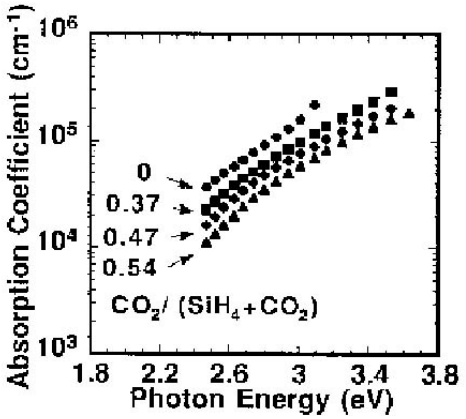


Fig. 1. Effect of the gas ratio, CO_2 /(SiH_4 + CO_2) on the absorption coefficient of p-type ,uc-iO films.

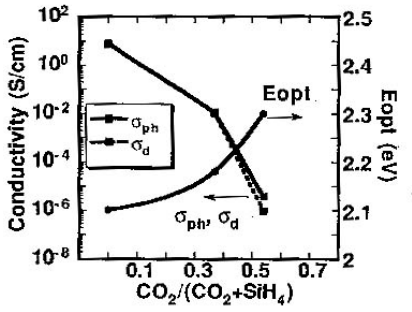


Fig. 2. Effect of the gas ratio, CO_2 /(SiH_4 + CO_2) on the Eopt, dark conductivity and photoconductivity of p-type yc-SiO films.

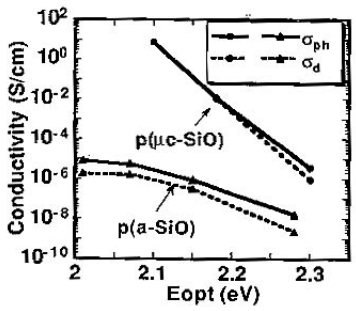


Fig. 3. The dark conductivity and photoconductivity as a function of Eopt for conventional p(a-SiO) and p(,u c-SiO) films

prepared at the substrate temperature lower than 100°C to achieve high VOC (> 0.85 V) and high performance [7]. However, it will be better for mass production if the p-layer can be deposited at a temperature as high as those of other layers.

We have investigated the effect of CO_2 gas ratio to the performance of the cell with p (,ac-SiO) films and its result is shown in Fig . 4. Here, p(,uc-SiO) layer was deposited at 150°C directly on the i-layer without any interface layer at p/I interface. The result obtained with the conventional microcrystalline Si p-layer prepared at 100°C is also shown in this figure for comparison. As can be seen from this figure, higher performance with higher VOC (> 0.92 V) and Jsc are obtained with p (,LtC-SiO) layer. It should be also noted here that its performance is better than the one with conventional ,ac-Si p-layer, even though the p(,uc-SiO) layer was deposited at 150°C.

Fig. 5 shows the effect of gas ratio to the collection efficiency. As seen from this figure, the response at short wavelength increases with an increase of oxygen due to the wider bandgap of the p-layer. Furthermore, it is confirmed that the cells with p(,ac-SiO) layer have better performance (higher VOC and Jsc) than the cells using conventional a-SiO layer as the p- and p/i interface layers. Up to now, an efficiency of 9.1% (VOC = 0.913 V, Jsc = 15.8 mA/cm2, FF = 0.632) was achieved for 1 cm2 area cell. It is expected that higher performance can be obtained by slightly boron-doping into the i-layer and by optimizing the p-layer thickness and the properties of p/i and ITO/p interfaces.

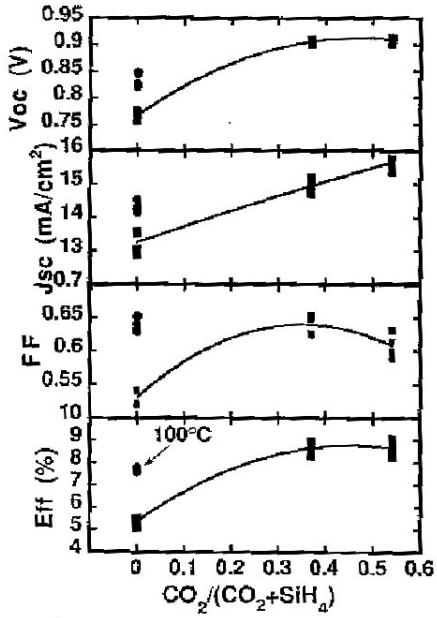


Fig. 4. Effect of the gas ratio, CO_2 /(SiH_4 + CO_2) on the cell performance of metal/n-i-p/lTO havingp(,gc-SiO) film as the p-layer.

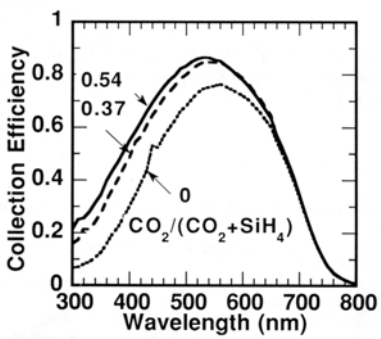


Fig. 5. Effect of the gas ratio, CO2/(SiH4+CO2) on the collection efficiency of metal/n-i-p/lTO having p(,u c-SiO) film as the p-layer.

4. Summary

We have found that a-SiO films with a microcrystalline Si phase can be easily fabricated by introducing an amount of CO_2 gas to the gas mixture of SiH_4 under the deposition condition of microcrystalline Si films. These films have lower absorption coefficients than conventional a-SiO films because of larger oxygen contents and some microcrystalline Si phase. However, their conductivities are still high since they contained some microcrystalline phase. By applying these films to the p-layer of metal/nip/ITO type cells, it was confirmed that higher perfor mance was obtained with this p (,uc-SiO) layer, comparing to the cells with conventional microcrystalline and conventional a-SiO p-layers. From these results, we consider

that this novel ,ac-SiO film is a promising material for the window layer of a-Si solar cells.

Acknowledgement

This work was supported by the New Energy and Industrial Technology Development Organization under the New Sunshine Program of the Ministry of International Trade and Industry.

References

[1] Y. Ichikawa, S. Fujikake, H. Ohta, T. Sasaki and H. Sakai, Proc. 22nd IEEE Photovoltaic onf., Las Vegas, 1991, p. 1296.

[2] Y. Hattori, D. Kruangam, K. Katoh, Y. Nitta, H. Okamoto and Y. Hamakawa, Proc. 19th IEEE Photovoltaic Conf., 1987, p. 689.

[3] K. Hanki, T. Hattori and Y. Hamakawa, Technical Digest of International PVSEC-3, 1987,p. 49.

[4] S. Guha and S.R. Ovshinsky, US Patent No. 4775425, 1988.

[5] A. Banerjee and S. Guha, Mat. Res. Soc. Symp. Proc., Vol. 192 (1990) p. 57.

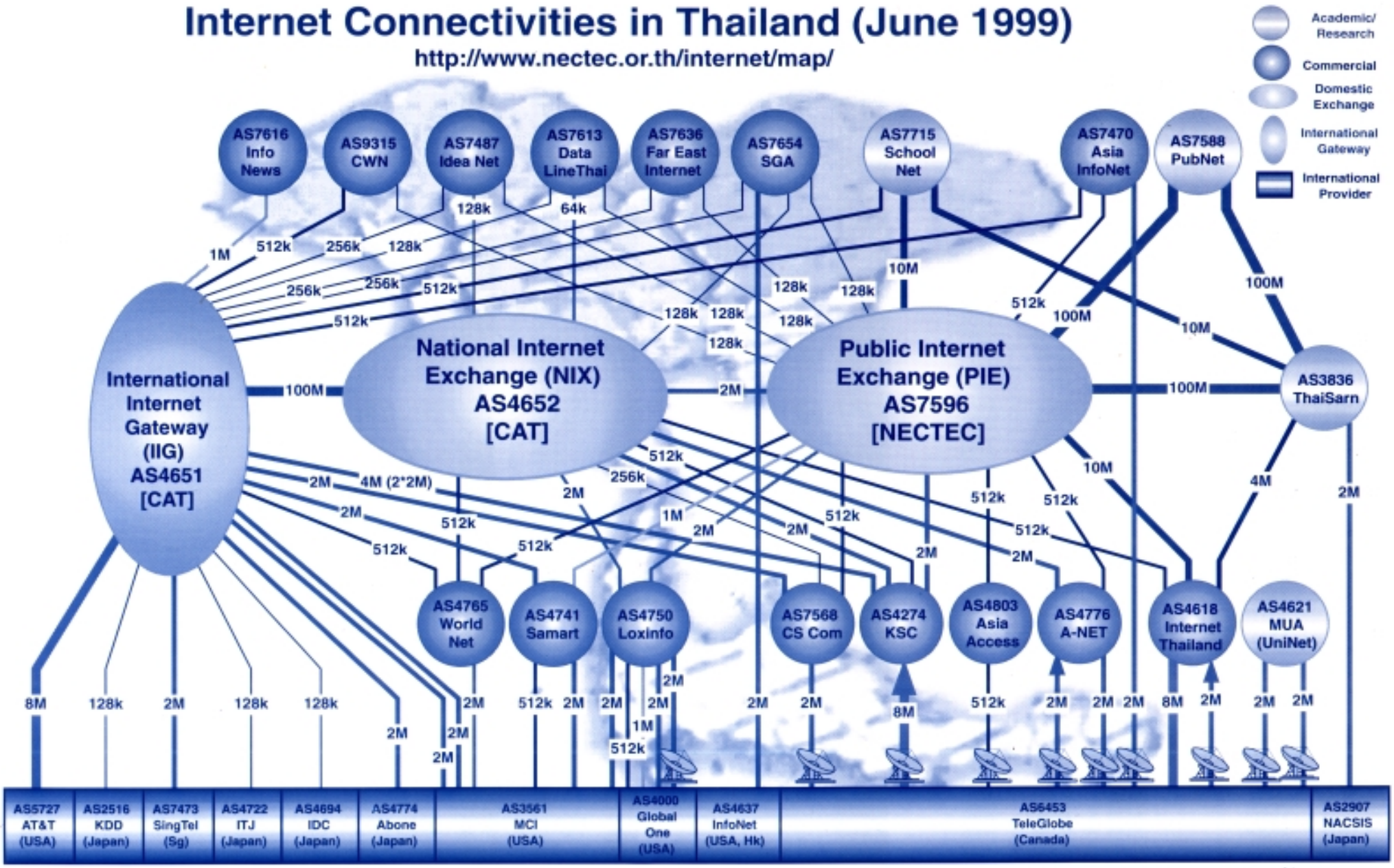
[6] S. Fujikake, H. Ohta, A. Asano, Y. Ichikawa and H. Sakai, MRS Spring Meeting, 1992, p.875.

[7] T. Ishihara et al., Proc. 19th IEEE Photovoltaic Conf., 1987, p. 749.

Porponth Sichanugrist received his B.Eng (Electronics and Electrical Engineering), M. Eng and D. Eng (Physical Electronics) from Tokyo Institute of Technology in 1981,1983 and 1986, respectively. He worked as a researcher and plant engineer in Showa Shell Sekiyu until 1992 and then he joined Fuji Electric for one year as a researcher. He came to join NECTEC in 1994 and was in charge of Thailand's Microelectronics R&D Center Project (TMEC Project) until 1997. He currently manages the R&D project on solar cells suited for tropical climate.

Internet Connectivities in Thailand (June 1999)

http://www.nectec.or.th/internet/map/



DISCLAIMER

Chart Date: 1999-06-01

ไทยสาร ThaiSarn