Combining Aggregation and Scheduling using an iterative Maximal Weight Matching Switch Scheduler

BRIAN BACH MORTENSEN
Department of COM, Networks Area
Technical University of Denmark
Oersteds Plads 343, 125, 2800 Kgs. Lyngby
Denmark

Abstract: - Hybrid electro-optical packet switches utilize optics in the backplane to switch optical packets from inputs to outputs on electronic line cards. The optical packets are traditionally considerably larger than minimum size IP packets. IP packets entering the switch must be formatted (segmented) and encapsulated in the optical packet format. This process is called packetisation or aggregation. This paper investigates a novel technique for aggregating IP packets into optical packets. It combines the well known i-OCF switch scheduler with the aggregation process. When a segment arrives at an empty optical packet, the optical packet is time stamped. This time stamp is then used in the i-OCF scheduler to compute a maximal weight match for the hybrid electro-optical packet switch matrix. In this paper it is illustrated that the optimum number of iterations, used for this particular application, is higher than Log(N) which is used for maximal size matching schedulers as i-SLIP. Furthermore, it is investigated how large a speedup is required in order to provide 100% throughput.

Key-Words: - i-OCF, MWM Scheduler, Optical Packet Switching (OPS), Packet Aggregation

1 Introduction

Medium to large scale packet switches with an aggregate throughput between 0.5 Tb/s and 5 Tb/s may potentially be implemented using hybrid electro-optical technologies in the future. Hybrid electro-optical packet switches utilize an optical switch matrix, instead of an electrical switch matrix found in conventional packet switches. This approach has been proposed in the literature due to potential advantages regarding power consumption and scalability [1] [2]. Furthermore, using optics in the switching backplane makes it possible to avoid expensive conversion of signals between electrical and optical domains. Figure 1 shows the basic architecture for this hybrid electro-optical packet switch. Traditional electrical packet switches may use high speed serial links (HSSL) between the line cards and the switch matrix. However, these links often use a large amount of power, harmful to the total power budget. Previous research projects have shown that implementing optical packet switches with a total capacity of 2.56Tb/s is viable using the well known broadcast and select architecture [3]. With this architecture all wavelengths are amplified and broadcasted to a space and wavelength selection unit. The space and wavelength selection unit utilizes SOA (Semiconductor Optical Amplifiers) gates to select a wavelength from a specific input port. The optical packet switch (OPS) in [3] operates in a time slotted scheme, thus employing fixed length optical packets to be transmitted over the backplane switch matrix. The time slotted scheme implies a guard band between optical packets to allow for reconfiguration of the optical switch matrix. Furthermore, overhead is needed in the optical packet in the form of a delimiter, which can be used by the receiver to recover the received data. A clock signal may be distributed between the different parts of the system. The clock signal frequency is helpful when extracting the received data at the output line cards. However, in order to do the data extraction phase information is also needed. Finding the phase of the received signal is much easier when the frequency is known. This means that the overhead used by the receiving line cards, may be reduced significantly compared to systems without a distributed clock signal.

The broadcast and select switch matrix used in [3] may also be reduced in terms of 3R regenerators, because the optical packets are terminated immediately after the switch matrix.

When using a time slotted approach as described above, a method that aggregates data (e.g. IP datagram) into the large sized optical packets must be defined. In this study a simple and viable method is selected. The method chops the incoming data into smaller fixed size segments. These segments are

then forwarded to the aggregation process that queues the segments in optical packets.

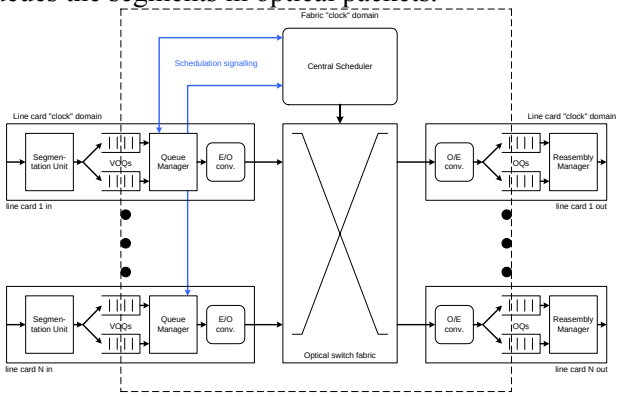

Figure 1 Hybrid Electro-Optical Packet switch architecture.

Each segment contains a very small header to keep information regarding the delineation of IP packets within the segment. This information is used in the egress line card to reassemble the IP packets. An optical packet containing three IP packets is illustrated in Fig. 2. The optimum number of segments in the optical packet depends on the optical packet size and the distribution of the incoming IP packet lengths. In this paper the segments are 64 bytes long, thus allowing minimum sized IP packets to be transported completely in one segment.

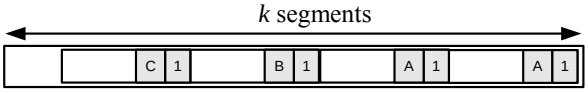

Figure 2 Optical packet format containing up to k segments

This segmentation method allows data from different IP datagrams to be transmitted through the switch in the same optical packet as illustrated in Fig. 2. Since some VOQ may have very little traffic, a technique that ensures fairness and bounded delay must be defined. In fact, a worst case situation could be when a minimum size IP datagram is encapsulated in a single segment, without any other segments arriving at that particular VOQ. The optical packet would never be filled, making it wait infinitely to be recognized by the queuing systems as an eligible optical packet. In this paper we timestamp the optical packet when the first segment is encapsulated. The timestamp is represented as a digital number which is distributed in a synchronous fashion by the scheduler. The number of bits representing the timestamp must be selected such that the time can loop around without causing problems seeing which packet is the oldest. Since a centralized scheduler is used, the timestamps may actually be organized as a set of lists managed by the scheduler. However, this has the impact that the line cards must only report the arrival of the first segment in an optical packet (announcing subsequent segments in the same optical packet, will cause a mismatch between actual number of packets in the VOQ and the number maintained in the scheduler set of lists).

2 Switch Model

This segmentation method allows data from different IP datagrams to be transmitted through the switch in the same optical packet as illustrated in Fig. 2. Since some VOQ may have very little traffic, a technique that ensures fairness and bounded delay must be defined. In fact, a worst case situation could be when a minimum size IP datagram is encapsulated in a single segment, without any other segments arriving at that particular VOQ. The optical packet would never be filled, making it wait infinitely to be recognized by the queuing systems as an eligible optical packet. In this paper we timestamp the optical packet when the first segment is encapsulated. The timestamp is represented as a digital number which is distributed in a synchronous fashion by the scheduler. The number of bits representing the timestamp must be selected such that the time can loop around without causing problems seeing which packet is the oldest. Since a centralized scheduler is used, the timestamps may actually be organized as a set of lists managed by the scheduler. However, this has the impact that the line cards must only report the arrival of the first segment in an optical packet (announcing subsequent segments in the same optical packet, will cause a mismatch between actual number of packets in the VOQ and the number maintained in the scheduler set of lists).

3 I-OCF Investigation and simulation

The simulation model was verified by a number of simulation studies used to compare with results presented in [4][5]. The primary validation effort is related to the i-OCF scheduler as this is a key component. A 16x16 switch with 4 iterations is simulated. Cell mode scheduling is used, avoiding influence from aggregation processes and the slot overhead. The traffic source for each input line card is Bernoulli distributed, with a probability ρ_i at input line card i of receiving a packet in each timeslot. Likewise the probability of not receiving a packet at each input line card is $1-\rho_i$. A traffic arrival rate matrix Γ (Equation 1) is introduced to find the distribution of packets for each line card.

The unbalance parameter in the traffic arrival rate matrix, determines the weight of unbalance between the line cards.

Equation 1

$$\Gamma = \begin{bmatrix} \delta & 1-\delta & 0 & 0 \\ 0 & \delta & 1-\delta & 0 \\ 0 & 0 & \delta & 1-\delta \\ 1-\delta & 0 & 0 & \delta \end{bmatrix}$$

The normalized throughput vs. the normalized load is shown in Fig. 3. The total offered load in the simulation is 100% (normalized offered load equals 1).

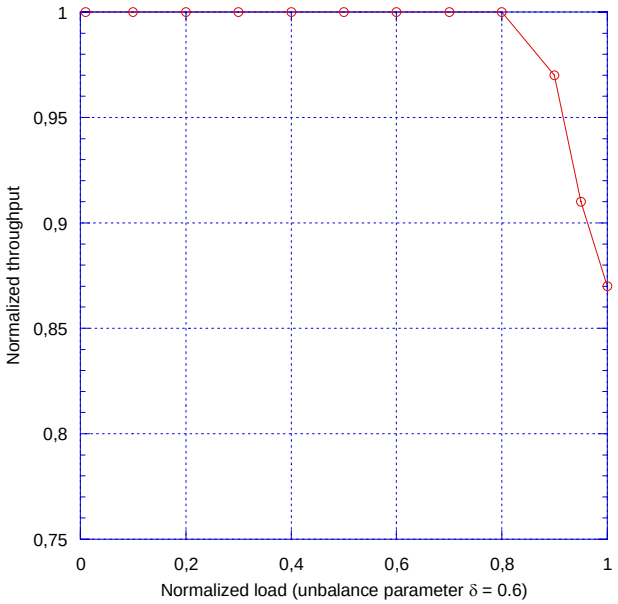

Figure 3 Result of simulation for 16x16 switch with a 4 iteration i-OCF scheduler.

It is noticed that the normalized throughput may go as low as 0.87 when the unbalance parameter is equal to 1. The results are identical to the results found in [5]. In order to see if the result is affected by the switch size, a simulation is carried out that shows the normalized throughput for a 32x32 port switch with 5 iterations of i-OCF. The results are shown in Fig. 4.

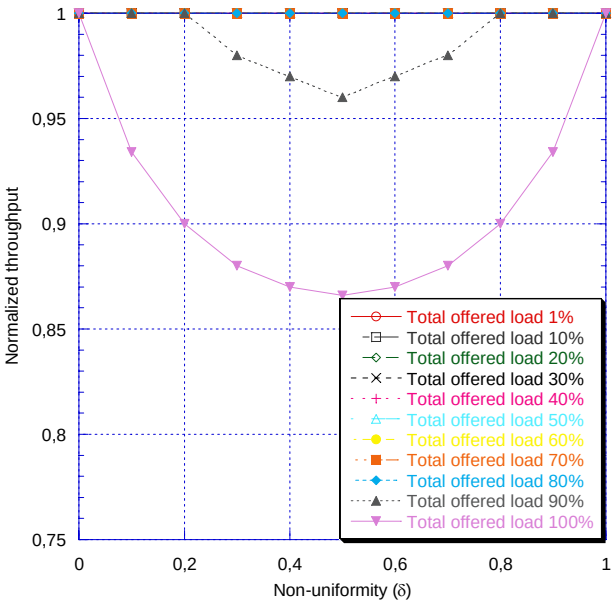

Figure 4 Result of simulation for 32x32 switch with a 5 iteration i-OCF scheduler.

The results are quite similar, showing a minimum normalized throughput of around 0.87 for unbalance parameter close to 0.5. The x-axis is changed between the Fig. 4 and Fig. 5. This is however only to make Fig. 4 comparable with the study in [5], and the preferred format in the rest of this paper.

In Fig. 4 (32x32 configuration) the number of iterations has been selected to 5, which are based on the normal assignment of iterations used with the iSLIP scheduler. In iSLIP the number of iterations (i) are given by $i = Log_2 N$, were N is the number of line cards in the switch. To the author's knowledge, no study has shown that this relation holds for iOCF. So there is a possibility that further iterations give a higher throughput (justifying the higher time complexity of the algorithm). A simulation is carried out to investigate the throughput performance of the iOCF scheduler, with 100% total offered load and different non-uniformity parameters and iteration numbers. The results of the simulations are illustrated in Fig. 5.

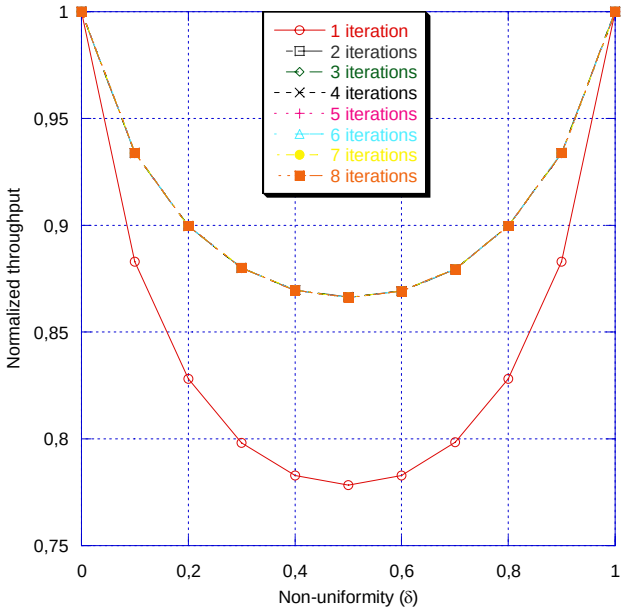

Figure 5 Normalized throughputs for different number of iterations with unbalanced Bernoulli traffic.

It is observed that increasing the number of iterations beyond 2 is not raising the normalized throughput for any unbalance factor. This seems intuitive since the traffic is only distributed between two distinct inputs for each output and vice versa. It is conjectured that it is the maximal matching properties of the scheduler that obstructs a maximum match.

In order to investigate the effect of the number of scheduler iteration cycles, a different traffic arrival rate matrix is used [6]. The diagonal and non-diagonal traffic rate elements are filled according to the following equations:

$$\lambda_{i,j} = \begin{cases} \lambda\left(\omega + \dfrac{1-\omega}{N}\right) & \text{if } i = i \\ \lambda\dfrac{(1-\omega)}{N} & \text{otherwise} \end{cases}$$

$\lambda_{i,j}$ is the traffic intensity from input i to j output , N being the number of line cards. ω is the degree of non-uniformity. The following conditions apply:

$$0 \le i, j < N$$

$$0 \le \omega \le 1$$

The offered load per input and output port is admissible when:

$$0 \le \lambda_{i,j} \le 1$$

And:

$$\lambda_i = \sum_{j=0}^{N-1} \lambda_{i,j} = \lambda_j = \sum_{i=0}^{N-1} \lambda_{i,j} = \lambda\left(\omega + N\frac{1-\omega}{N}\right) = \lambda$$

Using this non-uniform traffic profile, a simulation is carried out that shows the normalized throughput of the i-OCF scheduler for different average cell burst lengths. The result of the simulation is shown in Fig. 6. The cell trains are modeled using an ON-OFF model for each individual line card. The ON state average burst length is geometrical distributed and the OFF state average length is exponential distributed.

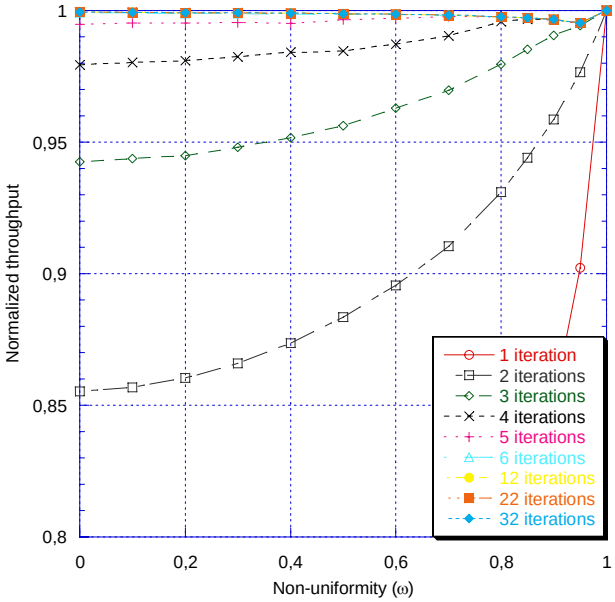

Figure 6 Normalized throughputs for different number of iterations with non-uniform traffic arrival and average cell burst length of 1. 32x32 switch size.

It is seen that increasing the number of iterations from 5 to 6 gives a small but never the less larger normalized throughput for non-uniformity parameters close to 0. Using more than 6 iterations does not give any significant improvement for this setup. Simulations with average cell burst lengths 8 and 16 are shown in Fig. 7 and Fig. 8. Again, it is observed that increasing the number of iterations from 5 to 6 gives an increase in normalized throughput for non-uniformity parameters close to zero.

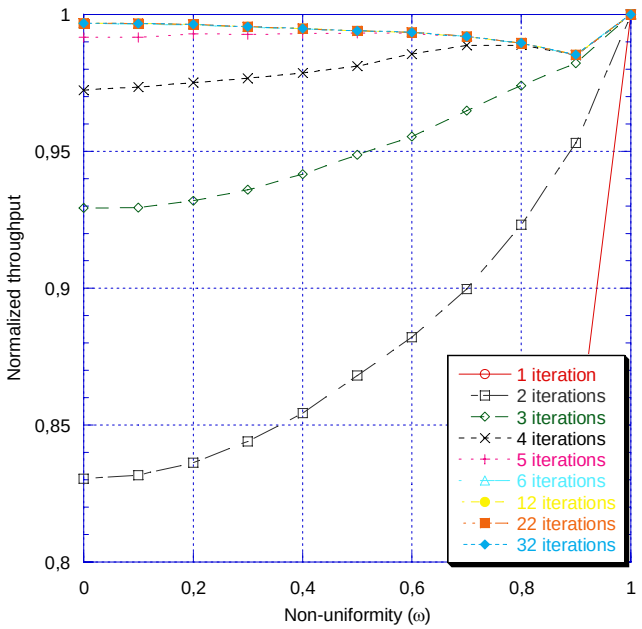

Figure 7 Normalized throughputs for different number of iterations with non-uniform traffic arrival and average on length of 8.

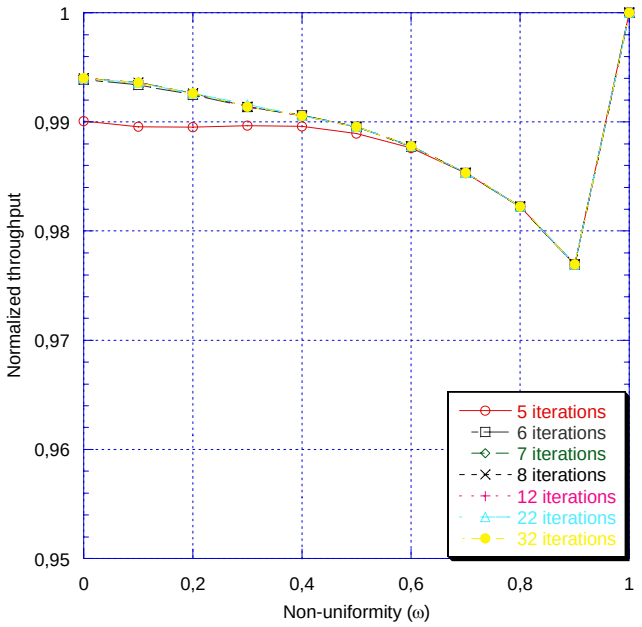

Figure 8 Normalized throughputs for different number of iterations with non-uniform traffic arrival and average on length of 16.

This result is quite intuitive, since the number of edges in the corresponding request graph gets higher, for lower values of non-uniformity. Therefore, more scheduler iterations are needed to find the maximal weight match. It is also seen that the normalized throughput is negatively impacted when the non-uniformity is around 0.9. This is not related to the number of scheduler iterations. Basically, this implies that a match was made in the beginning of the iterations, which blocks a possible maximum match later on in the iterations.

4 Combining Aggregation and Scheduler

In the previous simulations results, cell mode scheduling was used. This gave the possibility to investigate and compare the performance of the i-OCF scheduler. In the following a combination of the scheduler and aggregation of segments will be used, as explained in the beginning of this paper. Two simulations are carried out with the non-uniform traffic arrival rate matrix used above and defined in [6].

In this paper an optical slot length of 1μs is selected. The optical packet used in the backplane contains 8192 bits, which is the equivalent of 16 segments holding 64 bytes each. If the external and internal interface line speeds are equal (e.g. 10Gb/s) the maximum admissible total offered load is 81.92% due to simple bandwidth limitation. Inefficiencies in scheduling and aggregation can actually lower this value significantly, which was shown in the previous section. Selecting a packet size of only 8192 bits at 10Gb/s line speed with 1μs timeslots, is clearly very conservative, but the consequence may very well be that a real-world system would perform much better. The complete simulation parameter set is illustrated in TABLE I.

Table 1

Simulation parameter	Value
Line cards	32
Optical packet length	1024B
Segments per optical packet	16
Segment size	64B
Scheduler iterations	6
Time slot length	1us
Arrival process	On-Off model
On state distribution	Geometric
Off state distribution	Exponential
On state average length	1 and 16
Off state average length	Calculated from load
Non-uniformity ω	0 to 1

Fig. 9 and Fig. 10 illustrate that the average delay is tolerable for the selected simulations. However,

when the total offered load is 80% and non-uniformity is around 0.9 a higher average delay can be seen. This is related to the fact that the maximum total offered load is 81.92% due to the number of bits transported in one time slot (16 segments x 64 Bytes x 8). Furthermore, it was shown in the previous section that the i-OCF scheduler has a lower normalized throughput, when the non-uniformity is close to 0.9. Decreasing the reconfiguration overhead, thus using larger packet in the timeslots, is viable and gives a significantly higher admissible throughput. In the simulations results shown in this paper, the worst result is observed when using unbalanced Bernoulli arrival rate matrix with an unbalance parameter of 0.5. The normalized throughput is in that case reduced to approximately 0.87. With a speed up factor of 1.15, 100% throughput is admissible using cell mode scheduling. Having a reconfiguration overhead and hence losing up to 18% of the timeslot, requires a higher speed up to reach 100% throughput. In this case a speed up of 1.4 is needed.

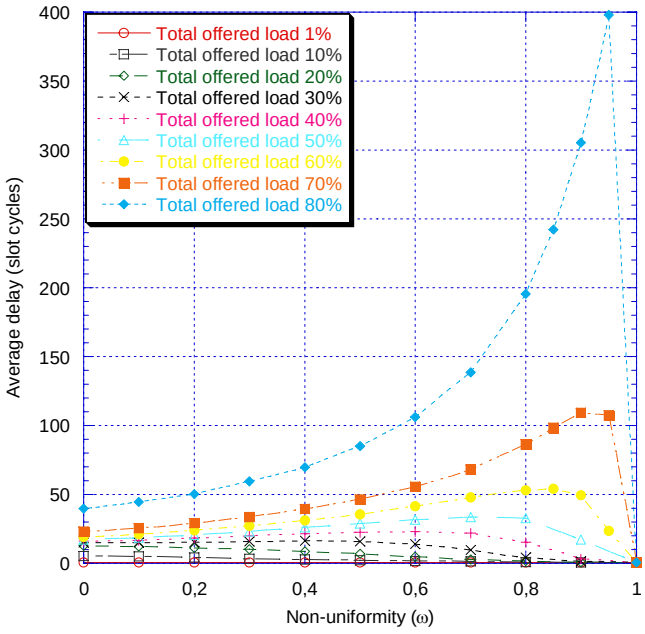

Figure 9 Average delay in slot cycles non-uniform traffic arrival and average burst segment length of 1.

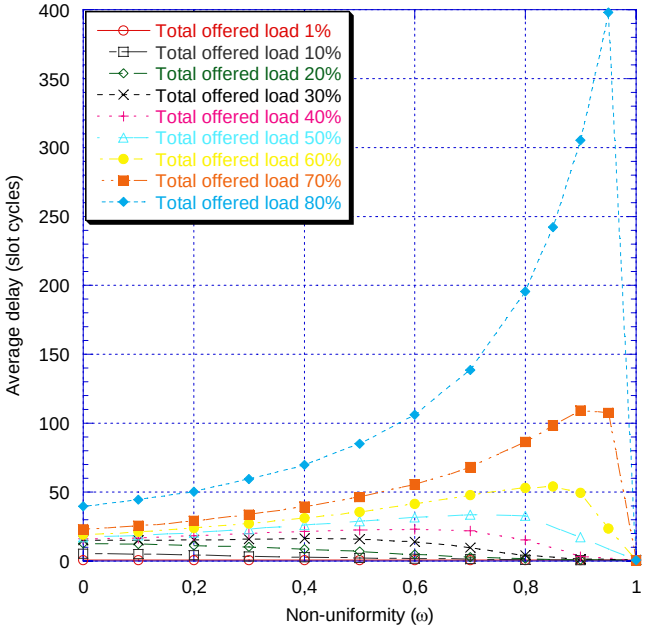

Figure 10 Average delay in slot cycles non-uniform traffic arrival and average burst segment length of 16.

5 Conclusion

Hybrid electro-optical packet switches are promising candidates for reaching aggregate switching bandwidth between 0.5 Tb/s and 5 Tb/s. The potential of reducing power consumption significantly and avoiding conversion between electrical and optical domains are main motivators for this architecture. In this paper a novel strategy for combining aggregating and scheduling using the i-OCF maximal weight match scheduler has been proposed. The strategy shows good performance, and has 100% throughput for the simulation scenarios in this paper, with a speedup of 1.4. Furthermore, it is illustrated through simulations, that getting maximum throughput with i-OCF requires more iterations of the algorithm, than the normal logarithmic relation between number of line cards and iterations used for iSLIP.

References:

[1] K. Kar, D. Stiliadis, T.V. Lakshman, L. Tassiulas, "Scheduling Algorithms for Optical Packet Fabrics", *IEEE Journal on selected areas in communication*, vol. 21, NO. 7, September 2003

[2] X. Li, M. Hamdi, On Scheduling Optical Packet Switches with Reconfiguration Delay", *IEEE Journal on selected areas in communication*, vol. 21, NO. 7, September 2003

[3] L. Dittmann, C. Develder, F. Neri, F. Callegati, Menber, IEEE, W. Koerber, A. Stavdas, M. Renaud, A. Rafel, J. Solé-Pareta, W. Cerroni, N. Legilou, L. Dembeck, B. Mortensen, M. Pickavet, N. Le Sauze, M. Mahony, B. Berde, and G. Eilenberger, "The European IST Project DAVID: A Viable Approach Toward Optical Packet Switching", *IEEE Journal on selected areas in communication*, vol 21, NO. 7, Sebtember 2003

[4] N. McKeown, "Scheduling Algorithm for Input-Queued Cell Switches". *Ph.D. Thesis*, University of California at Berkeley, 1995, USA.

[5] M. A. Marsan, A. Bianco, E. Filippi, P. Giacone, E. Leonardi, F. Neri, "A Comparison of Input Queuing Cell Switch architectures", *Proceedings of the 3rd International Workshop on Broadband Switching Systems*, Kingston, Canada, June 1999.

[6] R. Rojas-Cessa, E.Oki,Z.Jing, H.J.Chao, CIXB-1 Combined input-one-cell-crosspoint buffered, *Workshop on High performance Switching and Routing*, Dallas, Tx, USA, May 2001, 324-239.