Redundant Array of Independent Fabrics -An Architecture for Next Generation Network

Rongsen He and José G. Delgado-Frias
School of Electrical Engineering and Computer Science
Washington State University
Pullman, WA, 99163
{rhe, jdelgado}@eecs.wsu.edu

ABSTRACT. As the next generation network begins to incorporate the Internet, telecommunication and TV services, it becomes one of the most critical infrastructures for our society. Routers construct the skeleton of the network. Their kernel, the structure and configuration (scheduler) of the fabric, dominates the networks' performance, scalability, reliability and cost. Based on research in [1], we proposed an interleaved architecture of multistage switching fabrics in [2], which will meet the requirements for next generation routers. In this paper, we first assess its performance with a theoretical model which complements our simulation results in [2]. Moreover, the interleaved fabrics show great tolerance against internal hardware failures. Based on these properties, we propose the architecture of RAIF (Redundant Array of Independent Fabrics) for next generation network, which could get better performance and fault tolerance as RAID [3].

KEYWORDS. *MIN (multistage interconnection network), Switching fabric, Redundant Array of Independent Fabrics (RAIF).*

1. Introduction

The advantage of packet switching with statistical multiplexing makes the convergence inevitable of the Internet, telecommunication and TV service. For example, in the past few years, Britain has updated its entire telephone network to the Internet Protocol [4]. So there is no technical difference between the telephone network and the Internet in UK. At the same time, both telecommunication carriers and cable companies provide integrated voice, video, and data service for their customers with IPTV [5], which uses IP network to deliver TV program. Thus, the incorporation of next generation network requires that a large number of line cards be integrated in a single high performance router. However, most of present routers are based on single stage crossbar, which suffers from the scalable complexity with $O(N^2)$ (N is the fabric size or input/output number). As a result, these routers only support up to 16×16 interconnection in real applications such as the Cisco 12000 high-end router.

To address these issues of scalability and high performance, we proposed an interleaved architecture of switching fabrics in [2] as shown in Figure 1, which has *Y* panels as switching fabrics (from 1 to *Y*). Each panel is a *N*×*N* I-Cubeout (ICO) network (scalable with the complexity of $O(N\times\log_2 N)$) with recirculation to the last copy of the fabric as shown in Figure 2 [1]. The fabric is composed of *b*×*2b* switching elements (SEs), which has *b* remote outlets to connect to next stage and *b* local outlets to terminate the cells from the switching fabric to the destination queues. The Adjacent stages are interconnected according to the indirect n-cube connecting patterns such that the local SE scheduler just follows the shortest path algorithm with self-routing method as shown in [2]. If the cells fail to get to their destination queue after the primary output (the output of the last stage), we could reenter the cells into the last copy (to avoid intensive collisions at previous copies) of next panel in modular by recirculation. Thus the recirculation flows of panel *i* will go to panel ((*i mod Y*) +1). Because the recirculated cell could get the information of availability directly from prior SE's outlet latch indicator, we can choose the first available point (ICO^{FA}) [1] to fed them into the switching fabrics through the multiplexers. So as shown in Figure 1, at the *N* inputs, *N* demultiplexers distribute the input traffic into each panel synchronized with a clock. At clock cycle *t*, all the input cells at that time will enter the panel ((*t mod Y*) +1). After the multiple switching fabrics are interleaved by the recirculation, the scheme provides another opportunity to balance the traffic; this in turn effectively eases the hot flows after collisions.

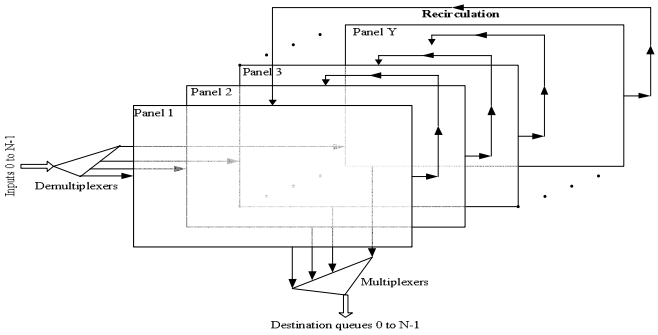

Figure 1. Interleaved multistage switching fabrics.

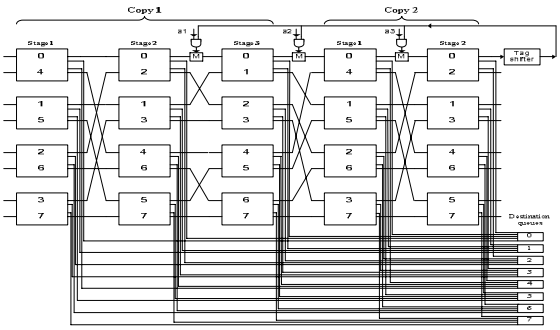

Figure 2. ICO_8 with recirculation

In [2], we already demonstrate the high performance of this novel scheme by simulation under uniform and hot-spot traffic. In Section 2, we will analyze its throughput by a theoretical model for more general cases. Besides with the scalability and high performance of the interleaved architecture, its fault tolerance is still under estimated. We will evaluate its high

reliability with different fault models in Section 3. Based on Section 2 and 3, we bring out the concept RAIF in Section 4. Finally, Section 5 concludes the paper.

2. Analytical model analysis

In the last section, we have introduced the interleaved multistage switching fabrics. Throughput or cell drop rate (throughput =1.0−cell drop rate) is one of the most important parameters to evaluate a switching fabric. Thus, an important issue is to determine the number of panels, which are enough for a real system to obtain good throughput with a reasonable hardware cost. Here, we use our analytical model under uniform traffic to address this issue; we corroborate the model's validity with simulation.

2.1 Analytical model for the single panel fabric

Normally, it is very difficult to analyze the interleaved multistage switching fabrics if the load is non-uniform between each panel or stage. For modeling simplicity, we assume that the traffic between each panel is evenly loaded and the traffic passing from each stage to next one is uniformly distributed to each port. Moreover, we assume no buffer inside the SE between the inter-stage links. With these modeling assumptions, the complex switching system could be decomposed into each single panel with relative independence; consequently, we just need to analyze the throughput of one of them.

We use the recursive method to get the analytical model for single panel fabric as [6, 7, and 8]. The load to the (k+1) stage is computed with the load that is not transferred to the output queues at the k stage. So if we know the random load starting at the first stage, we can compute the load to each stage through the whole fabric. Throughout the paper, the notations in Table 1 are used for the analytical model:

Table1. Notation used in this paper.

Notation	Explanation
X	the stage number for each panel from 1 to X.
Y	the total number of panels from 1 to Y.
N	the fabric size or the input/output number.
b	the switching element size with $b\times 2b$ crossbar.
n	stage number of a full fabric copy, $n=\log_b N$.
P_k	the load offered to stage k, and P_{X+1} is the load for recirculation.
$q_{k,d}$	the load offered to stage k due to cells that have distance d to their destination queues.
O_k	the flow extracted from the fabric to the destination queues at stage k.
F_k	the load from the stage k to next stage. To keep the flow balance, $F_k=P_k-O_k$.
p	the load from outside the fabric, it also means the probability that a cell is generated to the input port during each cycle. So $F_0=p$.
π	the cell drop rate of the fabric.

With the ICO^{FA} mechanism, the cells are dropped if all the recirculation points are not available. In order to calculate the π, we need to compute the load to each stage as follows:

$$F_k = P_k - O_k \text{ and } F_0 = p. \quad (2)$$

$$P_{k+1} = F_k \text{ when } 1 \le k+1 < X-n+1. \quad (3)$$

$$P_{k+1} = F_k + \text{recirculation, when } X-n+1 \le k+1 \le X.$$

To calculate P_{X-n+1} to P_X with recirculation, we use a recursive expression of the form:

$$P_{X-n+1} = F_{X-n} + (1-F_{X-n})\times P_{X+1}$$

$$\vdots$$

$$P_X = F_{X-1} + (1-F_{X-1})\times F_{X-2}\times\cdots\times F_{X-n}\times P_{X+1} \quad (4)$$

$$P_{X+1} = F_X = P_X - O_X \quad (5)$$

Thus,

$$\pi = P_{X+1} - (1-F_{X-n})\times P_{X+1} - \cdots - (1-F_{X-1})\times F_{X-2}\times\cdots\times F_{X-n}\times P_{X+1}$$
$$= P_{X+1}\times F_{X-1}\times F_{X-2}\times\cdots\times F_{X-n} \quad (6)$$

Moreover, each P_k for $1\leqslant k\leqslant X+1$ is composed of $q_{k,d}$ with:

$$P_k = \sum_{d=0}^{n-1} q_{k,d} \quad (7)$$

To evaluate (2)-(5), it is necessary to compute the O_k first. A tagged cell in stage k can exit the fabric only if its distance becomes 0 to the destination queues. Furthermore, one of the following conditions must be met: (i) the tagged cell is the only one requiring a local outlet of the SE or (ii) more than one cell require a local outlet of the SE, but the tagged cell is chosen over the others. Then

$$O_k = q_{k,0}\times[\sum_{m=0}^{b-1}\sum_{h=0}^{m}\binom{b-1}{m}\cdot\binom{m}{h}\cdot q_{k,0}^{h}\cdot(P_k-q_{k,0})^{m-h}\cdot V(h)\cdot(1-P_k)^{b-1-m}] \quad (8)$$

Except the tagged cell at one SE inlet, there are other (b-1) SE inlets from which, h has cells that require local outlets with 0 distance, m-h has cells that require remote outlets to next stage with nonzero distance, and (b-1-m) has no cell for this cycle. $V(h)$ is the probability the tagged cell is chosen in the conflict that may occur if some of the h cells require the same local outlet. So

$$V(h) = \sum_{l=0}^{h}\binom{h}{l}(\frac{1}{b})^{l}\cdot(1-\frac{1}{b})^{h-l}\cdot\frac{1}{l+1} \quad (9)$$

To proceed the load from P_k to P_{k+1}, we need the conditional probability to compute $q_{k+1,d}$ from $q_{k,d}$ distribution for $d=0, 1, \ldots, n-1$.:

$$q_{k+1,d} = \sum_{j=0}^{n-1} P\{q_{k+1,d} \mid q_{k,j}\} q_{k,j} \quad (10)$$

$P\{q_{k+1,d} \mid q_{k,j}\}$ is the conditional probability that a cell has distance d in stage k+1 after it has been switched from stage k where it has distance j. Because we use the shortest algorithm with deflection scheme in the SE's local scheduler, most of the $P\{q_{k+1,d} \mid q_{k,j}\}$ parameters are zero. Depending on the different values of d, three cases are distinguished as follows for (10):

1) $q_{k+1,0} = (q_{k,0} - O_k) + P\{q_{k+1,0} \mid q_{k,1}\} q_{k,1}$ (11)

2) $q_{k+1,j} = P\{q_{k+1,j} \mid q_{k,j+1}\} q_{k,j+1}$, when $0<j<n-1$. (12)

3) $q_{k+1,n-1} = \sum_{j=0}^{n-2}[1-P\{q_{k+1,j} \mid q_{k,j+1}\}] q_{k,j+1}$ (13)

In (11), ($q_{k,0}$ -O_k) are the flows, which failed to reach their destination queues due to collisions, but they still have a zero distance to next stage with the shortest path. In (13), $q_{k+1,n-1}$ collects all the deflected flows with distance n-1. Because we assume that the traffic passing from each stage to next one is

uniformly distributed, we ignore the effects of ($q_{k,0}$ - O_k) to $P\{q_{k+1,j} \mid q_{k,j+1}\}$ and compute $P\{q_{k+1,j} \mid q_{k,j+1}\}$ when $0 \leqslant j < n-1$ as:

$$P\{q_{k+1,j} \mid q_{k,j+1}\} = \sum_{m=0}^{b-1} \sum_{h=0}^{b-1-m} \binom{b-1}{m} \cdot \binom{b-1-m}{h} \cdot q_{k,j+1}^{h} \cdot (\sum_{i=1}^{j} q_{k,i})^{m} \cdot T(b-1-m-h, j+1) \cdot [1-D(m,h)] \quad (14)$$

Except the tagged cell at one SE inlet, there are (b-1) other SE inlets from which, h has cells that require remote outlets with distance j+1, m has cells that require remote outlets with distance from 1 to j.

T(z,i) is the probability that z inlets of the SE has the conditions as follows: (1) Empty, (2) Kept busy by cells with distance 0, or (3) Kept busy by cells with distance $d>i$. Taking into account these conditions we have that:

$$T(z,i) = \sum_{l=0}^{z} \sum_{r=0}^{z-l} \binom{z}{l} \cdot \binom{z-l}{r} \cdot q_{k,0}^{z-l-r} \cdot (\sum_{t=i+1}^{n-1} q_{k,t})^{l} \cdot (1-P_k)^{r} \quad (15)$$

D(m,h) is the probability that the tagged cell is deflected, if m cells with lower distance and h cells with equal distance are switched to next stage by the SE:

$$D(m,h) = \sum_{l=0}^{m} \binom{m}{l} \cdot (\frac{1}{b})^{l} \cdot (1-\frac{1}{b})^{m-l} + (1-\frac{1}{b})^{m} \cdot \sum_{l=1}^{h} \binom{h}{l} \cdot (\frac{1}{b})^{l} \cdot (1-\frac{1}{b})^{h-l} \cdot \frac{l}{l+1} \quad (16)$$

From (7) to (16), we could compute the load of each stage without recirculation. For the last copy of the fabric from stage (X-n+1) to X, we should count in the recirculation load from the final stage P_{X+1}. So, based on (4) and (11)-(13), the distributions for $0 \leqslant j \leqslant n-1$ when $X-n+1 \leqslant k \leqslant X$ are computed:

$$q_{k,j} = q_{k,j} + (1-F_{k-1}) \times F_{k-2} \times \cdots \times F_{X-n} \times (\sum_{i=0}^{n-1} e_{k,j,i} \cdot q_{X+1,i}) \quad (17)$$

In (17), the $q_{k,j}$ on the right hand comes from previous stage and is calculated with equations (11)-(13). The second term belongs to the recirculation load. The coefficient $e_{k,j,i}$ is determined by the fabric structure (N and b) and the recirculation point. For each stage k between $X-n+1 \leqslant k \leqslant X$, the $[e_{k,j,i}]_{j \times i}$ form a 2-dimentional coefficient matrix with row j and column i, $0 \leqslant j, i \leqslant n-1$. As an example, when N=256, b=4 and n=4, the four coefficient matrixes are shown below:

$$[e_{X-3,j,i}]_{j \times i} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad [e_{X-2,j,i}]_{j \times i} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \\ \frac{3}{4} & \frac{3}{4} & \frac{3}{4} & \frac{3}{4} \end{bmatrix},$$

$$[e_{X-1,j,i}]_{j \times i} = \begin{bmatrix} \frac{1}{4} & 0 & \frac{1}{16} & 0 \\ 0 & 0 & 0 & \frac{1}{16} \\ \frac{3}{4} & 0 & \frac{3}{16} & \frac{3}{16} \\ 0 & 1 & \frac{3}{4} & \frac{3}{4} \end{bmatrix} \quad [e_{X,j,i}]_{j \times i} = \begin{bmatrix} \frac{1}{4} & 0 & 0 & \frac{1}{64} \\ \frac{3}{4} & 0 & 0 & \frac{3}{64} \\ 0 & 1 & 0 & \frac{3}{16} \\ 0 & 0 & 1 & \frac{3}{4} \end{bmatrix}.$$

We could compute each item in matrix $[e_{k,j,i}]_{j \times i}$ from Table 2. Because there is no address change when cells are recirculated back to stage X-3, $[e_{X-3,j,i}]_{j \times i}$ is the identity matrix. Otherwise, the item $e_{k,j,i}$ depends on the recirculation point. For example, corresponding with $q_{X+1,3}$ to stage X, the third bit position A_2A_1must be nonzero. After the cyclic rotation of the address to stage X, there are four cases as follows:

1) When A_8A_7=A_6A_5=A_4A_3=0, $e_{X,0,3}$ =1/64 as distance 0.
2) When $A_8A_7 \neq 0$, and A_6A_5=A_4A_3=0, $e_{X,1,3}$ =3/64 as distance 1.
3) When $A_6A_5 \neq 0$, and A_4A_3=0, $e_{X,2,3}$ =3/16 as distance 2.
4) When $A_4A_3 \neq 0$, $e_{X,3,3}$ =3/4 as distance 3.

Table 2 Address cyclic rotation for recirculation (N=256, b=4).

Stage \ Distance	0	1	2	3
From $X+1$	A_8A_7	A_6A_5	A_4A_3	A_2A_1
To X-3	A_8A_7	A_6A_5	A_4A_3	A_2A_1
To X-2	A_6A_5	A_4A_3	A_2A_1	A_8A_7
To X-1	A_4A_3	A_2A_1	A_8A_7	A_6A_5
To X	A_2A_1	A_8A_7	A_6A_5	A_4A_3

Based on equation (2)-(17), we compute P_k, $q_{k,d}$ and O_k recursively until they become steady. Then with (6), the drop rate π (or throughput) of the single panel fabric is obtained.

2.2 Analytical model for the interleaved switching fabrics

We have introduced an analytical model for the single panel fabric in last section and assumed that the traffic between panels is evenly loaded. Thus, each panel runs at load (p/Y), in which p is the load from outside and Y is the number of panels. The total cell drop rate (or throughput) is the sum of that from each panel. For example, when Y=2 and p=1.0, each panel runs at load= 0.5 and the total cell drop rate equals to 2 × (cell drop rate at load 0.5 for single panel). When Y=3 and p=0.9, each panel runs at load= 0.3 and the total cell drop rate equals to 3 × (cell drop rate at load 0.3 for single panel). The benefit of the interleaved switching fabrics comes from that it avoids the high non-linear increase during load portion ($0.5 \leqslant p \leqslant 1$) for the single panel fabric. The interleaved architecture, which runs with (p/Y) $\leqslant$0.5 for each panel, replaces this non-linear portion with linear increase of Y for throughput.

2.3 Validation of the analytical model

The accuracy of the analytical models in Sections 2.1 and 2.2 has been assessed by comparing its results with those from simulations of the system. Figure 3 shows the cell drop rate under uniform traffic with N=256, b=4 and X=4. In this paper, we use the notation SX/PY where X specifies the number of stages per panel and Y the number of panels. We choose N=256, b=4 and X=4 because it is a full copy of fabric which will make the results more pronounced.

From Figure 3, there is a difference between the analytical model and simulation results for S4/P1 when load is $0.3<p<0.9$. As mentioned in [6, 7], the basic reason is that the traffic between the stages is unbalanced and the assumption of uniform distribution does not hold any more. When load is very low at p<0.3, the traffic between stages could be still considered as uniform. When load is high enough after p>0.9, all inter-stage links are saturated with traffic from previous stage or recirculation. Thus the traffic between stages could be considered as uniform again from outside view. As a result, the diagram shows satisfactory matching between models and simulation for these two load portions. As to S4/P2 and S4/P3, the lower load to each panel, doubled or tripled recirculation

path, and the interleaved connections make the traffic uniform between stages. Thus, there is a good match between the models and simulation.

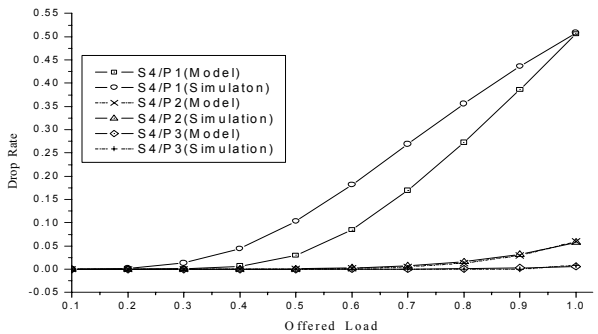

Figure 3. Drop rate vs offered load for analytical models and simulation, N=256, b=4 and X=4.

To address the design issue at the beginning of Section 2, we investigate the optimal number Y for good throughput over hardware cost. For Y and $Y+1$, the load to each panel is p/Y and $p/(Y+1)$ respectively. The load difference is $p/Y - p/(Y+1) = p/(Y^2+Y)$, which will decrease quickly with increase of Y. That means the *marginal* gain of throughput will decrease accordingly with the hardware cost around $1/Y^2$. So Y=2 is the optimum number for throughput over hardware cost. Based on the analysis above, we choose Y=2 for the extensive simulations in Sections 3 and [2]. However, Y=3 or higher is preferred for fault tolerant reasons (Section 4).

3. Fault tolerance of the interleaved architecture

A critical design aspect of high performance routers is their reliability. Though the Internet itself is designed to tolerate failure of some router nodes, lost of core routers still results in considerable congestion to other routers with unbalanced traffic. Moreover, some subnets, which connect through the failure node, will be made unreachable. On the other side, VLSI moves into the nanometer range, this in turn provides faster systems and higher integration. However, the devices suffer from extreme process variation, particle-induced transient errors, and transistor wear-out. In the near future it will be unlikely to avoid having faults in VLSI systems.

As to our interleaved switching fabrics, its parallel architecture already has built-in redundancy that in turn provides fault tolerance. Our scheme treats a faulty element in a similar fashion as the hot congestion area, and deflects the traffic away from it. There are link and SE failures inside the switching fabrics. SE hardware is much more complex and, therefore, more prone to faults than the internal link connections. An internal link fault could also be modeled as a SE fault since a faulty link renders the following SE as a nonworkable unit. Thus in this paper, we use the SE fault model to evaluate its detrimental effects to system performance.

Following the simulation models in [1, 2], we choose fabric speedup ξ=2 and each SE output queue (either local or remote one) is equipped with 12-cell buffers. For all results, 200,000 system clocks are simulated, which are long enough to get steady state results.

3.1 Single fault model

In the single fault model, we have a SE that is faulty which could not accept any cells. Thus, the cells that need to pass the faulty SE are deflected in prior stage. If the faulty SE is located in last copy of the fabrics, the recirculated cells need to jump this faulty point in case of the ICO^{FA} approach. However, we have observed that the stage location of the faulty SE determines the degradation to the performance other than the row location. Thus, in our simulations the faulty SE is placed at the same row but in different stages. For the results reported here we have chosen row 31 (and different stages) for symmetry purpose. Because hot–spot traffic saturates part of fabrics along its path, the effect will depend on the location of the faulty SE. Thus, we will just use uniform traffic for fault tolerant test throughout the paper.

In this section, we compare how a fault impacts the performance of the single panel and interleaved double panels; both single and double panels have the same length of 6 stages. Figure 4 and 5 show the results with an increasing load in the x-axis. The fault stage locations are labeled in parenthesis.

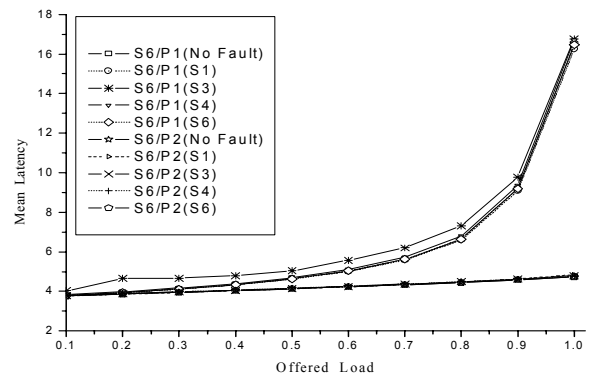

Figure 4. Mean latency vs offered load for single fault test.

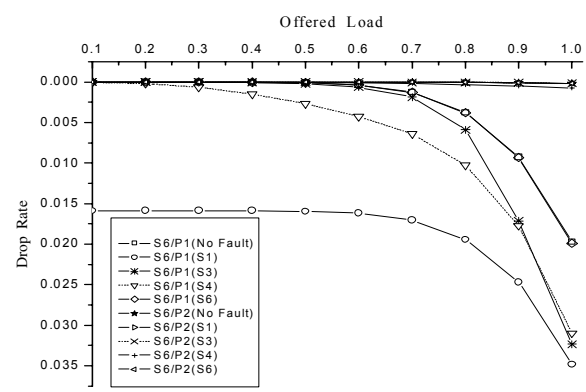

Figure 5. Drop rate vs offered load for single fault test.

It is observed that the fault location determines the degradation to S6/P1 especially for drop rate. In Figure 5, if the faulty SE is in the first stage, the drop rate will start from 1.59%. There are total 64 SEs in each stage, 1/64= 1.56%. One faulty SE means 1.56% of the input traffic will be lost immediately without switching through the fabrics. Thus, the simulation results match the theoretical value and prove the

first stage is the most critical for single panel fabric.Another critical stage of S6/P1 is the first stage of last copy, which also merges with the first entrance point of the FA approach. If the fault happens to be this stage, both the latency and drop rate deteriorate noticeably, with a 3.23% drop rate as opposed to 1.97% for fault-free situation. Finally, the performance is insensitive to the last stage fault of S6/P1, considering most of cells have been switched to destination queues in prior stages.

As expected, the faulty SE exhibits negligible impacts to performance of S6/P2 regardless of their fault locations. The interleaved fabrics in parallel not only substantially enhance the performance, but also tolerate the single hardware failure. Even for the fatal fault in first stage of S6/P1, S6/P2 still gives the inputs another chance to divert the flows into the fabrics.

3.2 Multiple fault model

The multiple fault model is considerably more complicated than the single fault model, because of the abundant combinations of the number of faults and locations. However, the faults in the identical stage of different copies will generate a switching bottleneck and make the performance to deteriorate significantly, since all of them correct the same position of the tag bits. Thus, the simulations in Figures 6 and 7 depict this situation. As before, we specify the fault location in the parenthesis. For single panel, S8/P1(S4+S8) means that we choose faults at stage 4 and 8. For interleaved double panels, S6/P2[(S2+S6)/P1+S2/P2] means that we choose faults at stage 2 and 6 of panel 1 and stage 2 of panel 2. S6/P2[(S2+S6)/(P1+P2)] means that we choose faults at stage 2 and 6 for both panels, etc. Since faults have a strong tendency to happen in continuous (or nearby) areas, we assume there are continuous 4 faulty SEs in each stage and locate them at row 30,31,32,33 for symmetry purpose.

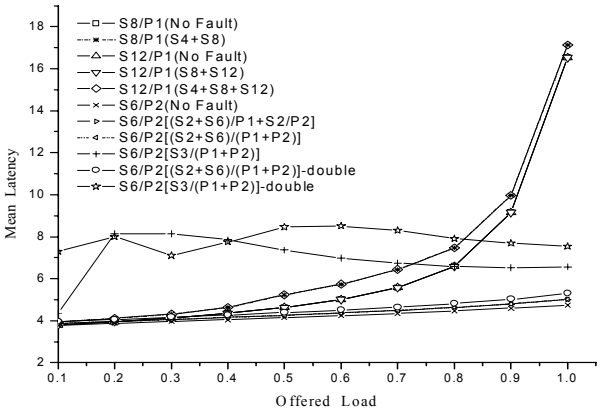

Figure 6. Mean latency vs offered load for multiple faults test.

As it can be observed in Figures 6 and 7, S8/P1(S4+S8) and S12/P1(S4+S8+S12) are considerably impacted by this fault model. First of all, more delay latency pulls the curve above the fault-free ones. Then their drop rate jumps quickly from 9.8×10^{-8} of fault–free to 3.4% at full load p=1.0. However, the redundant stage in first copy of S12/P1(S8+S12) still provides capabilities to tolerate faults in stage 8 and 12.

For S6/P2, stages 2 and 6 of S6/P2 (total of 4 affected stages) correct the second position of the tag bits. It is reasonable that one or more redundant stages in S6/P2[(S2+S6)/P1] and S6/P2[(S2+S6)/P1+S2/P2] will compensate the faults with slight degradations in latency and drop rate. Stage 3 of S6/P2 (total of 2 affected stages) corrects the third position of the tag bits. If both of them fail, one can expect inferior performance as S8/P1(S4+S8) and S12/P1(S4+S8+S12) exhibit before. However, it is important to notice that both S6/P2[S3/(P1+P2)] and S6/P2[(S2+S6)/(P1+P2)] cause a negligible increase in drop rate, 0.086% as opposed to 0.013% of fault-free.

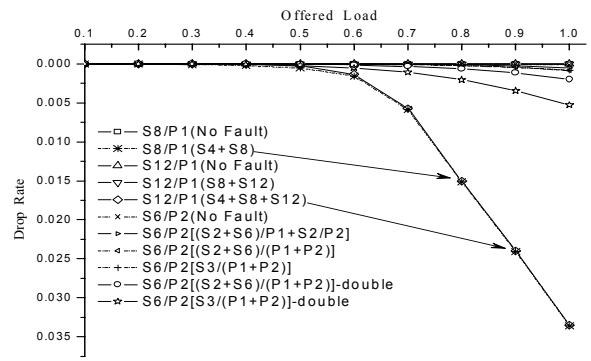

Figure 7. Drop rate vs offered load for multiple faults test.

In order to understand why the interleaved architecture is fault tolerant, we consider first the two points where cells are dropped. One is in the SE itself when all local buffers are full. For example, if SE at row 0 stage 2 in Figure 2 fails, all flows from input 0 and 4 will merge and go through port 4 of SE at row 0 stage 1. The intense flows fill the local buffers quickly and make dropping cells unavoidable. The second point is at the FA entrance points. If all FA entrance points are not available, the recirculated cells will be dropped. Compared with the single panel architecture, S6/P2 firstly reduces the traffic to half for each panel, and then it doubles the FA recirculation points which broaden the switching path and mitigate collisions. Though correcting the deflected cells of S6/P2[S3/(P1+P2)] increases latency a little to 6.6 cycles as opposed to 4.7 cycles of fault-free as shown in Figure 6; this latency is still far below its counterpart of S12/P1. Moreover, with total four stages to correct the second position of tag bits, S6/P2[(S2+S6)/(P1+P2)] still achieve remarkable performance with low latency and drop rate.

Furthermore, we have even performed simulations with some extreme cases which double the faults from row 28 to 35 to a total of 8 faults in specific stages. In Figures 6 and 7, S6/P2[(S2+S6)/(P1+P2)]-double and S6/P2[S3/(P1+P2)]-double show this situation at full load p=1.0, with 5.3 cycles latency and 0.2% drop rate for S6/P2[(S2+S6)/(P1+P2)]-double, and 7.5 cycles latency and 0.5% drop rate for S6/P2[S3/(P1+P2)]-double. Again, the interleaved architecture shows much stronger capability of tolerating fault than its counterpart of the single panel fabric.

3.3 Broken test for fabrics

In this section, we will conduct the broken test to our interleaved switching fabrics. We use the S6/P2 for the simulation and test the critical stage *combinations* S6/P2[(S2+S6)/(P1+P2)] and S6/P2[S3/(P1+P2)]. We will

adopt *C1* to denote for S6/P2[S3/(P1+P2)] and *C2* for S6/P2[(S2+S6)/(P1+P2)] as abbreviations. We double the faulty SEs each time from 4 to 64. The faulty SEs are located symmetrically in central of each stage as before. Figures 8 and 9 depict these broken tests gradually to extreme.

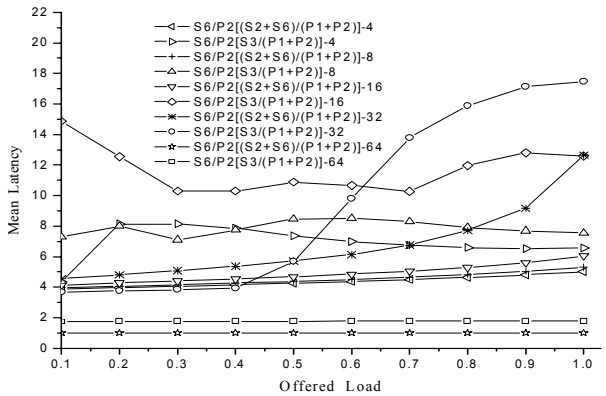

Figure 8. Mean latency vs offered load for broken test

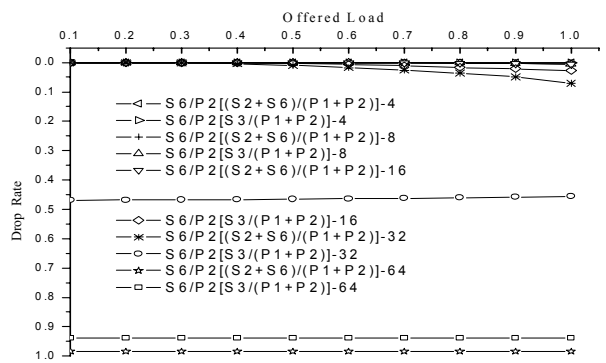

Figure 9. Drop rate vs offered load for broken test.

So under the worst case when 256/4=64 SEs are faulty (it means the whole stage are nonworkable), the drop rates of *C2*-64 and *C1*-64 hold 98.4% and 93.75% respectively, which match the theoretical values of (1-1/64) and (1-1/16). And the mean latency of *C2*-64 and *C1*-64 keep 1.0 and 1.77 cycles respectively, which match the theoretical values of 1.0 and $[1\times(1/64)+2\times(1/16-1/64)]/(1/16) = 1.75$ cycles. When 32 SEs are faulty per stage, *C1*-32 keeps the drop rate around 46% because half of (1-1/16) flows will never have a chance to be switched on the third stage. Similar as Figures 6 and 7, *C2* still exhibit satisfactory performance even half of SEs are broken in the specific stages. For other broken test cases, our interleaved architecture shows its remarkable reliability!

4. RAIF (Redundant Array of Independent Fabrics)

From the simulations and analysis in [2] and sections above, as a good example of interleaved architecture, S6/P2 shows a better performance and much stronger capability to tolerate internal hardware failures than the single panel architecture. Specifically, each panel in S6/P2 will be a switching board which is relatively independent as mentioned in Section 2. Even under the worst case scenario where one panel is broken; the other panel in S6/P2 will allow the router to continue running with some performance degradation. On the other hand, in the case of S12/P1 whole system will malfunction. Moreover, inspired by RAID (Redundant Array of Independent Disks) technology [3], we could build a Redundant Array of Independent Fabrics (RAIF) by upgrading the S6/P1 with more panels in parallel. Thus, each switching panel in RAIF works as similar as a hard disk in RAID. The extra panels could work as RAIF 0 (similar as RAID level 0): working in parallel as S6/P2. Other alternative, one additional panel works as RAIF 1 (similar as RAID level 1): stand by; this will help to lower power consumption while there is no fault and this panel will replace a malfunctioning one when fault happens. Combined with RAIF 0 and 1, we could build the RAIF 2 with Y panels ($Y>2$): $Y-1$ panels working in parallel and one panel stand by for fault tolerance. In general, the RAIF provide a flexible scalability with fault tolerance and graceful performance degradation.

5. Concluding Remarks

In [2], we have presented a novel architecture of interleaved switching fabrics for scalable high performance routers. In this paper, we present an analytical model to assess its throughput. The benefit of the interleaved architecture comes from avoiding the high non-linear increase during load portion ($0.5 \leqslant p \leqslant 1$) for the single panel fabric. The simulations in [2] also demonstrate our analysis here.

Moreover, our scheme treats a faulty element in a similar fashion as the hot congestion area, and deflects the traffic away from it. Extensive simulations under different faulty models have reveled that the interleaved multistage switching fabrics are highly fault tolerant against internal hardware failures that single panel fabric does not achieve. Under the worst case, the single panel fabric drops all packets when faults are located at the first stage. In general, it is possible to build a reliable, scalable high performance switching system using a Redundant Array of Independent Fabrics (RAIF) scheme in a similar fashion as RAID [3].

References

[1] N.-F. Tzeng, "Multistage-based switching fabrics for scalable routers," *IEEE Transactions on Parallel and Distributed Systems,* Vol. 15, No. 4, pp. 304-318, April 2004.

[2] R. He and J. Delgado-Frias, "Interleaved multistage switching fabrics for scalable high performance routers," *Proceedings of 49th IEEE GLOBECOM conference*, San Francisco, CA, November, 2006.

[3] D. A. Patterson, G. A. Gibson, and R. H. Katz, "A case for redundant arrays of inexpensive disks (RAID)," *ACM SIGMOD Conf. Proc., Chicago,* pp. 109-116, June. 1988.

[4] S. Cherry, "Noting But Net," *IEEE Spectrum,* Vol. 44, No. 1, pp. 22-26, Jan.2007.

[5] S. Jr. Ortiz, "Phone Companies Get into the TV Business," *IEEE Computer,* Vol. 39, No. 10, pp. 12-15, Oct.2006.

[6] S. Bassi, M. Decina, P. Giacomazzi, and A. Pattavina, "Multistage shuffle networks with shortest path and deflection routing for high performance ATM switching: The open-loop Shuffleout," *IEEE Trans on Communication,* Vol. 42, No. 10, pp. 2881-2889, Oct.1994.

[7] M. Decina, P. Giacomazzi, and A. Pattavina,, "Multistage shuffle networks with shortest path and deflection routing for high performance ATM switching: The closed-loop Shuffleout," *IEEE Trans on Communication,* Vol. 42, No. 11, pp. 3034-3044, Nov. 1994.

[8] A. Pattavina, *Switching Theory: Architecture and Performance in Broadband ATM Networks,* Wiley, 1998.