VLAN-based Minimal Paths in PC Cluster with Ethernet on Mesh and Torus

Tomohiro Otsuka   Michihiro Koibuchi[*]   Akiya Jouraku   Hideharu Amano
Department of Information and Computer Science, Keio University
3-14-1, Hiyoshi, Kohoku-ku, Yokohama, 223-8522 Japan
{terry,koibuchi,jouraku,hunga}@am.ics.keio.ac.jp

Abstract

In a PC cluster with Ethernet, well-distributed multiple paths among hosts can be obtained by applying VLAN technology. In this paper, we propose VLAN topology sets and path assignment methods in mesh and torus. The proposed VLAN-based methods on mesh require N^{M-1} and $\lfloor N^{M-1}/2 \rfloor + 1$ VLANs to provide balanced minimal paths and partially balanced ones respectively, where N is the number of switches per dimension and M is the number of dimensions. Similarly, those on torus require $2N^{M-1}$ and $N^{M-1}+2$ VLANs respectively. Simulation results show that the proposed methods improve up to 902% and 706% of throughput respectively.

1. Introduction

Ethernet has been used to connect hosts in PC clusters, because of its high performance per cost. Unlike the early Beowulf clusters, recent PC clusters employ system software[6][7] which enables zero- or one-copy communication used in system area networks (SANs)[2][3]. In background of enabling the simplified software stack at hosts to provide low-latency and high-bandwidth communication, high-throughput switching fabrics are recently implemented; indeed, a packet is not often discarded. In addition, link bandwidth of Ethernet is rapidly increased, such as GbE or 10GbE standardization, as CPU computation power is increased. Thus, Ethernet has become an alternative network for high-performance PC clusters, and Ethernet topology and its routing paths will become one of crucial components to build a large-scale system.

However, most of PC clusters using Ethernet have employed simple tree-based topologies, since the spanning tree protocol (STP) logically requires acyclic topologies for dynamic host configuration. This is because Ethernet technology is not originally designed for high-performance computing or parallel computing.

Thus, even when building a typical topology for parallel computers, such as a torus, PC clusters with Ethernet must accept non-minimal embedded-tree paths, and links which do not belong to a spanning tree cannot be used. That is, well-distributed minimal paths, such as the dimension-order routing[1] studied for parallel computers and SANs cannot be applied. We consider that this is the reason why various topologies for parallel processing have not been focused in research field of PC clusters with Ethernet.

Kudoh et al. proposed to apply VLAN technology to PC clusters with Ethernet so as to employ multiple paths between switches under various topologies including fat-tree, mesh and hyper crossbar[4][5]. They also showed that VLAN-based routing made the best use of link bandwidth under well-distributed paths.

VLAN technology is originally not for increasing network throughput, but for partitioning hosts into multiple groups. However, VLAN can also be used to provide multiple paths between hosts to increase throughput as follows: all VLAN groups are extended to include all hosts, and different link sets are assigned to each VLAN topology. In this case, all pairs of hosts can communicate via any VLAN group. Thus, multiple paths which include different links are available between each pair of hosts.

Although each VLAN topology is logically a tree, by introducing multiple VLANs each of which consists of a different set of links, a flexible physical path set including all links can be employed. For example, Figure 1 shows three examples of VLAN topologies, which include all hosts and different link sets. As shown in this figure, minimal paths can be taken by selecting a suitable VLAN topology from (b), (c) and (d) at a source host.

Note that each path is assigned into a single VLAN, and each source host must indicate a VLAN number corresponding to a path. Thus, Ethernet physical topology is free from tree-based structures with VLAN technology. VLAN-based routing can be conducted by L2 Ethernet[5], and it is likely to be supported by the low-level communication library in PC clusters[6].

[*] Presently with National Institute of Informatics, koibuchi@nii.ac.jp

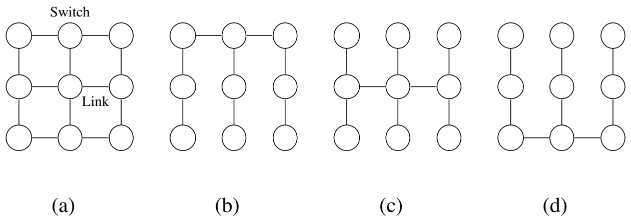


Figure 1. Examples of VLAN topologies for 3×3 2-D mesh

Although IEEE 802.1Q VLAN tag field can identify 4,094 ($2^{12}-2$) VLANs, commercial cost-effective Ethernet switches support only a limited number of VLANs, which would be a limiting factor in an implementation and extension of PC clusters. Since balanced minimal paths, which decrease packet collisions, are essential to improve performance in PC clusters, an efficient strategy to employ VLAN topologies is required for PC clusters with Ethernet.

In this paper, we propose VLAN topology sets and path assignment methods to them in mesh and torus. The rest of this paper is organized as follows. In Section 2, we show VLAN topology sets and path assignment methods in mesh, and we also show them in torus in Section 3. In Section 4, evaluation results of VLAN-based paths and the STP-based paths are shown, and in Section 5, the conclusion is presented.

2. VLAN-based Minimal Paths on Mesh

In this section, we show two VLAN-based methods to take balanced or partially balanced minimal paths. The first one ensures minimal and well-distributed paths, which are the same as those of the dimension-order routing (DOR)[1], which is known as a method to make well-distributed paths. In the dimension-order routing on two-dimensional (2-D) mesh, packets are forwarded to x-direction with required hops first, and then forwarded to y-direction. The second method guarantees minimal paths with a slight loss of path uniformity, while the number of required VLANs is about a half of that in the first method.

2.1. Preliminary

Figure 2 shows 4×4 2-D mesh and three examples of VLAN topologies for it. In the figure, each vertex and arc represents Ethernet switch and link respectively. Although it is possible that some hosts are connected to each switch, they are omitted here.

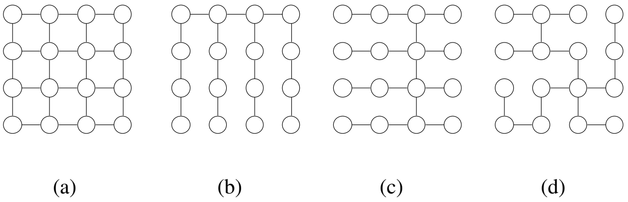


Figure 2. 4×4 2-D mesh and examples of its VLAN topologies

Definition 2.1 (2-D mesh) *Assign a number with a two-dimensional coordinate (x, y) where $0 \leq x < N$ and $0 \leq y < N$ to each switch. By connecting vertex (x, y) with vertexes $(x+1, y)$, $(x-1, y)$, $(x, y+1)$ and $(x, y-1)$, if $x+1 < N$, $x-1 \geq 0$, $y+1 < N$ and $y-1 \geq 0$ respectively, an $N \times N$ two-dimensional mesh is formed.* □

Two-dimensional mesh treated here is commonly defined as an N-ary 2-cube, and its numbering is as follows:

$$\begin{array}{cccc} (0,0) & (1,0) & \cdots & (N-1,0) \\ (0,1) & (1,1) & \cdots & (N-1,1) \\ \vdots & \vdots & \ddots & \vdots \\ (0,N-1) & (1,N-1) & \cdots & (N-1,N-1) \end{array}$$

Each of VLAN topologies in Figure 2 is a spanning tree of physical network (a), and consists of N^2 switches and N^2-1 links. As shown in Figure 2, there are various alternative VLAN topologies (trees) in mesh. However, it is difficult to establish a simple minimal-path set using trees with low regularity, such as (d), in combination with other VLAN topologies. Therefore, VLAN topologies in the proposed methods are based on simple topologies similar to (b) or (c). In order to identify each VLAN topology used in the proposed methods, we use the following notation.

Definition 2.2 (linear connection in 2-D mesh) *A vertical connection in 2-D mesh is represented as $l(x, -)$. That is, vertex (x, y) is connected with vertexes $(x, y+1)$ and $(x, y-1)$, if $y+1 < N$ and $y > 0$ respectively, in the connection. Similarly, a horizontal connection in 2-D mesh is represented as $l(-, y)$. That is, vertex (x, y) is connected with $(x+1, y)$ and $(x-1, y)$, if $x+1 < N$ and $x > 0$ respectively.* □

A VLAN topology can be formed by a single linear connection in a dimension and all linear connections in the opposite dimension. For example, the VLAN (b) in Figure 2 consists of connections $l(0, -)$, $l(1, -)$, $l(2, -)$, $l(3, -)$ and $l(-, 0)$, while the VLAN (c) consists of connections $l(2, -)$, $l(-, 0)$, $l(-, 1)$, $l(-, 2)$ and $l(-, 3)$. Such a VLAN topology is represented as the following notation.

Definition 2.3 (VLAN topology in 2-D mesh) *Each of the VLAN* $\mathrm{VL}(-, y_0)$ *and* $\mathrm{VL}(x_0, -)$ *consists of all switches (vertexes) and the following set of linear connections:*

$$\begin{aligned} \mathrm{VL}(-, y_0): &\quad \left\{ l(x, -) \;\middle|\; 0 \le x < N \right\} \cup \left\{ l(-, y_0) \right\} \\ \mathrm{VL}(x_0, -): &\quad \left\{ l(-, y) \;\middle|\; 0 \le y < N \right\} \cup \left\{ l(x_0, -) \right\} \end{aligned}$$

□

2.2. Minimal Paths for the DOR in 2-D Mesh

Definition 2.4 (DOR VLANs in 2-D mesh) *The DOR VLAN set consists of the following* N *VLANs in* $N \times N$ *mesh:*

$$\left\{ \mathrm{VL}(-, y) \;\middle|\; 0 \le y < N \right\}$$

□

A path from a source switch (x_S, y_S) is assigned into the VLAN $\mathrm{VL}(-, y_S)$. An example of the DOR VLANs is shown in Figure 3. Figure 3 shows that four VLANs, $\mathrm{VL}(-, 0)$, $\mathrm{VL}(-, 1)$, $\mathrm{VL}(-, 2)$ and $\mathrm{VL}(-, 3)$, are employed to take the DOR paths in 4×4 2-D mesh.

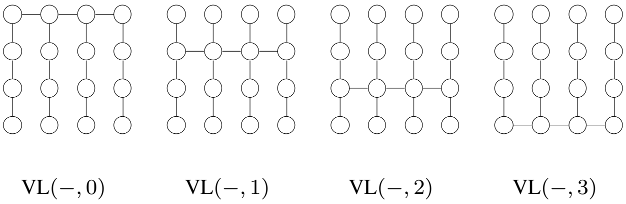


Figure 3. The DOR VLANs in 4×4 2-D mesh

Theorem 2.1 *The DOR VLANs provide the same minimal-path set as that of the dimension-order routing in 2-D mesh.*

Proof A VLAN $\mathrm{VL}(-, y_S)$ consists of a horizontal connection $l(-, y_S)$ and all of N vertical connections. Thus, on $\mathrm{VL}(-, y_S)$, all paths from a source switch (x_S, y_S) are minimal along the dimension-order routing. Since there are N VLANs $\left\{ \mathrm{VL}(-, y) \;\middle|\; 0 \le y < N \right\}$, paths from a source switch (x_S, y_S) on $\mathrm{VL}(-, y_S)$ are the same as that of the dimension-order routing. □

Note that an Ethernet switch does not have dedicated channel buffers and flow control mechanism for VLANs, unlike virtual channels in interconnection networks for parallel computers. Thus, the path distribution among VLANs hardly affects network performance, and VLAN selection for minimal paths from a source switch (x_S, y_S) to a destination switch (x_D, y_D), which can be assigned to different VLANs, is trivial.

2.3. Minimal Paths with Partial DOR (PDOR) in 2-D Mesh

We show the second method whose path set is similar to that of the dimension-order routing.

Definition 2.5 (PDOR VLANs in 2-D mesh) *The PDOR VLAN set consists of the following* $\lfloor N/2 \rfloor + 1$ *VLANs in* $N \times N$ *mesh:*

$$\left\{ \mathrm{VL}(-, 2i{+}1) \;\middle|\; 0 \le i < \lfloor N/2 \rfloor \right\} \cup \left\{ \mathrm{VL}(x_0, -) \right\}$$

□

Figure 4 shows an example of the PDOR VLANs in 4×4 2-D mesh. This method employs only three VLANs, $\mathrm{VL}(-, 1)$, $\mathrm{VL}(-, 3)$ and $\mathrm{VL}(1, -)$, by taking different minimal-path set from that of the DOR. The selection of the value x_0 is trivial, because links of the vertical connection in $\mathrm{VL}(x_0, -)$ are never used by paths.

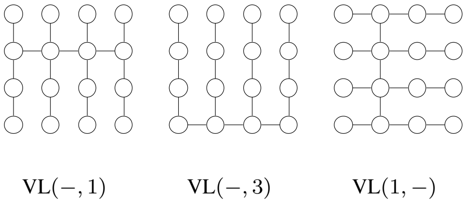


Figure 4. The PDOR VLANs in 4×4 2-D mesh

Compared with the DOR VLANs, $\lfloor (N{+}1)/2 \rfloor$ VLANs $\left\{ \mathrm{VL}(-, 2i) \;\middle|\; 0 \le i < \lfloor (N+1)/2 \rfloor \right\}$ are deleted, and $\mathrm{VL}(x_0, -)$ is newly employed.

As illustrated in $\mathrm{VL}(-, 1)$ and $\mathrm{VL}(-, 3)$ in Figure 4, on the VLAN $\mathrm{VL}(-, 2i{+}1)$, all paths from a switch $(x, 2i{+}1)$ to all destination switches take minimal paths along the dimension-order routing. On the other hand, paths from a switch $(x, 2i)$ cannot be along the dimension-order routing for all destinations, because of lack of the VLAN $\mathrm{VL}(-, 2i)$. Thus, in order to take minimal paths, each source switch (x_S, y_S) uses one of appropriate VLANs selected by the following procedure for a destination switch (x_D, y_D).

```
if $y_S$ mod $2 = 1$ then use $\mathrm{VL}(-, y_S)$ ;
else if $y_D < y_S$ then use $\mathrm{VL}(-, y_S{-}1)$ ;
else if $y_D > y_S$ then use $\mathrm{VL}(-, y_S{+}1)$ ;
else $\{y_D = y_S\}$ use $\mathrm{VL}(x_0, -)$ ;
```

Theorem 2.2 *The PDOR VLANs provide a minimal-path set in 2-D mesh.*

Proof Since $\mathrm{VL}(x_0, -)$ has all horizontal connections, paths from a source switch $(x_S, 2i)$ to a destination switch

$(x, 2i)$ on VL$(x_0, -)$ are minimal along one of the horizontal connections. For other destinations, paths from a source switch $(x_S, 2i)$ are minimal via switch $(x_S, 2i+1)$ on VL$(-, 2i+1)$ for $y_D > y_S$, and those are minimal via switch $(x_S, 2i-1)$ on VL$(-, 2i-1)$ for $y_D < y_S$.

On the other hand, according to Theorem 2.1, paths from a source switch $(x_S, 2i+1)$ on VL$(-, 2i+1)$ are minimal. Since both $(x_S, 2i+1)$ and $(x_S, 2i-1)$ belong to $\big\{(x_S, 2i{+}1) \bigm| 0 \leq i < \lfloor N/2 \rfloor\big\}$, the PDOR VLANs provide a minimal-path set in 2-D mesh. □

For example, assuming that the source switch is $(0, 0)$ and the destination switch is $(3, 2)$ in Figure 4, the path is along the following order using VL$(-, 1)$.

$$(0,0) \rightarrow (0,1) \rightarrow (1,1) \rightarrow (2,1) \rightarrow (3,1) \rightarrow (3,2)$$

With this method, the path set is slightly different from that of the dimension-order routing, since it differs from the dimension-order routing only when a source switch is $(x_S, 2i)$. However, the difference is only for the first step toward y-dimension. Therefore, paths of this method are still well-distributed. Its influence will be evaluated in Section 4.

2.4. Generalization (M-dimensional Mesh)

We simply show a generalization of the VLAN-based minimal paths for N^M M-dimensional mesh.

By extending Definition 2.1, we assign an M-dimensional coordinate $(x_0, x_1, \ldots, x_{M-1})$ to each switch, where $0 \leq x_0, x_1, \ldots, x_{M-1} < N$.

Similarly, by simply extending Definition 2.2, $l(x_0, x_1, \ldots, x_{i-1}, -, x_{i+1}, \ldots, x_{M-1})$ is stated as a linear connection which is parallel with i-th axis and has N vertexes $\big\{(x_0, x_1, \ldots, x_i, \ldots, x_{M-1}) \bigm| 0 \leq x_i < N\big\}$.

A VLAN

$$\begin{aligned}
&\mathrm{VL}\big(x_0, x_1, \ldots, x_{i_0-1}, -, x_{i_0+1}, \ldots, x_{M-1} \\
&\qquad\qquad\qquad \bigm| (i_0, i_1, \ldots, i_{M-1})\big) \\
&\big(i_0, i_1, \ldots, i_{M-1} \in \{0, 1, \ldots, M-1\},\ \ i_j \neq i_k\ (j \neq k)\big)
\end{aligned}$$

consists of the following $\left(N^M - 1\right) / (N-1)$ connections (one parallel with i_0-th axis, N connections parallel with i_1-th axis, and so on).

$$\begin{aligned}
&\big\{l(x_0, x_1, \ldots, x_{i_0-1}, -, x_{i_0+1}, \ldots, x_{M-1})\big\} \\
\cup\ &\big\{l(x_0, x_1, \ldots, x_{i_1-1}, -, x_{i_1+1}, \ldots, x_{M-1}) \\
&\qquad\qquad \bigm| 0 \leq x_{i_0} < N\big\} \\
\cup\ &\big\{l(x_0, x_1, \ldots, x_{i_2-1}, -, x_{i_2+1}, \ldots, x_{M-1}) \\
&\qquad\qquad \bigm| 0 \leq x_{i_0}, x_{i_1} < N\big\} \\
&\qquad\qquad \vdots \\
\cup\ &\big\{l(x_0, x_1, \ldots, x_{i_{M-1}-1}, -, x_{i_{M-1}+1}, \ldots, x_{M-1}) \\
&\qquad\qquad \bigm| 0 \leq x_{i_0}, x_{i_1}, \ldots, x_{i_{M-2}} < N\big\}
\end{aligned}$$

Definition 2.6 (DOR VLANs in M-dimensional mesh) *The DOR VLAN set consists of the following N^{M-1} VLANs in N^M M-dimensional mesh:*

$$\begin{aligned}
&\big\{\mathrm{VL}(-, x_1, x_2, \ldots, x_{M-1} \mid A) \\
&\qquad \bigm| 0 \leq x_1, x_2, \ldots, x_{M-1} < N\big\} \\
&\qquad\qquad \big(A = (0, 1, \ldots, M-1)\big)
\end{aligned}$$

□

All paths from a source switch $(x_{0_S}, x_{1_S}, \ldots, x_{(M-1)_S})$ are along the dimension-order routing on a VLAN VL$(-, x_{1_S}, x_{2_S}, \ldots, x_{(M-1)_S} \mid A)$. It is a simple extension of the case in the two-dimensional mesh.

Next, we shift to the second method whose path set is similar to that of the dimension-order routing.

Definition 2.7 (PDOR VLANs in M-dimensional mesh) *The PDOR VLAN set consists of the following $\lfloor N^{M-1}/2 \rfloor{+}1$ VLANs in N^M M-dimensional mesh:*

$$\begin{aligned}
&\big\{\mathrm{VL}(-, x_1, x_2, \ldots, x_{M-1} \mid A) \\
&\qquad \bigm| 0 \leq x_1, x_2, \ldots, x_{M-1} < N, \\
&\qquad\quad \sum_{k=1}^{M-1} x_k \equiv 1 \bmod 2 \Big\} \\
\cup\ &\big\{\mathrm{VL}(x_0, x_1, \ldots, x_{M-2}, - \mid B)\big\} \\
&\qquad\qquad \big(A = (0, 1, \ldots, M-1), \\
&\qquad\qquad\ B = (M-1, M-2, \ldots, 0)\big)
\end{aligned}$$

□

All paths from a source switch $(x_{0_S}, x_{1_S}, \ldots, x_{(M-1)_S})$ to a destination switch $(x_{0_D}, x_{1_D}, \ldots, x_{(M-1)_D})$ are minimal using one of appropriate VLANs selected by the following procedure.

```
if $\sum_{k=1}^{M-1} x_{k_S}$ mod $2 = 1$ then
  use VL$(-, x_{1_S}, x_{2_S}, \ldots, x_{(M-1)_S} \mid A)$ ;
else begin
  $selected := false$ ;
  for $i := 1$ to $M-1$ do
    if $x_{i_D} > x_{i_S}$ then begin
      use VL$(-, x_{1_S}, x_{2_S}, \ldots, x_{(i-1)_S},$
                $x_{i_S}+1, x_{(i+1)_S}, \ldots, x_{(M-1)_S} \mid A)$ ;
      $selected := true$ ; break ;
    end else if $x_{i_D} < x_{i_S}$ then begin
      use VL$(-, x_{1_S}, x_{2_S}, \ldots, x_{(i-1)_S},$
                $x_{i_S}-1, x_{(i+1)_S}, \ldots, x_{(M-1)_S} \mid A)$ ;
      $selected := true$ ; break ;
    end
  end
  if $selected \neq true$ then
    use VL$(x_0, x_1, \ldots, x_{M-2}, - \mid B)$ ;
end
```

Note that the VLAN $\mathrm{VL}(x_0, x_1, \ldots, x_{M-2}, - \mid B)$ is selected only if the following condition is held:

$$\sum_{k=1}^{M-1} x_{k_S} \equiv 0 \bmod 2, \quad x_{i_S} = x_{i_D} \ \ (1 \leq i < M)$$

3. VLAN-based Minimal Paths on Torus

In this section, we show the two VLAN-based methods (DOR and PDOR) on torus. Although the number of VLANs in the PDOR VLAN set is about a half of that in the DOR VLAN set as well as in mesh, it is not the minimum number of VLANs to take minimal paths in torus. However, the method for the minimum number of VLANs causes traffic imbalance, and not treated here.

3.1. Preliminary

In this subsection, we state some notations to represent VLAN topologies in torus. Figure 5 shows 4×4 2-D torus (a) and three examples of its VLAN topologies, (b), (c) and (d). Unlike a mesh, there are wrap-around links in a torus (snipped off in this figure, however these two lines are actually linked).

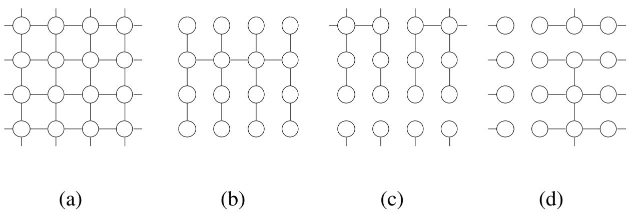


Figure 5. 4×4 **2-D torus and examples of its VLAN topologies**

Definition 3.1 (2-D torus) *Assign a number with a two-dimensional coordinate* (x, y) *where* $0 \leq x < N$ *and* $0 \leq y < N$ *to each switch. By connecting vertex* (x, y) *with four vertexes* $\big((x \pm 1+N) \bmod N,\ y\big)$ *and* $\big(x,\ (y \pm 1+N) \bmod N\big)$*, an* $N \times N$ *two-dimensional torus is formed.* □

Two-dimensional torus treated here is commonly defined as an N-ary 2-cube, and its numbering is the same as that of the mesh.

The VLAN (b) in Figure 5 is the identical with a VLAN $\mathrm{VL}(-, 1)$ used in the previous section. However, the VLAN (c) and (d) cannot be represented by the notation of VLAN topologies for a mesh (see Section 2.1). Thus, we state the notation of VLAN topologies for a torus.

As shown in Figure 5(a), there are N links including a wrap-around link in each dimension, and they form a loop. Therefore, one of these N links must be cut off in a VLAN topology.

Definition 3.2 (linear connection in 2-D torus) *A vertical connection, which consists of* N *vertexes and* $N-1$ *links, is represented as* $l(x_0, - : y_0)$ *in 2-D torus. That is, vertex* (x, y) *is connected with the neighboring two vertexes* $\big(x,\ (y \pm 1+N) \bmod N\big)$*, except for a (cutting off) link between vertexes* (x_0, y_r) *and* $\big(x_0,\ (y_r+1) \bmod N\big)$*, where* $y_r = (y_0 + \lfloor N/2 \rfloor) \bmod N$.

Similarly, a horizontal connection, which consists of N *vertexes and* $N-1$ *links, is represented as* $l(- : x_0, y_0)$ *in 2-D torus. That is, vertex* (x, y) *is connected with the neighboring two vertexes* $\big((x \pm 1+N) \bmod N,\ y\big)$*, except for a link between vertexes* (x_r, y_0) *and* $\big((x_r+1) \bmod N,\ y_0\big)$*, where* $x_r = (x_0 + \lfloor N/2 \rfloor) \bmod N$. □

As shown in Figure 6, the vertex (x_0, y_0) is centered in each horizontal connection.

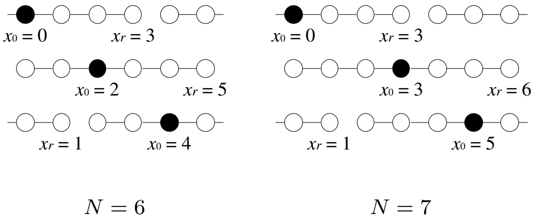


Figure 6. Examples of horizontal connections in 2-D torus

According to Definition 3.2, the VLAN (b) in Figure 5 consists of connections $l(0, - : 1)$, $l(1, - : 1)$, $l(2, - : 1)$, $l(3, - : 1)$ and $l(- : 1, 1)$, while the VLAN (c) consists of connections $l(0, - : 0)$, $l(1, - : 0)$, $l(2, - : 0)$, $l(3, - : 0)$ and $l(- : 3, 0)$, and the VLAN (d) consists of connections $l(2, - : 2)$, $l(- : 2, 0)$, $l(- : 2, 1)$, $l(- : 2, 2)$ and $l(- : 2, 3)$.

Definition 3.3 (VLAN topology in 2-D torus) *Each of the VLAN* $\mathrm{VL}(- : x_0, y_0)$ *and* $\mathrm{VL}(x_0, - : y_0)$ *consists of all switches and the following set of linear connections:*

$$\begin{aligned}
&\mathrm{VL}(- : x_0, y_0) : \\
&\quad \big\{ l(x, - : y_0) \bigm| 0 \leq x < N \big\} \cup \big\{ l(- : x_0, y_0) \big\} \\
&\mathrm{VL}(x_0, - : y_0) : \\
&\quad \big\{ l(- : x_0, y) \bigm| 0 \leq y < N \big\} \cup \big\{ l(x_0, - : y_0) \big\}
\end{aligned}$$

□

According to Definition 3.3, VLANs (b), (c) and (d) in Figure 5 are represented as VL$(- : 1, 1)$, VL$(- : 3, 0)$ and VL$(2, - : 2)$, respectively.

In addition, we use two kinds of term "distance" from a source switch (x_S, y_S) to a destination switch (x_D, y_D) on each x- or y-coordinate as follows.

$$\begin{aligned} d^+(x_S, x_D) &= (x_D - x_S + N) \bmod N \\ d^-(x_S, x_D) &= (x_S - x_D + N) \bmod N \end{aligned}$$

Each of $d^+(x_S, x_D)$ and $d^+(y_S, y_D)$ is the distance on positive direction of x- or y-axis, while each of $d^-(x_S, x_D)$ and $d^-(y_S, y_D)$ is the distance on negative direction. For example, distances from switch $(0, 3)$ to switch $(3, 1)$ on 4×4 torus are as follows:

$$\begin{aligned} d^+(x_S, x_D) = 3, \quad d^-(x_S, x_D) = 1, \\ d^+(y_S, y_D) = 2, \quad d^-(y_S, y_D) = 2 \end{aligned}$$

3.2. Minimal Paths for the DOR in 2-D Torus

In this subsection, we show VLANs based paths along the dimension-order routing in 2-D torus.

Definition 3.4 (DOR VLANs in 2-D torus) *The DOR VLAN set consists of the following* $2N$ *VLANs in 2-D torus:*

$$\left\{ \text{VL}(- : x, y) \;\middle|\; x = a, b, \;\; 0 \le y < N \right\}$$
$$(a = \lfloor (N-1)/2 \rfloor, \;\; b = N-1)$$

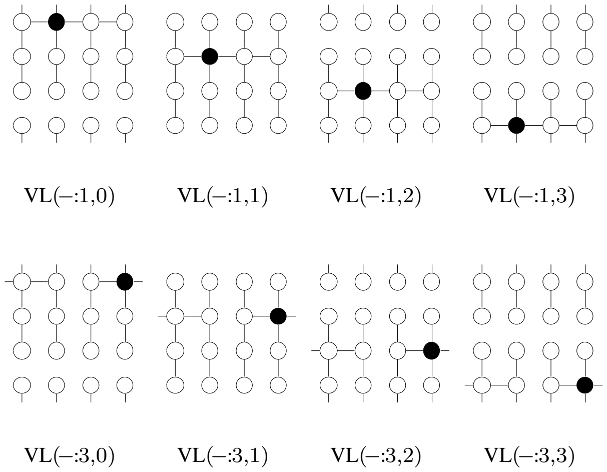
□

This method is similar to that on the mesh, however, a larger number of VLANs is used due to wrap-around links. An example of the DOR VLANs is shown in Figure 7. Figure 7 shows that eight VLANs, VL$(- : 1, 0)$, VL$(- : 1, 1)$, VL$(- : 1, 2)$, VL$(- : 1, 3)$, VL$(- : 3, 0)$, VL$(- : 3, 1)$, VL$(- : 3, 2)$ and VL$(- : 3, 3)$, are employed to take the DOR paths in 4×4 2-D torus.

In this method, paths from a switch (x_S, y_S) take one of two VLANs VL$(- : a, y_S)$ and VL$(- : b, y_S)$. For example, in Figure 7, all paths from the switch $(0, 0)$ use VL$(- : 1, 0)$ for the destination switch $(1, y_D)$ $(y_D \in \{0, 1, 2, 3\})$, while they use VL$(- : 3, 0)$ for the destination switch $(3, y_D)$.

Assuming that a source switch is (x_S, y_S) and a destination switch is (x_D, y_D), the procedure for selecting an appropriate VLAN is described as follows.

```
function $select\_ab\,(s,\ d,\ N : integer) : integer$
begin
  if $d^+(s, d) \le d^-(s, d)$ then begin
    if $s < \lfloor N/2 \rfloor$ then $select\_ab := a$ ;
    else $select\_ab := b$ ;
  end else $\{d^+(s, d) > d^-(s, d)\}$ begin
    if $s < \lfloor N/2 \rfloor$ then $select\_ab := b$ ;
    else $select\_ab := a$ ;
  end
end
use VL$(- : select\_ab\,(x_S, x_D, N),\ y_S)$ ;
```


VL(–:1,0) VL(–:1,1) VL(–:1,2) VL(–:1,3)

VL(–:3,0) VL(–:3,1) VL(–:3,2) VL(–:3,3)

Figure 7. The DOR VLANs in 4×4 2-D torus

For example, if the source switch is $(0, 0)$ and the destination switch is $(3, 2)$ in Figure 7, the path is along the following order using VLAN VL$(- : 3, 0)$. $(b = 3,\ d^+(x_S, x_D) = 3,\ d^-(x_S, x_D) = 1)$

$$(0, 0) \rightarrow (3, 0) \rightarrow (3, 1) \rightarrow (3, 2)$$

Lemma 3.1 *As long as a path is along a single dimension in a torus, it can be formed by two linear connections, each of which is centered on* $a = \lfloor (N-1)/2 \rfloor$ *and* $b = N-1$ *respectively.*

Proof Assume that the path is along x-dimension and two linear connections are $l(- : a, y_0)$ and $l(- : b, y_0)$. $l(- : a, y_0)$ lacks just the wrap-around link between $(N-1, y_0)$ and $(0, y_0)$. Since the maximum distance within x-dimension (the small one of $d^+(x_S, x_D)$ and $d^-(x_S, x_D)$) on the torus is $\lfloor N/2 \rfloor$, the minimal-path set which cannot be covered by $l(- : a, y_0)$ is across the following vertex set.

$$\left\{ (x, y_0) \;\middle|\; x = x_1, x_1{+}1, \ldots, N{-}1,\ 0, 1, \ldots, x_2 \right\}$$
$$\left(x_1 = N - \lfloor N/2 \rfloor, \;\; x_2 = (N-1+\lfloor N/2 \rfloor) \bmod N \right)$$

Note that $x_1 > x_2$ is held. $l(- : b, y_0)$ includes the above vertex set, since $x_r = (b+\lfloor N/2 \rfloor) \bmod N = x_2$. Therefore, the two linear connections $l(- : a, y_0)$ and $l(- : b, y_0)$ are sufficient for minimal paths for all pairs of a source switch and a destination switch among $\left\{ (x, y_0) \;\middle|\; 0 \le x < N \right\}$. □

Theorem 3.1 *The DOR VLANs provide the same minimal-path set as that of the dimension-order routing in 2-D torus.*

Proof Each of VLANs $\mathrm{VL}(-:a, y_S)$ and $\mathrm{VL}(-:b, y_S)$ consists of a horizontal linear connection $l(-:a, y_S)$ or $l(-:b, y_S)$ and N vertical linear connections $\{l(x, y_S) \mid 0 \le x < N\}$. According to Lemma 3.1, a path between each pair of switches among $\{(x, y_S) \mid 0 \le x < N\}$ is minimal along the horizontal linear connection. In each vertical linear connections, all paths from (x, y_S) to (x, y) $(0 \le y < N)$ are minimal, because (x, y_S) is the center of the vertical linear connection. Thus, by selecting an appropriate VLAN of $\mathrm{VL}(-:a, y_S)$ and $\mathrm{VL}(-:b, y_S)$, all paths from a source switch (x_S, y_S) are minimal. Since there are $2N$ VLANs $\{\mathrm{VL}(-:x, y) \mid x = a, b,\ \ 0 \le y < N\}$, paths from a source switch (x_S, y_S) using an appropriate one of $\mathrm{VL}(-:a, y_S)$ and $\mathrm{VL}(-:b, y_S)$ are the same as that of the dimension-order routing. □

3.3. Minimal Paths with Partial DOR (PDOR) in 2-D Torus

The second method for the PDOR VLAN set in the torus is also similar to that in the mesh.

Definition 3.5 (PDOR VLANs in 2-D torus) *The PDOR VLAN set consists of the following* $2\lfloor (N{+}1)/2 \rfloor + 2$ *VLANs in* $N \times N$ *torus:*

$$\begin{aligned} &\{\mathrm{VL}(-:x, 2i) \mid x = a, b,\ \ 0 \le i < \lfloor (N{+}1)/2 \rfloor\} \\ \cup\ &\{\mathrm{VL}(a, -:y_a),\ \mathrm{VL}(b, -:y_b)\} \\ &\qquad\qquad (a = \lfloor (N{-}1)/2 \rfloor,\ \ b = N{-}1) \end{aligned}$$

□

Figure 8 shows an example of the PDOR VLANs in 4×4 2-D torus. This method employs only six VLANs, $\mathrm{VL}(-:1, 0)$, $\mathrm{VL}(-:1, 2)$, $\mathrm{VL}(-:3, 0)$, $\mathrm{VL}(-:3, 2)$, $\mathrm{VL}(1, -:1)$ and $\mathrm{VL}(3, -:3)$.

Compared with the DOR VLANs, $\mathrm{VL}(-:1, 2i{+}1)$ and $\mathrm{VL}(-:3, 2i{+}1)$ $(i = 0, 1, \ldots)$ are deleted, and $\mathrm{VL}(1, -:y_a)$ and $\mathrm{VL}(3, -:y_b)$ are newly employed. The selection of the values y_a and y_b is trivial, because links of the vertical connections in $\mathrm{VL}(1, -:y_a)$ and $\mathrm{VL}(3, -:y_b)$ are never used by paths.

In this method, paths from switch $(x_S, 2i)$ $(0 \le i < \lfloor N/2 \rfloor)$ are along the dimension-order routing for all destination switches using one of two VLANs $\mathrm{VL}(-:a, 2i)$ and $\mathrm{VL}(-:b, 2i)$ in the same way in the first (DOR) method. On the other hand, minimal paths from switch $(x_S, 2i{+}1)$ are achieved by using one of appropriate VLANs, $\mathrm{VL}(-:a, 2i)$, $\mathrm{VL}(-:b, 2i)$, $\mathrm{VL}\big(-:a, (2i{+}2) \bmod N\big)$, $\mathrm{VL}\big(-:b, (2i{+}2) \bmod N\big)$, $\mathrm{VL}(a, -:y_a)$ and $\mathrm{VL}(b, -:y_b)$.

Assuming that a source switch is (x_S, y_S) and a destination switch is (x_D, y_D), the procedure for selecting an appropriate VLAN for this method is described as follows (function *select_ab* was described in Section 3.2).

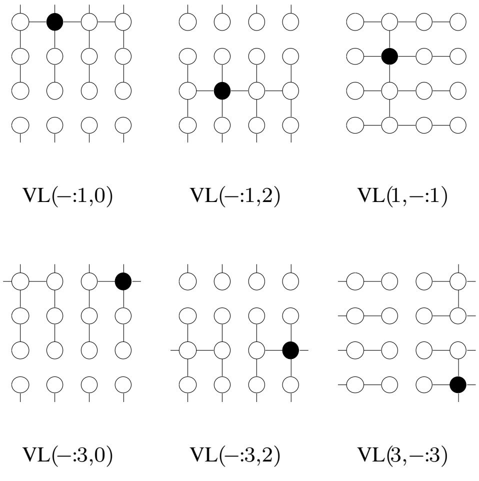


Figure 8. The PDOR VLANs in 4×4 2-D torus

```
ab := select_ab $(x_S, x_D, N)$ ;
if $y_S$ mod $2 = 0$ then use $\mathrm{VL}(-:ab, y_S)$ ;
else begin
  if $y_D = y_S$ then use $\mathrm{VL}(ab, -:y_{ab})$ ;
  else if $d^+(y_S, y_D) \le d^-(y_S, y_D)$ then
    use $\mathrm{VL}(-:ab,\ (y_S{+}1) \bmod N)$ ;
  else $\{d^+(y_S, y_D) > d^-(y_S, y_D)\}$
    use $\mathrm{VL}(-:ab,\ y_S{-}1)$ ;
end
```

Theorem 3.2 *The PDOR VLANs provide a minimal-path set in 2-D torus.*

Proof Since $\mathrm{VL}(a, -:y_a)$ and $\mathrm{VL}(b, -:y_b)$ have N horizontal linear connections $\{l(-:a, y) \mid 0 \le y < N\}$ and $\{l(-:b, y) \mid 0 \le y < N\}$ respectively, paths from a source switch $(x_S, 2i+1)$ to a destination switch $(x, 2i+1)$ $(0 \le x < N)$ on an appropriate one of $\mathrm{VL}(a, -:y_a)$ and $\mathrm{VL}(b, -:y_b)$ are minimal (according to Lemma 3.1). For other destinations, paths from a source switch $(x_S, 2i{+}1)$ are via the switch $\big(x_S,\ (2i{+}2) \bmod N\big)$ with $\mathrm{VL}\big(-:a,\ (2i+2) \bmod N\big)$ or $\mathrm{VL}\big(-:b,\ (2i+2) \bmod N\big)$ for $d^+(y_S, y_D) \le d^-(y_S, y_D)$, or those are via the switch $(x_S, 2i)$ with $\mathrm{VL}(-:a, 2i)$ or $\mathrm{VL}(-:b, 2i)$ for $d^+(y_S, y_D) > d^-(y_S, y_D)$.

On the other hand, according to Theorem 3.1, paths from a source switch $(x_S, 2i)$ using appropriate one of $\mathrm{VL}(-:a, 2i)$ and $\mathrm{VL}(-:b, 2i)$ are minimal. Since both $\big(x_S,\ (2i{+}2) \bmod N\big)$ and $(x_S, 2i)$ belong to $\{(x_S, 2i) \mid 0 \le i < \lfloor (N{+}1)/2 \rfloor\}$, the PDOR VLANs provide minimal-path set in 2-D torus. □

For example, if the source switch is $(0, 3)$ and the destination switch is $(1, 1)$ in Figure 8, the path is along the following order using VLAN VL$(- : 1, 0)$. ($a = 1$, $d^+(x_S, x_D) = 1$, $d^-(x_S, x_D) = 3$, $d^+(y_S, y_D) = 2$, $d^-(y_S, y_D) = 2$)

$$(0,3) \rightarrow (0,0) \rightarrow (1,0) \rightarrow (1,1)$$

With this method, the path set is slightly different from that of the DOR VLANs, since it differs from the dimension-order routing when the source switch is $(x_S, 2i+1)$. However, as well as the PDOR VLANs on a mesh, the difference is possible only for the first step toward y-dimension. Therefore, paths of this method are still well-distributed. Its influence will be evaluated in Section 4.

3.4. Generalization (M-dimensional Torus)

We simply show a generalization of the VLAN-based minimal paths for N^M M-dimensional torus.

By extending Definition 3.1, we assign an M-dimensional coordinate $(x_0, x_1, \ldots, x_{M-1})$ to each switch, where $0 \le x_0, x_1, \ldots, x_{M-1} < N$.

Similarly, by simply extending Definition 3.2, $l(x_0, x_1, \ldots, x_{i-1}, - : x_i, x_{i+1}, \ldots, x_{M-1})$ is stated as a connection which is parallel with i-th axis and is centered on vertex $(x_0, x_1, \ldots, x_{M-1})$.

A VLAN

$$\begin{aligned}\text{VL}\big(&x_0, x_1, \ldots, x_{i_0-1}, - : x_{i_0}, x_{i_0+1}, \ldots, x_{M-1}\\ &\big| \; (i_0, i_1, \ldots, i_{M-1})\big)\\ \big(&i_0, i_1, \ldots, i_{M-1} \in \{0, 1, \ldots, M-1\}, \;\; i_j \ne i_k \; (j \ne k)\big)\end{aligned}$$

consists of following $\left(N^M - 1\right)/(N-1)$ connections (one parallel with i_0-th axis, N connections parallel with i_1-th axis, and so on).

$$\begin{aligned}&\big\{l(x_0, x_1, \ldots, x_{i_0-1}, - : x_{i_0}, x_{i_0+1}, \ldots, x_{M-1})\big\}\\ \cup \;\; &\big\{l(x_0, x_1, \ldots, x_{i_1-1}, - : x_{i_1}, x_{i_1+1}, \ldots, x_{M-1})\\ &\qquad \big| \; 0 \le x_{i_0} < N\big\}\\ \cup \;\; &\big\{l(x_0, x_1, \ldots, x_{i_2-1}, - : x_{i_2}, x_{i_2+1}, \ldots, x_{M-1})\\ &\qquad \big| \; 0 \le x_{i_0}, x_{i_1} < N\big\}\\ &\qquad \vdots\\ \cup \;\; &\big\{l(x_0, x_1, \ldots, x_{i_{M-1}-1}, - : x_{i_{M-1}}, x_{i_{M-1}+1}, \ldots, x_{M-1})\\ &\qquad \big| \; 0 \le x_{i_0}, x_{i_1}, \ldots, x_{i_{M-2}} < N\big\}\end{aligned}$$

Definition 3.6 (DOR VLANs in M-dimensional torus) *The DOR VLAN set consists of the following $2N^{M-1}$ VLANs in N^M M-dimensional torus:*

$$\begin{aligned}\big\{&\text{VL}(- : x_0, x_1, x_2, \ldots, x_{M-1} \mid A)\\ &\big| \; x_0 = a, b, \;\; 0 \le x_1, x_2, \ldots, x_{M-1} < N\big\}\\ &\big(A = (0, 1, \ldots, M-1),\\ &\;\; a = \lfloor (N-1)/2 \rfloor, \;\; b = N-1\big)\end{aligned}$$

□

All paths from a source switch $(x_{0_S}, x_{1_S}, \ldots, x_{(M-1)_S})$ to a destination switch $(x_{0_D}, x_{1_D}, \ldots, x_{(M-1)_D})$ are minimal along the dimension-order routing using (function *select_ab* was described in Section 3.2):

$$\begin{aligned}\text{VL}\big(&- : \textit{select\_ab}\,(x_{0_S}, x_{0_D}, N),\\ &x_{1_S}, x_{2_S}, \ldots, x_{(M-1)_S} \mid A\big)\end{aligned}$$

If N is an odd number, the PDOR VLAN set becomes complicated and is similar to the case of an even number $N+1$ due to wrap-around links. Here we show the PDOR VLAN set only in the case that N is an even number.

Definition 3.7 (PDOR VLANs in M-dimensional torus) *The PDOR VLAN set consists of the following $N^{M-1} + 2$ VLANs in N^M M-dimensional torus:*

$$\begin{aligned}&\big\{\text{VL}(- : x_0, x_1, x_2, \ldots, x_{M-1} \mid A)\\ &\qquad \Big| \; x_0 = a, b, \;\; 0 \le x_1, x_2, \ldots, x_{M-1} < N,\\ &\qquad \left.\sum_{k=1}^{M-1} x_k \equiv 0 \bmod 2\right\}\\ \cup \;\; &\big\{\text{VL}(x_0, x_1, \ldots, x_{M-2}, - : x_{M-1} \mid B) \; \big| \; x_0 = a, b\big\}\\ &\qquad \big(A = (0, 1, \ldots, M-1),\\ &\qquad \;\; B = (M-1, M-2, \ldots, 0),\\ &\qquad \;\; a = \lfloor (N-1)/2 \rfloor, \;\; b = N-1\big)\end{aligned}$$

All paths from a source switch $(x_{0_S}, x_{1_S}, \ldots, x_{M-1_S})$ to a destination switch $(x_{0_D}, x_{1_D}, \ldots, x_{M-1_D})$ are minimal using one of appropriate VLANs selected by the following procedure.

```
$ab := \textit{select\_ab}\,(x_{0_S}, x_{0_D}, N)$ ;
**if** $\sum_{k=1}^{M-1} x_{k_S}$ **mod** $2 = 0$ **then**
  use VL$(- : ab, x_{1_S}, x_{2_S}, \ldots, x_{(M-1)_S} \mid A)$ ;
**else begin**
  $\textit{selected} := \textit{false}$ ;
  **for** $i := 1$ **to** $M-1$ **do**
    **if** $x_{i_S} = x_{i_D}$ **then continue** ;
    **else if** $d^+(x_{i_S}, x_{i_D}) \le d^-(x_{i_S}, x_{i_D})$ **then begin**
      use VL$(- : ab, x_{1_S}, x_{2_S}, \ldots, x_{(i-1)_S},$
                  $(x_{i_S} + 1) \bmod N,$
                  $x_{(i+1)_S}, \ldots, x_{(M-1)_S} \mid A)$ ;
      $\textit{selected} := \textit{true}$ ; **break** ;
    **end else** $\{d^+(x_{i_S}, x_{i_D}) > d^-(x_{i_S}, x_{i_D})\}$ **begin**
      use VL$(- : ab, x_{1_S}, x_{2_S}, \ldots, x_{(i-1)_S},$
                  $(x_{i_S} - 1 + N) \bmod N,$
                  $x_{(i+1)_S}, \ldots, x_{(M-1)_S} \mid A)$ ;
      $\textit{selected} := \textit{true}$ ; **break** ;
    **end**
  **end**
  **if** $\textit{selected} \ne \textit{true}$ **then**
    use VL$(ab, x_1, x_2, \ldots, x_{M-2}, - : x_{M-1} \mid B)$ ;
**end**
```

The VLAN $\mathrm{VL}(ab, x_1, x_2, \ldots, x_{M\text{–}2}, - : x_{M\text{–}1} \mid B)$ is selected only if the following condition is held:

$$\sum_{k=1}^{M\text{–}1} x_{k_S} \equiv 1 \bmod 2, \quad x_{i_S} = x_{i_D} \ \ (1 \leq i < M)$$

4. Performance Evaluation

In this section, we evaluate the performance of the proposed two VLAN-based methods by software simulation.

4.1. Simulation Condition

The following three methods are evaluated: the DOR VLANs based routing (DOR_VB), the PDOR VLANs based routing (PDOR_VB) and the STP-based routing (STP_B). For the comparison, we also evaluate the STP-based method, which uses only one VLAN providing the minimum average path hops among available ones. Thus, its topology is equal to a single spanning tree.

The following six topologies are employed: 4×4 2-D mesh/torus, 8×8 2-D mesh/torus, and $4 \times 4 \times 4$ 3-D mesh/torus. Uniform traffic, in which a host sends a packet to the randomly selected host, is used as a traffic pattern.

We have used a generic flit-level network simulator written in C++. A switch-based Ethernet with point-to-point links is employed. In the network, adjacent switches are connected with just one link each other, and one host is attached to each switch. Every switch uses cut-through as the switching technology, and hosts inject a frame independently of each other.

A simple model consisting of channel buffers, a crossbar, link controllers, a routing table and control circuits is used for the switching fabric. As timing parameters, at least ten clock cycles are required for routing and crossbar set in each switch, and five clock cycles are consumed for link delay (transferring a flit to the next switch or host). We set the frame header size to 6 flits, and payload size to 128 flits. In the simulator, we assume that each flit size is 4 bytes. Thus, total frame size is 536 bytes (134 flits). The simulation time is set to 100,000 clock cycles ignoring the first 10,000 clock cycles.

We use accepted traffic and latency as performance measures. Accepted traffic is the flit reception rate. We define throughput as the maximum accepted traffic. Whereas, latency is the elapsed time in clock cycles after the generation of a frame at a source host until it is delivered at a destination host.

4.2. Simulation Results

Figure 9 shows the latency versus the accepted traffic of three routing methods in each topology.

First, we focus on evaluation results on the mesh (Figure 9(a)(b)(c)). They clearly demonstrate that both of proposed methods improve throughput as compared with the STP-based routing. The improvement is enhanced as the number of dimensions and switches are increased. In particular, Figure 9(c) shows the DOR VLANs based routing improves throughput up to 753% as compared with the STP-based routing. The reason is that the STP-based routing uses only links in a spanning tree, increasing non-minimal paths and traffic concentration due to the non-uniform path distribution. It can be said that the proposed methods for minimal paths are quite efficient to improve network performance. In all conditions, the DOR VLANs based routing achieves higher throughput than the PDOR VLANs based routing up to 75%. The reason is that the PDOR VLANs based routing is different from the DOR VLANs based routing in some paths, leading non-uniform path distribution which increases traffic concentration.

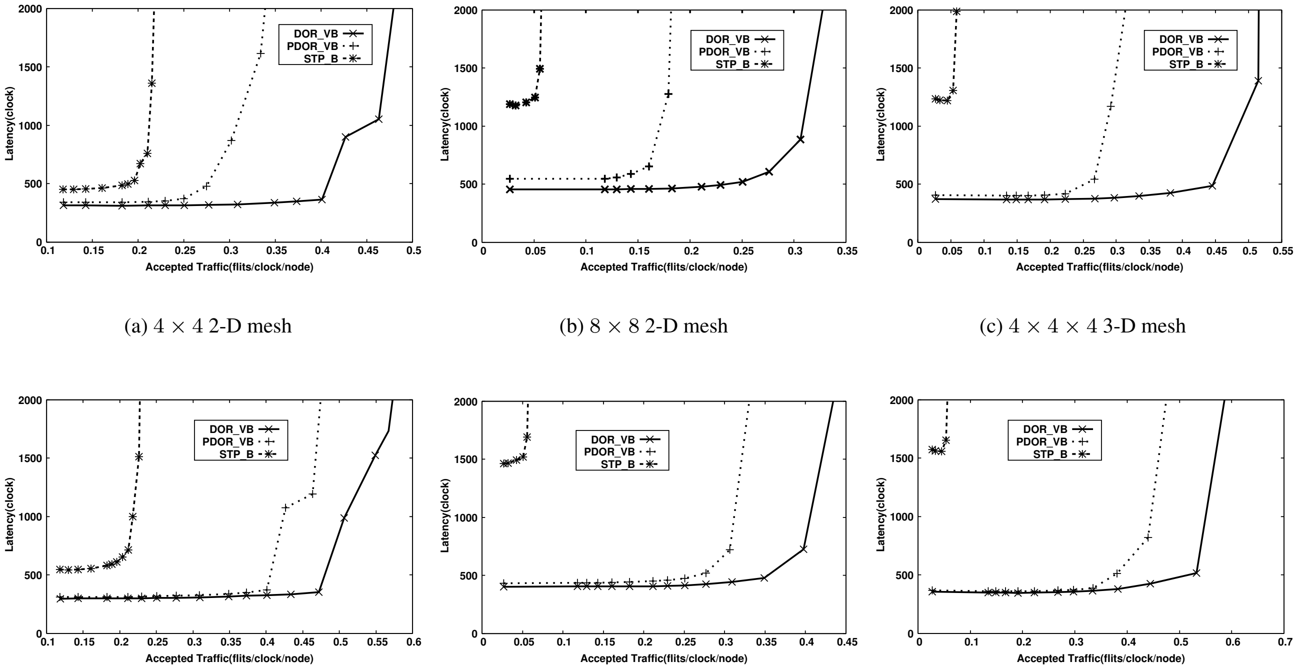
Next, we focus on evaluation results on the torus (Figure 9(d)(e)(f)). As well as results on the mesh, They demonstrate that the proposed methods achieve higher throughput than the STP-based routing, and the DOR VLANs based routing achieves the highest throughput in all conditions. In addition, the improvement on throughput by the proposed methods increases than that in the mesh topologies. In particular, Figure 9(f) shows the DOR VLANs based routing increases throughput up to 902% as compared with the STP-based routing. The reason is explained as follows: since a torus is a symmetric topology due to its wrap-around links, the number of minimal paths of the proposed methods increases, and the paths are distributed more uniformly than that in the mesh. However, the number of minimal paths and the path distribution of the STP-based routing are almost the same in the mesh and the torus, because the STP-based routing can use only links in a spanning tree. As a result, the improvement of the proposed methods in the torus increases as compared with that in the mesh.

On the other hand, the performance gap between the two proposed methods becomes smaller than that in the mesh. The reason is that the difference of the path distribution is smaller than that in the mesh due to the symmetric property of torus.

To sum up, the proposed methods drastically improve throughput as compared with the STP-based routing in all conditions, and the DOR VLANs based routing achieves the highest throughput. The throughput is strongly affected by not only the number of minimal paths but also the uniformity of path distribution.

5. Conclusions

Ethernet has been used to connect hosts in PC clusters by employing system software which enables zero- or one-



(a) 4 × 4 2-D mesh

(b) 8 × 8 2-D mesh

(c) 4 × 4 × 4 3-D mesh

(d) 4 × 4 2-D torus

(e) 8 × 8 2-D torus

(f) 4 × 4 × 4 3-D torus

Figure 9. Accepted traffic and latency under each topology

copy communication. Unlike interconnection networks in parallel computers, links which do not belong to a spanning tree cannot be used for routing due to the limitation of the spanning tree protocol (STP), and simple tree-based topologies have been employed. Thus, even when building a typical topology, such as a torus, clusters with Ethernet must accept non-minimal embedded-tree paths. However, by applying VLAN technology, all links in a cluster with Ethernet can be used to take minimal and/or balanced paths.

In this paper, we proposed VLAN topology sets and path assignment methods to them. The proposed VLAN topology sets on mesh require N^{M-1} and $\lfloor N^{M-1}/2 \rfloor + 1$ VLANs to provide balanced minimal paths and partially balanced ones respectively, where N is the number of switches per dimension and M is the number of dimensions. Similarly, those on torus require $2N^{M-1}$ and $N^{M-1}+2$ VLANs respectively. Simulation results show that the proposed balanced minimal paths improve up to 902% of throughput compared with the STP-based paths, and the proposed partially balanced minimal paths with a slight loss of path uniformity still improve up to 706% of throughput. We are currently planning to evaluate the proposed VLAN-based minimal paths and the STP-based paths on a real PC cluster with 16- or 64-switch Gigabit Ethernet.

References

[1] W. J. Dally and C. L. Seitz. Deadlock-Free Message Routing in Multiprocessor Interconnection Networks. *IEEE Trans. Comput.*, 36(5):547–553, May 1987.

[2] I.T.Association. Infiniband architecture. specification volume 1,release 1.0.a. *available from the InfiniBand Trade Association, http://www.infinibandta.com*, June 2001.

[3] Myricom, Inc. `http://www.pccluster.org/`.

[4] T.Kudoh, H.Tezuka, M.Matsuda, Y.Kodama, O.Tatebe, and S.Sekiguchi. VLAN-based Routing: Multi-path L2 Ethernet Network for HPC Clusters. In *Proceedings of 2004 IEEE International Conference on Cluster Computing (Cluster'2004)*, 2004.

[5] T.Kudoh, M.Matsuda, H.Tezuka, T.Shimizu, Y.Kodama, O.Tatebe, and S.Sekiguchi. VLAN-based Multi-path L2 Ethernet Network for Clusters. *IPSJ Transactions on Advanced Computing System*, 45(SIG 6):35–43, May 2004 (in Japanese).

[6] T.Takahashi, S.Sumimoto, A.Hori, H.Harada, and Y.Ishikawa. PM2: High Performance Communication Middleware for Heterogeneous Network Environment. In *SC2000*, pages 52–53, Nov. 2000.

[7] Y.Ishikawa, H.Tezuka, A.Hori, S.Sumimoto, T. Takahashi, F. O'Carroll, and H. Harada. RWC PC Cluster II and SCore Cluster System Software – High Performance Linux Cluster. In *5th Annual Linux Expo*, pages 55–62, May 1999.