김일성종합대학학보
정보과학

주체108(2019)년
제65권 제4호

Fat-tree구조를 리용한 자료중심망에 대한 연구

리경심, 리일남

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《새로운 과학기술분야를 개척하기 위한 사업도 전망성있게 밀고나가야 합니다.》(《**김정일**선집》 증보판 제11권 138페지)

오늘날 대규모클라스터들에서의 주되는 장애는 마디사이통신대역폭제한이다.

대규모클라스터들을 위한 통신구조에는 두가지 형태가 있다.

우선 InfiniBand[1]나 Myrinet[1]와 같이 하드웨어와 통신규약들을 전문화한 구조이다. 이러한 구조는 비용이 많이 들고 TCP/IP응용프로그람들과 호환이 되지 않으므로 일반적으로 리용하지 않는다.

다음으로 계층형위상구조를 리용한 구조이다. 상품이써네트스위치와 경로기들을 리용하고 TCP/IP응용프로그람, 조작체계, 하드웨어들과 호환가능하므로 이 구조를 많이 리용한다.

계층형위상구조[2]에서는 대역폭문제를 해결하기 위하여 웃준위에 성능이 높은 비싼 10Gbps스위치를 배치하지만 규모가 확대되는데 따라 뿌리계층에서의 병목현상이 해결되지 못하고있다.

따라서 비싼 스위치가 아니라 눅은 일반스위치를 합리적으로 배치하여 적당한 비용으로 대역폭제한을 해결하기 위한 Fat-tree구조[1]가 제안되였다.

그러나 Fat-tree구조를 리용한다고 해도 단일경로조종을 하는 경우 앞붙이개수가 너무 크기때문에 Fat-tree구조의 우점을 리용할수 없게 한다.

론문에서는 Fat-tree위상구조의 우점인 여유경로를 리용할수 있는 합리적인 경로조종방법인 2준위경로조종방법을 제안하였다.

1. 주 소 화

론문에서는 망의 모든 주소를 전용주소 10.0.0.0/8내에서 할당한다.

포드스위치들은 다음과 같은 주소형식을 가진다.

(10.pod.switch.1)

여기서 pod는 포드의 번호([0, $k-1$])를 가리키며 switch는 포드내에서 스위치의 번호([0, $k-1$], 왼쪽에서 오른쪽으로, 우에서 아래로)를 가리킨다.

핵심부스위치에서는 10.k.j.i형식의 주소를 할당하는데 여기서 j와 i는 $(k/2)^2$개의 핵심부스위치의 자리표를 나타낸다.

포드에 련결된 말단호스트의 주소는 자기와 련결된 스위치에 따라 10.pod.switch.ID의 형식으로 설정하는데 ID는 부분망내에서의 위치([2, $k/2+1$], 왼쪽에서 오른쪽으로)를 표현한다.

모든 낮은준위스위치는 $k/2$개의 말단호스트들로 구성된 마스크가 /24인 부분망을 담당한다.

그림 1에 $k=4$인 경우 Fat-tree구조의 주소화형식을 보여주었다.

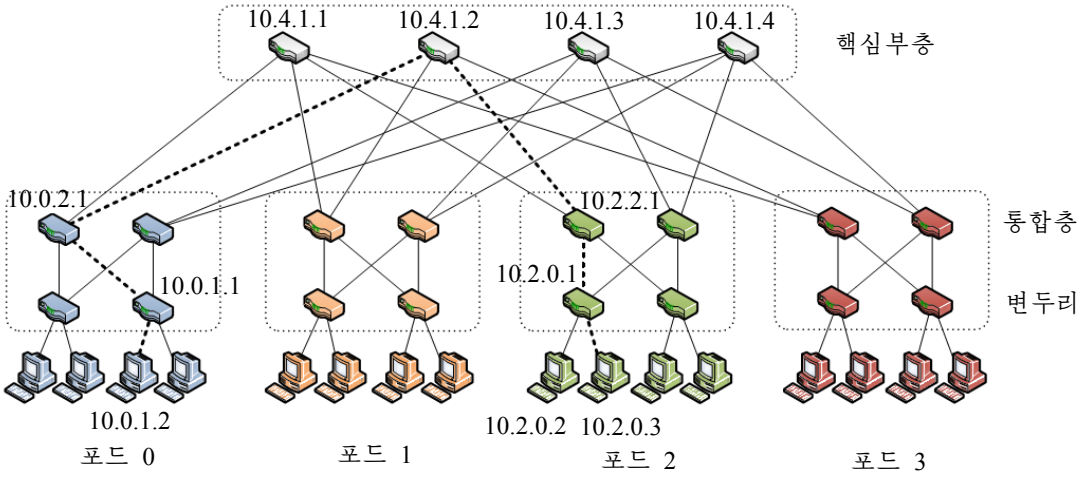

그림 1. $k=4$인 경우 Fat-tree구조의 주소화형식

이 방법은 주소공간을 랑비하지만 경로조종표를 간단하게 한다.

합리적인 경로조종을 위하여 2준위앞붙이표에 기초한 경로조종표를 제안한다.

2준위경로조종표의 첫번째 표에는 두번째 표의 주소를 가리키는 1개의 항목을 추가한다. 첫번째 표에서 앞붙이는 2준위뒤붙이가 없는 경우 즉시 처리되며 두번째 표는 첫번째 표와 련결되게 된다.

2준위경로조종표를 사용하여 10.0.1.2인 원천지에서 10.2.0.3인 목적지까지 파케트들을 전송한다.(그림 2)

Prefix	Output port
10.2.0.0/24	0
10.2.1.0/24	1
0.0.0.0/0	

→

Suffix	Output port
0.0.0.2/8	2
0.0.0.3/8	3

그림 2. 2준위경로조종표실례

이 경로조종표는 10.2.2.1스위치에 해당한 표이다.

2준위경로조종표는 경로기의 비교대기시간을 약간 지연시키지만 하드웨어에서의 병렬앞붙이검색에서는 표의 크기가 매우 작으므로 지연이 크지 않다.

그림 2에서 보는바와 같이 모든 포드스위치들의 경로표에는 앞붙이나 뒤붙이의 개수가 $k/2$보다 작다.

이써네트스위치들에서 내용주소화기억방식(CAM)을 리용하여 2준위경로조종을 수행한다.

TCAM(Ternary Content-Addressable Memory)은 집중탐색에 리용하며 비트패턴에 대한 탐색에 리용되는 접근알고리듬보다 빠르다.

TCAM은 단순박자순환의 모든 항목들을 병렬로 탐색하는 탐색방법을 사용하고있다.(그림 3)

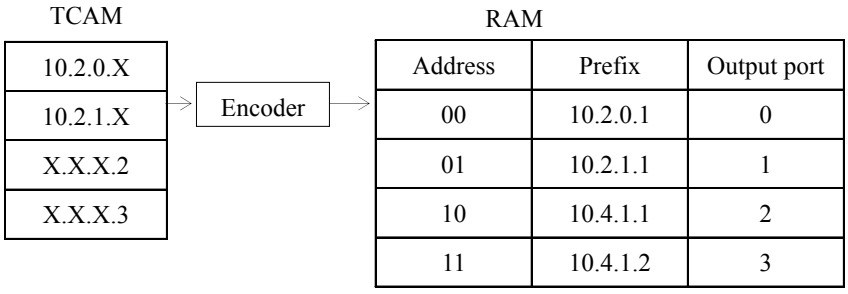

그림 3. 2층경로조종표의 TCAM실현실례

TCAM은 특수한 위치에서 0, 1비교연산을 진행하는 비보호비트들을 보관하는 방법으로 경로표의 뒤붙이들을 보관한다.

TCAM은 용량이 작은 기억장치이지만 매 비트에 대해 대단히 빠르고 비싸다.

TCAM으로 2준위경로조종표를 실현하여 속도를 제고한다.

2. 2준위경로조종표생성알고리듬

Fat-tree에서의 웃준위, 낮은준위스위치는 통신흐름분배기로 동작한다.

모든 포드와 핵심부스위치들에서 전송표들을 생성하기 위한 단계들은 다음과 같다.

① 포드스위치

개별적인 포드스위치에는 같은 포드에 포함된 부분망들에 대한 앞붙이들이 있다.

포드사이통화에 대해서는 호스트식별자를 일치시키는 2준위표에 /0인 마스크를 앞붙이로 추가한다.

포드의 웃준위스위치의 경로조종표알고리듬은 포드내의 웃준위스위치들의 경로조종표를 생성하는 과정을 보여준다.

포드의 웃준위스위치의 경로조종표알고리듬은 다음과 같다.

```
for each pod x in [0, k－1]do
 for each switch z in [(k/2), k－1]do
      for each subnet i in [0, (k/2)－1]do
           addPrefix(10.x.z.1, 10.x.i.0/24, i);
      end
      addPrefix(10.x.z.1, 0.0.0.0/0, 0);
      for each host ID i in[2, (k/2)+1]do
```

```
                addSuffix(10.x.z.1, 0.0.0.i/8, (i－2+z)mod (k/2)+(k/2));
        end
    end
end
```

포드의 웃준위스위치의 경로조종표알고리듬은 같은 호스트식별자를 가진 하나의 호스트에로 전송요구가 있을 때 각이한 낮은준위스위치로부터 통화량이 집중되여 같은 웃준위스위치로 빠져나가는것을 피하도록 한다.

낮은준위포드스위치들에 대하여 볼 때 포드의 웃준위스위치의 경로조종표알고리듬의 3행에서 그 부분망의 통화가 절환되기때문에 마스크가 /24인 부분망의 앞붙이들을 생략할수 있으며 포드내부와 포드사이통화는 웃준위스위치에서 고르롭게 분할되게 된다.

② 핵심부스위치

매 핵심부스위치들이 모든 포드(포구 i가 포드 i에 련결)에 련결되기때문에 핵심부스위치들은 핵심부스위치의 경로조종표알고리듬에서 보여주는바와 같이 목적지포드를 가리키는 마스크가 /16인 앞붙이들을 포함한다. 이 알고리듬은 크기가 k인 표들을 생성한다.

핵심부스위치의 경로조종표알고리듬은 다음과 같다.

```
for each j in [1, (k/2)] do
 for each i in [1, (k/2)] do
        for each destination pod x in[0, k－1]do
              addPrefix(10.k.j.i, 10.x.0.0/16, x);
        end
 end
end
```

3. 분 석

① 흐름일정관리기의 시간과 기억공간요구

표 1에 2.33GHz PC에서 동작하는 흐름일정관리기의 시간과 기억공간요구를 보여주었다.

표 1. 흐름일정관리기의 시간과 기억공간요구

k	호스트개수	평균응답시간/μs	련결상태기억	흐름상태기억
4	16	50.9	64B	4KB
16	1 024	55.3	4KB	205KB
24	3 456	116.8	14KB	691KB
32	8 192	237.6	33KB	1.64MB
48	27 648	754.43	111KB	5.53MB

k가 변할 때 배치요구를 처리하는 평균시간과 련결상태 및 흐름상태자료구조에 필요한 전체 기억을 측정하였다.

27 648개의 호스트망에서 일정관리기는 5.6MB의 기억을 요구하며 0.8ms이내에 흐름을 결정할수 있다.

② 대역폭

표 2에 망의 통합층대역폭을 보여주었다.

대역폭측정을 통하여 알수 있는바와 같이 모든 포드사이통신에서 계층형위상구조를 리용하면 리상적인 대역폭이 28%이다.

표 2. 망의 통합층대역폭

실 험	계층형위상구조/%	Fat-tree의 2준위경로조종표/%
우연통신	53.4	75
stride(1)	100	100
stride(2)	78.1	100
stride(4)	27.9	100
stride(8)	28.0	100
포드사이입력	28.0	50.6
같은 ID출력	27.8	38.5

표 2에서 stride(i)는 식별자가 x인 호스트와 식별자가 (x+i)mod16인 호스트와의 통신이다.

2준위경로조종표를 리용하면 우연통신에서 리상적인 대역폭이 75%이다. 포드사이통신인 경우에 50%로서 통합층에서 같은 출력포구로 전진하므로 결국 전체적인 기대값은 75%로 된다.

맺 는 말

Fat-tree구조를 리용한 자료중심망에서 유효대역폭을 확장하기 위해 2준위경로조종표를 제안하였다.

계층형위상구조를 리용하는 경우 리상적인 대역폭이 28%일 때 론문에서 제안한 방법은 리상적인 대역폭을 75%로 확장하였다.

론문에서 제안한 방법은 현재 존재하는 기술들보다 상당히 적은 비용으로 확장가능한 대역폭을 제공한다.

참 고 문 헌

[1] Gary Lee; Understanding Cloud-based Data Center Networks, Elsevier, 75～180, 2014.

[2] Kailash Jayaswal; Administering Data Centers, Wiley, 340～410, 2006.

주체108(2019)년 8월 5일 원고접수

A Study on Data Center Network Using Fat-tree Topology

Ri Kyong Sim, Ri Il Nam

In this paper two-level routing table was proposed of spraying the traffic to solve the communication bandwidth limitation in Data Center Network using Fat-tree topology.

As the result, using two-level routing table improves the bandwidth, to 75% and reduces for quarter in cost as compared with hierarchical design.

Key words: bandwidth, Data Center Network