Advanced Multicast Switching

Jon Turner
Computer Science & Engineering
Washington University

www.arl.wustl.edu/~jst
Copy-then-Route Multicast

- Input lookup gives fanout and Multicast ID (MI)
- Copy net sends cells to consecutive output ranges
  » distributes load by changing starting point of range
  » unicast cells distributed similarly
  » needs no extra speedup
- Multicast routing tables
  » entry per multicast session
  » output numbers assigned cyclically with wrap-around
  » highly redundant for small fanout sessions
- VOQs before routing net for traffic scheduling
Reducing Routing Table Redundancy

- Every multicast session requires an entry in every routing table - inefficient for small fanout sessions
- Can improve efficiency by routing “small fanout” sessions to restricted subset of routing tables
  - perform load distribution within subsets of size $O(n^{1/2})$
  - reduced redundancy to $O(n^{1/2})$
  - requires speedup of $(1+B)$ due to imperfect distribution of small fanout sessions
- Adding and removing endpoints
  - $n$ setup cells for large fanout connections, $n^{1/2}$ for small
  - hardware update support can reduce to $O(F)$
- Can extend to >2 fanout classes (memory vs. bw)
Two-Pass Copy-then-Route

- Unicast cells use one pass
- Multicast cells use two passes
  - rotate starting point to balance load
  - cyclic assignment of output numbers with wrap-around
  - resequencing done cumulatively across both passes
  - speedup of $1+\delta$ where $\delta$ is fraction of outgoing traffic in multicast sessions

- Can use same methods to reduce redundancy
  - with two fanout classes $O(n^{1/2})$ redundancy, speedup of $1+2\delta+B$
  - with $r$ classes $O(n^{1/r})$ redundancy, speedup of $1+r\delta+(r-1)B$

- Two stage traffic regulation for multicast
  - control total multicast traffic volume in first pass
System Level Traffic Regulation

- Unicast uses rate-controlled VOQs at all inputs
- Two stage regulation of multicast
  - first pass regulates total multicast traffic volume to avoid congestion
    - use per multicast flow queues at inputs to prevent hogging
  - separate set of multicast VOQs for second pass
    - objective is to minimize delay variation for multicast traffic
    - essential for reliable resequencing
    - if multiple fanout classes used, need separate set of multicast VOQs for each class
- Can provide strong performance guarantees if total multicast traffic is not excessive
  - can isolate “compliant flows” from greedy flows
Multicast Static Routing Networks

- Routing tables in each SE specify branching
- Multicast Index (MI) supplied by input port
- MI used to lookup bit vector specifying branching
  » may use direct lookup or hash
Nonblocking Multicast

■ A session is triple \((x, Y, \omega)\) where \(Y\) is an output set
  » route is subtree of network connecting \(x\) to all of \(Y\)

■ Compatibility
  » set of sessions is compatible if total weight per input/output is \(\leq 1\)
  » set of routes is compatible if total weight per internal link is \(\leq S\)

■ A state of a network is a set of compatible sessions and corresponding compatible routes

■ Nonblocking properties
  » strict – in any state, can add new session or extend existing one
  » wide-sense – can add new session or extend session in all states that can be reached by a specific routing algorithm
  » reroutable – can add new session in all reachable states
  » rearrangeable – can route any set of compatible sessions
Nonblocking Condition for Clos

**Theorem.** For multicast sessions $C^3_{n,d,r}$ is wide-sense nonblocking if

$$
 r > \left\lfloor \frac{F \cdot \frac{d-B}{S-B} + \frac{d-B}{S-B}}{b} \right\rfloor + \left\lfloor \frac{d-b}{b} \right\rfloor \quad \text{if } b > S-B
$$

where $F = \min\{r, n/d\}$.

**Proof sketch.** If routing algorithm uses greedy branching in third stage, number of middle stage switches not reachable from an input is $\leq \left\lfloor F(d-B)/(S-B) \right\rfloor - 1$ and number not reachable from an output is $\leq (d-B)/(S-B) - 1$.

If $d-n^{1/2}$ and $S$ is a constant (not dependent on $n$), need $O(n)$ middle stage switches – impractically expensive.

Can derive similar condition for Benes – also not practical.
Cascaded Benes Networks

- Limit fanout to $f$ in each Benes network
- If $f \leq d$, and there are $\log_f n$ subnets, no blocking when

$$S \geq (f + 1)(1 - 1/d)(k - 1) + (f + 1)/d + B$$

when $f=4$, cost is about $(5/4)\log_2 n$ times the cost of nonblocking unicast network (so $10\times$ for $n=256$)
- Can merge boundary stages to reduce cost
Multipass Variant

- Binary branching in each pass
- Re-route to modify session
  - extend session by inserting additional binary branch point
  - retract session by removing binary branch point
  - time required comparable to unicast
- With Benes network, no blocking if

\[ S \geq 2(1 + \delta)((1 - 1/d)(k - 1) + 1/d) + (1 - (1 + \delta)/d)B \]

where \( \delta \) is multicast traffic fraction
- Only method for which network cost and time per operation is roughly comparable with unicast
Pippenger Network

- Recursive construction
  - branch as needed in first stage
  - nonblocking unicast switches act as concentrators
  - repeat in smaller subnets
- Using Benes networks for concentrators need
  \[ S \geq 2(1 - 1/d)(k - 1) + (1 - 1/d)B + 2/d \]
- Cost is roughly \( \log_2 n \) times cost of unicast net
- Can also use to construct rearrangeable network
Making Clos Reroutably Nonblocking

- Limit branching in first stage to $f \leq \Gamma$

- **Theorem.** $C^3_{n,d,r}$ is reroutably nonblocking if
  $$r > \left\lceil \frac{\Gamma(d-B)}{(S-B)-1} \right\rceil + \left\lceil \frac{(d-B)}{(S-B)-1} \right\rceil (n/d)^{1/f}$$
  if $b=0$ and first stage fanout is limited to $f$.

  Can choose $f$ to minimize $r$.

  Key to proof is a lemma concerning the set covering problem:
  - Given a set $A = \{a_1, \ldots, a_t\}$, and a collection $S = \{S_1, \ldots, S_n\}$
    where each $S_i$ is a subset of $A$, find the smallest possible number of sets $S_i$ whose union equals $A$.

  - The **greedy algorithm** for set covering selects sets one at a time, always picking the next set that covers the most previously uncovered elements.
Set Covering Lemma

Lemma. Let $A = \{a_1, \ldots, a_t\}$ and $S = \{S_1, \ldots, S_p\}$ be an instance of set covering in which every $a_i$ appears in $\geq p-q$ sets for some $q$. If $p > qt^{1/2}$ then the greedy solution uses $\leq t$ sets.

Proof. Let $h$ be the number of sets in the greedy solution and assume sets are numbered so that for $1 \leq i \leq h$, $S_i$ is the set chosen in step $i$. Define $U_i = S_1 \cup \ldots \cup S_i$ and let $D_i = S_i - U_{i-1}$, $s_i = |S_i|$, $u_i = |U_i|$, $d_i = |D_i|$. Then,

$$ps_i \geq \sum_{j=1}^p s_j \geq (p-q)t$$

so $u_i \geq (1-q/p)t = (1-x_0)t$

where $x_j = (q-j)/(p-j)$. Next, note that

$$(p-1)d_i \geq \sum_{j=1}^p D_i \geq (p-q)(l-u_i) \quad \text{so} \quad d_i \geq (1-x_0)(l-u_i)$$

$$u_i = u_{i-1} + d_i \geq (1-x_0)l + x_0u_{i-1} \geq (1-x_0)t + x_0(1-x_0)t = (1-x_0)t$$

Similarly, we find that for $i \leq h$, $u_i \geq (1-x_{i-1} \ldots x_1x_0)t$. In particular

$$u_i \geq (1-x_{i-1} \ldots x_1x_0)t \geq (1-x_0)^{i}t \geq (1-(1/t))^i - t - 1$$

So, $U_i$ has $>t-1$ elements and since $|A| = t$, $|U_i| = t$. 


Proof of Nonblocking Theorem

To prove theorem, show that one can setup a route of weight \( B \) from an input \( x \) to some arbitrary set of outputs.

Let \( p \) be the number of second stage switches reachable from \( x \) and note that \( p \geq r - f(d-B)/(S-B) - 1 > (d-1)/(S-B) - 1 (n/d)^{1/r} \), using the bound on \( r \) in the statement of the theorem.

Note that each third stage switch can be blocked from reaching at most \( (d-B)/(S-B) - 1 \) second stage switches.

To prove the theorem, apply the lemma by letting \( A \) be the set of switches in the third stage (so \( t = n/d \)). Define \( S_j \) to be the set of third stage switches that can be reached by the \( j \)-th second stage switch that is reachable from \( x \).

Observe that each third stage switch appears in at least \( p - [(d-B)/(S-B) - 1] \) of the \( S_j \), so let \( q = [(d-B)/(S-B) - 1] \).

Now, we see that \( qt^{1/r} = [(d-B)/(S-B) - 1] (n/d)^{1/r} \) and since \( p \) is larger than this, the lemma tells us that we can reach all the third stage switches through just \( f \) of the second stage switches.
Other Reroutable Networks

- Cascaded pair of Bcncs nets is reroutable if
  \[ S \geq 2(1 - 1/d)(k - 1) + (1 - 1/d)B + 2/d \]
  and branching is allowed only in second network
  - new sessions are routed through most lightly loaded part of second network
- can merge stages at boundary
  - if \( k = 2 \), this yields five stage network with speedup \( = 2 + B \)
  - can eliminate one stage at the expense of higher speedup
    \[ S \geq (3/2) + (1-1/d)B + ((5/4)-(B/d))^{1/2} \approx 2.62 + B \]
- Two pass variant requires speedup
  \[ S \geq 2(1 + \delta)(1 - 1/d)(k - 1) + B + 2(1 + \delta)/d \]
  where \( \delta \) is multicast fraction