(How to Implement) Basic Communication Operations

Alexandre David
B2-206

Overview

- One-to-all broadcast & all-to-one reduction (4.1).
- All-to-all broadcast and reduction (4.2).
- All-reduce and prefix-sum operations (4.3).
- Scatter and Gather (4.4).
- All-to-All Personalized Communication (4.5).
- Circular Shift (4.6).
- Improving the Speed of Some Communication Operations (4.7).

Collective Communication Operations

- Represent regular communication patterns.
- Used extensively in most data-parallel algorithms.
- Critical for efficiency.
- Available in most parallel libraries.
- Very useful to "get started" in parallel processing.

Collective: involve group of processors.

The efficiency of data-parallel algorithms depends on the efficient implementation of these operations.

Recall: t_s+mt_w time for exchanging a *m*-word message with cut-through routing.

All processes participate in a single **global** interaction operation or subsets of processes in **local** interactions.

Goal of this chapter: good algorithms to implement commonly used communication patterns.

Reminder

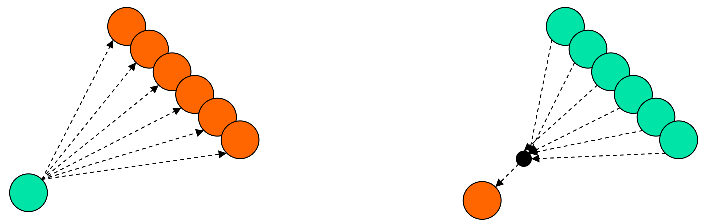
- Result from previous analysis:
  - Data transfer time is roughly the same between *all pairs* of nodes.
  - Homogeneity true on modern hardware (randomized routing, cut-through routing…).
    - t_s+mt_w
    - Adjust t_w for congestion: effective t_w.
- Model: bidirectional links, single port.
- Communication with point-to-point primitives.

Broadcast/Reduction

- One-to-all broadcast:
  - Single process sends identical data to all (or subset of) processes.
- All-to-one reduction:
  - Dual operation.
  - *P* processes have *m* words to send to one destination.
  - Parts of the message need to be *combined*.

Reduction can be used to find the sum, product, maximum, or minimum of sets of nmbers.

Broadcast/Reduction



Broadcast Reduce

This is the logical view, what happens from the programmer's perspective.

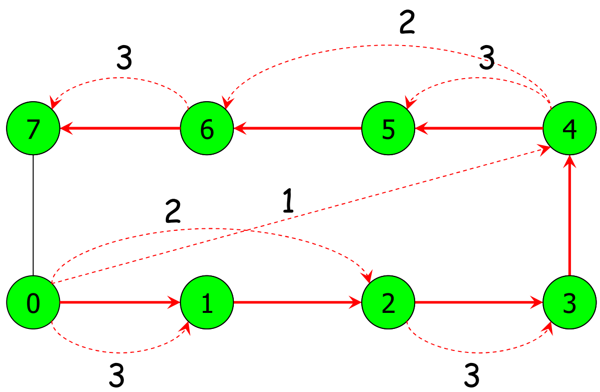
One-to-All Broadcast – Ring/Linear Array

- Naïve approach: send sequentially.
  - Bottleneck.
  - Poor utilization of the network.
- Recursive doubling:
  - Broadcast in $\log p$ steps (instead of p).
  - Divide-and-conquer type of algorithm.
  - Reduction is similar.

Source *process* is the bottleneck. Poor utilization: Only connections between single pairs of nodes are used at a time.

Recursive doubling: All processes that have the data can send it again.

Recursive Doubling



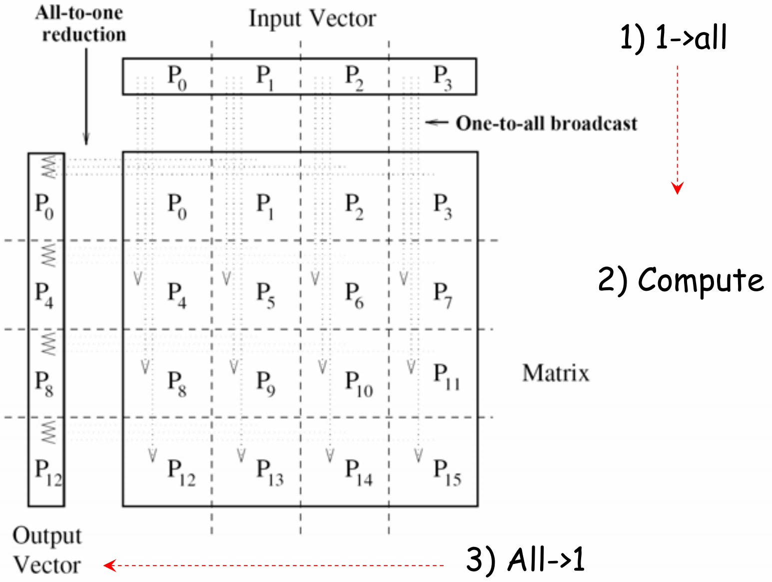
Note:

•The nodes do not snoop the messages going “through” them. Messages are forwarded but the processes are not notified of this because they are not destined to them.

•Choose carefully destinations: furthest.

•Reduction symmetric: Accumulate results and send with the same pattern.

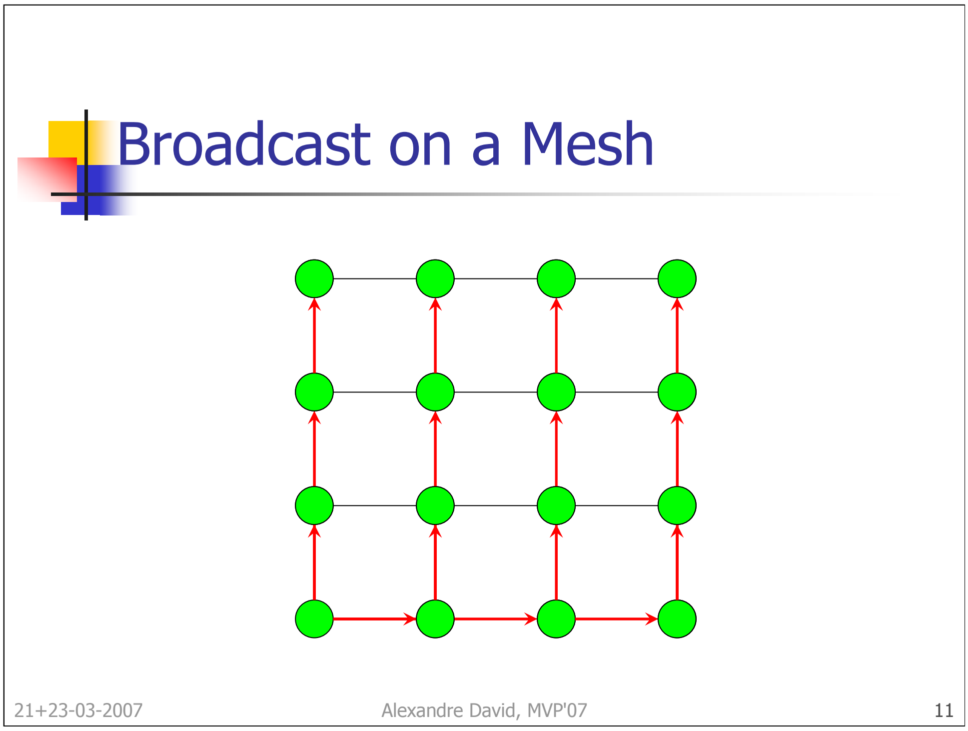
Example: Matrix*Vector



Although we have a matrix & a vector the broadcast are done on arrays.

One-to-All Broadcast – Mesh

- Extensions of the linear array algorithm.
  - Rows & columns = arrays.
  - Broadcast on a row, broadcast on columns.
  - Similar for reductions.
  - Generalize for higher dimensions (cubes…).

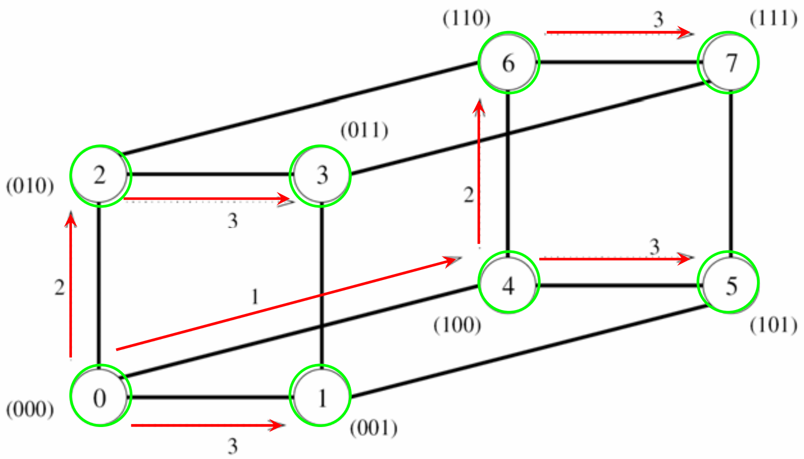


1. Broadcast like linear array.
2. Every node on the linear array has the data and broadcast on the columns with the linear array algorithm, *in parallel*.

One-to-All Broadcast – Hypercube

- Hypercube with 2^d nodes = d-dimensional mesh with 2 nodes in each direction.
- Similar algorithm in d steps.
- Also in $\log p$ steps.
- Reduction follows the same pattern.

Broadcast on a Hypercube



Better for congestion: Use different links every time. Forwarding in parallel again.

All-to-One Broadcast – Balanced Binary Tree

- Processing nodes = leaves.
- Hypercube algorithm maps well.
- Similarly good w.r.t. congestion.

Broadcast on a Balanced Binary Tree

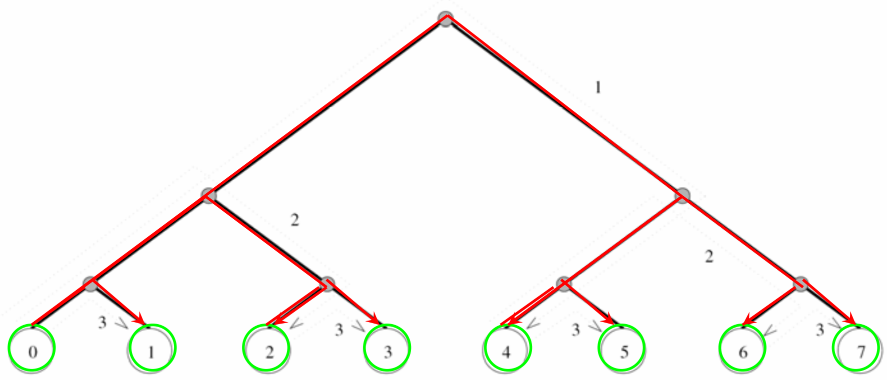


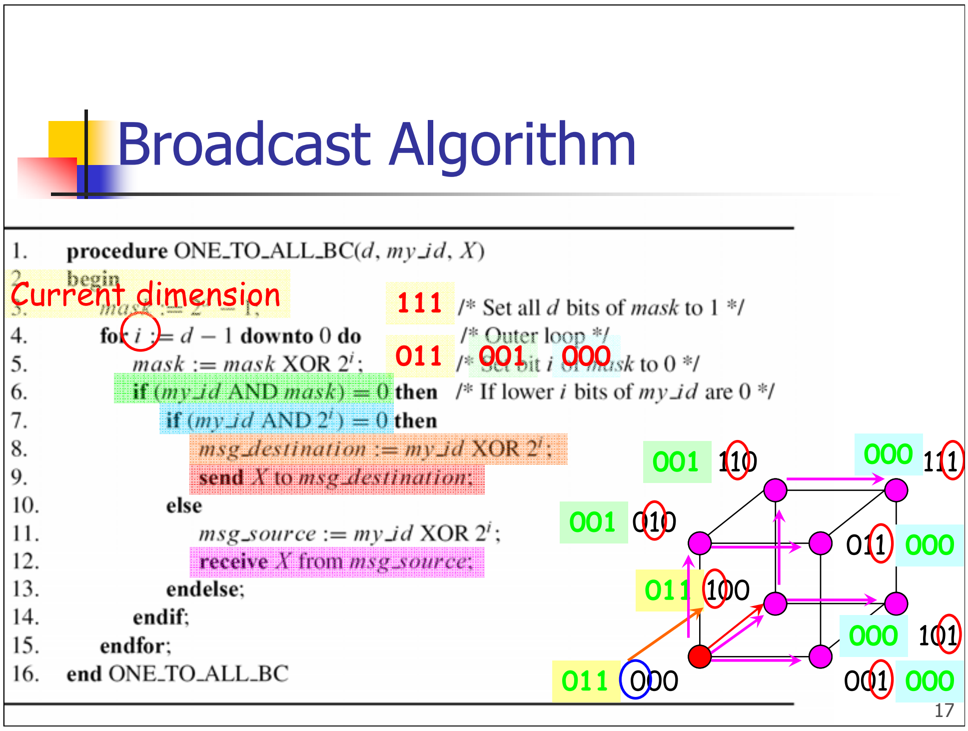
Figure 4.7 One-to-all broadcast on an eight-node tree.

Divide-and-conquer type of algorithm again.

Algorithms

- So far we saw pictures.
- Not enough to implement.
- Precise description
  - to implement.
  - to analyze.
- Description for hypercube.
- Execute the following procedure on all the nodes.

For sake of simplicity, the number of nodes is a power of 2.

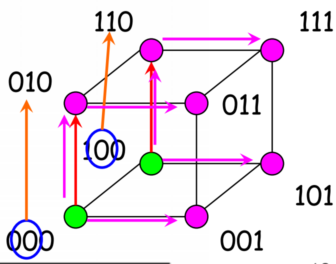


my_id is the label of the node the procedure is executed on. The procedure performs **d** communication steps, one along each dimension of the hypercube.

Nodes with zero in **i** least significant bits (of their labels) participate in the communication.

Broadcast Algorithm

```
procedure ONE_TO_ALL_BC(d, my_id, X)
begin
   mask := 2^d - 1;                       /* Set all d bits of mask to 1 */
   for i := d - 1 downto 0 do             /* Outer loop */
      mask := mask XOR 2^i;               /* Set bit i of mask to 0 */   001  000
      if (my_id AND mask) = 0 then        /* If lower i bits of my_id are 0 */
         if (my_id AND 2^i) = 0 then
            msg_destination := my_id XOR 2^i;
            send X to msg_destination;
         else
            msg_source := my_id XOR 2^i;
            receive X from msg_source;
         endelse;
      endif;
   endfor;
end ONE_TO_ALL_BC
```



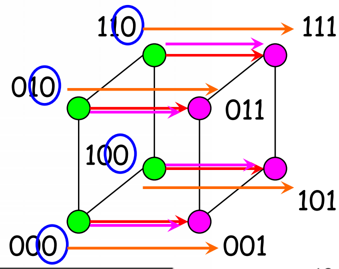
18

my_id is the label of the node the procedure is executed on. The procedure performs **d** communication steps, one along each dimension of the hypercube.

Nodes with zero in **i** least significant bits (of their labels) participate in the communication.

Broadcast Algorithm

```
procedure ONE_TO_ALL_BC(d, my_id, X)
begin
    mask := 2^d - 1;                          /* Set all d bits of mask to 1 */
    for i := d - 1 downto 0 do                /* Outer loop */
        mask := mask XOR 2^i;                 /* Set bit i of mask to 0 */
        if (my_id AND mask) = 0 then          /* If lower i bits of my_id are 0 */
            if (my_id AND 2^i) = 0 then
                msg_destination := my_id XOR 2^i;
                send X to msg_destination;
            else
                msg_source := my_id XOR 2^i;
                receive X from msg_source;
            endelse;
        endif;
    endfor;
end ONE_TO_ALL_BC
```



19

my_id is the label of the node the procedure is executed on. The procedure performs **d** communication steps, one along each dimension of the hypercube.

Nodes with zero in **i** least significant bits (of their labels) participate in the communication.

Notes:

- Every node has to know when to communicate, i.e., call the procedure.
- The procedure is distributed and requires only point-to-point synchronization.
- Only from node 0.

Algorithm For Any Source

```
procedure GENERAL_ONE_TO_ALL_BC(d, my_id, source, X)
begin
   my_virtual_id := my_id XOR source;
   mask := 2^d - 1;
   for i := d - 1 downto 0 do      /* Outer loop */
      mask := mask XOR 2^i;   /* Set bit i of mask to 0 */
      if (my_virtual_id AND mask) = 0 then
         if (my_virtual_id AND 2^i) = 0 then
            virtual_dest := my_virtual_id XOR 2^i;
            send X to (virtual_dest XOR source);
   /* Convert virtual_dest to the label of the physical destination */
         else
            virtual_source := my_virtual_id XOR 2^i;
            receive X from (virtual_source XOR source);
   /* Convert virtual_source to the label of the physical source */
         endelse;
   endfor;
end GENERAL_ONE_TO_ALL_BC
```

20

XOR the source = renaming relative to the source. Still works because of the sub-cube property: changing 1 bit = navigate on one dimension, keep a set of equal bits = sub-cube.

Reduce Algorithm

```
procedure ALL_TO_ONE_REDUCE(d, my_id, m, X, sum)
begin
    for j := 0 to m − 1 do sum[j] := X[j];
    mask := 0;
    for i := 0 to d − 1 do
        /* Select nodes whose lower i bits are 0 */
        if (my_id AND mask) = 0 then
            if (my_id AND 2^i) ≠ 0 then
                msg_destination := my_id XOR 2^i;
                receive X from msg_source;
                for j := 0 to m − 1 do
                    sum[j] :=sum[j] + X[j];
            endelse;
        mask := mask XOR 2^i;    /* Set bit i of mask to 1 */
    endfor;
end ALL_TO_ONE_REDUCE
```

In a nutshell:
reverse the previous one.

Cost Analysis

p processes $\rightarrow$ log$p$ steps (point-to-point transfers in parallel).
Each transfer has a time cost of $t_s+t_w m$.
Total time: $T=(t_s+t_w m)$ log$p$.

All-to-All Broadcast and Reduction

- Generalization of broadcast:
  - Each processor is a source and destination.
  - Several processes broadcast different messages.
- Used in matrix multiplication (and matrix-vector multiplication).
- Dual: all-to-all reduction.

How to do it?

If performed naively, it may take up to *p* times as long as a one-to-all broadcast (for *p* processors).

Possible to concatenate all messages that are going through the same path (reduce time because fewer t_s).

All-to-All Broadcast and Reduction

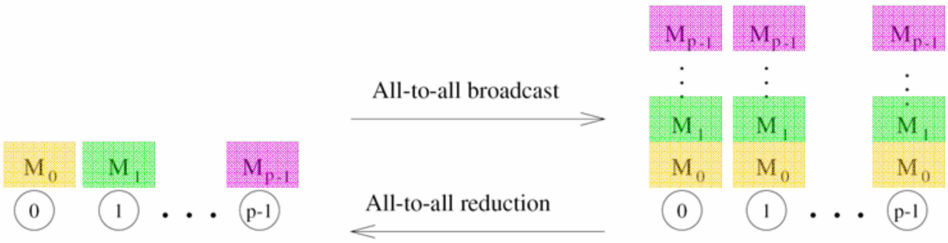
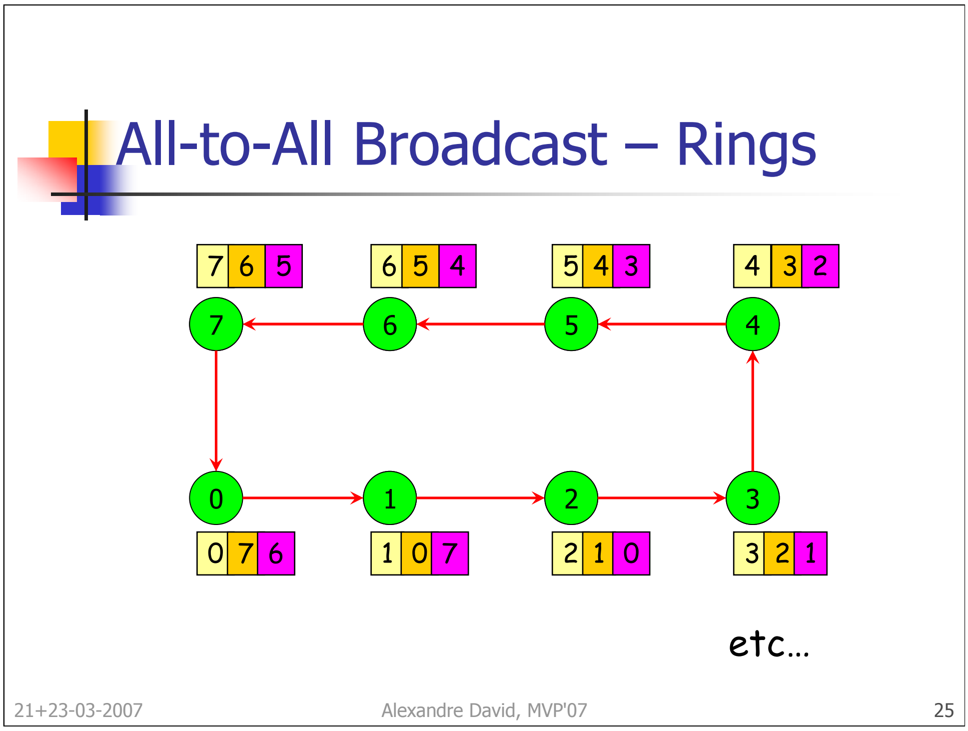


Figure 4.8 All-to-all broadcast and all-to-all reduction.



All communication links can be kept busy until the operation is complete because each node has some information to pass. One-to-all in log*p* steps, all-to-all in *p-1* steps instead of *p* log*p* (naïve).

How to do it for linear arrays? If we have bidirectional links (assumption from the beginning), we can use the same procedure.

All-to-All Broadcast Algorithm

```
procedure ALL_TO_ALL_BC_RING(my_id, my_msg, p, result)
begin
    left := (my_id − 1) mod p;          Ring: mod p.
    right := (my_id + 1) mod p;         Receive & send – point-to-point.
    result := my_msg;
    msg := result;                      Initialize the loop.
    for i := 1 to p − 1 do
        send msg to right;
        receive msg from left;          Forward msg.
        result := result ∪ msg;         Accumulate result.
    endfor;
end ALL_TO_ALL_BC_RING
```

Algorithm 4.4 All-to-all broadcast on a p-node ring.

All-to-All Reduce Algorithm

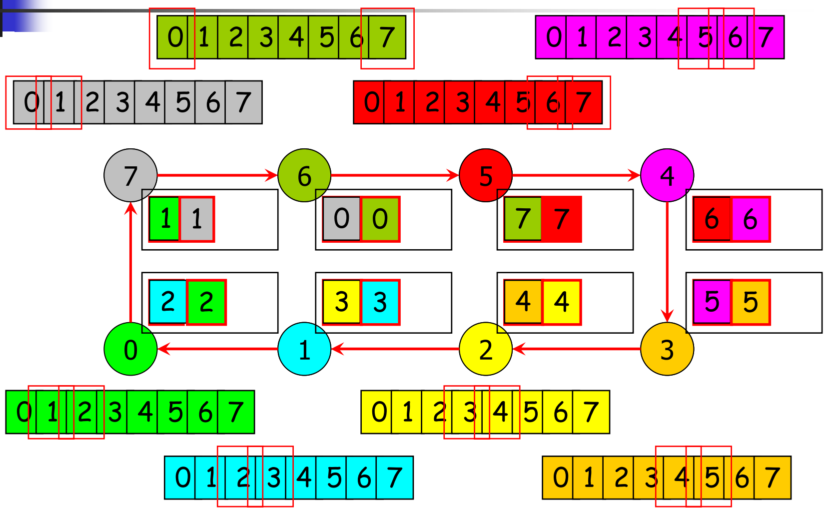
```
procedure ALL_TO_ALL_RED_RING(my_id, my_msg, p, result)
begin
    left := (my_id − 1) mod p;
    right := (my_id + 1) mod p;
    recv := 0;
    for i := 1 to p − 1 do
        j := (my_id + i) mod p;
        temp := msg[j] + recv;
        send temp to left;
        receive recv from right;
    endfor;
    result := msg[my_id] + recv;
end ALL_TO_ALL_RED_RING
```

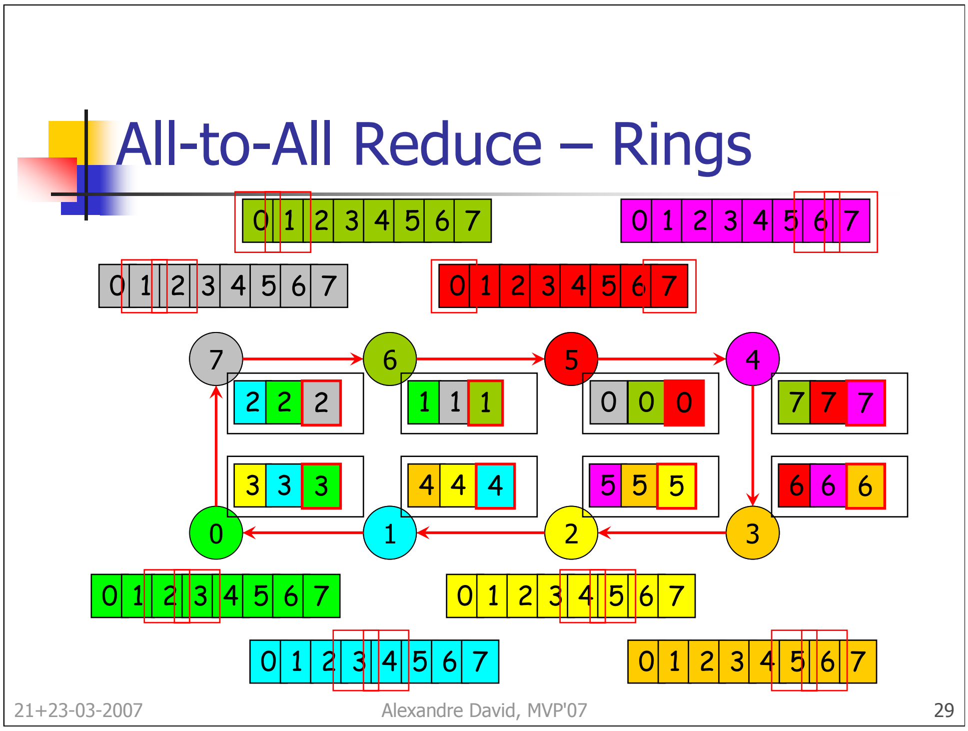
Accumulate and forward.

Last message for *my_id*.

Algorithm 4.5 All-to-all reduction on a p-node ring.

All-to-All Reduce – Rings



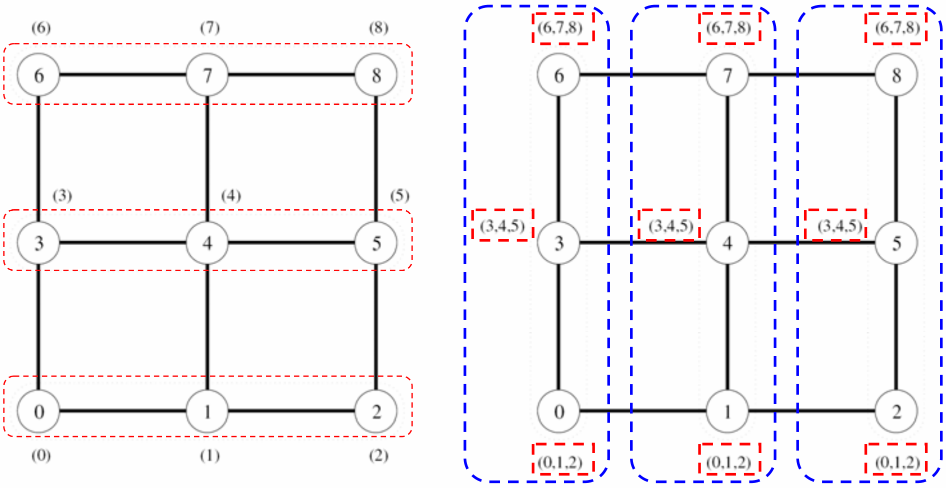

All-to-All Reduce – Rings
21+23-03-2007
Alexandre David, MVP'07
29

p-1 steps.

All-to-All Broadcast – Meshes

- Two phases:
  - All-to-all on rows – messages size m.
    - Collect sqrt(p) messages.
  - All-to-all on columns – messages size sqrt(p)*m.

All-to-All Broadcast – Meshes



Algorithm

procedure ALL_TO_ALL_BC_MESH(my_id, my_msg, p, $result$)
begin

/* Communication along rows */
    $left := my\_id - (my\_id \bmod \sqrt{p}) + (my\_id - 1) \bmod \sqrt{p}$;
    $right := my\_id - (my\_id \bmod \sqrt{p}) + (my\_id + 1) \bmod \sqrt{p}$;
    $result := my\_msg$;
    $msg := result$;
    for $i := 1$ **to** $\sqrt{p} - 1$ **do**
        send msg to $right$;
        receive msg from $left$;
        $result := result \cup msg$;
    endfor;

/* Communication along columns */
    $up := (my\_id - \sqrt{p}) \bmod p$;
    $down := (my\_id + \sqrt{p}) \bmod p$;
    $msg := result$;
    for $i := 1$ **to** $\sqrt{p} - 1$ **do**
        send msg to $down$;
        receive msg from up;
        $result := result \cup msg$;
    endfor;
end ALL_TO_ALL_BC_MESH

All-to-All Broadcast - Hypercubes

- Generalization of the mesh algorithm to log*p* dimensions.
- Message size doubles at every step.
- Number of steps: log*p*.

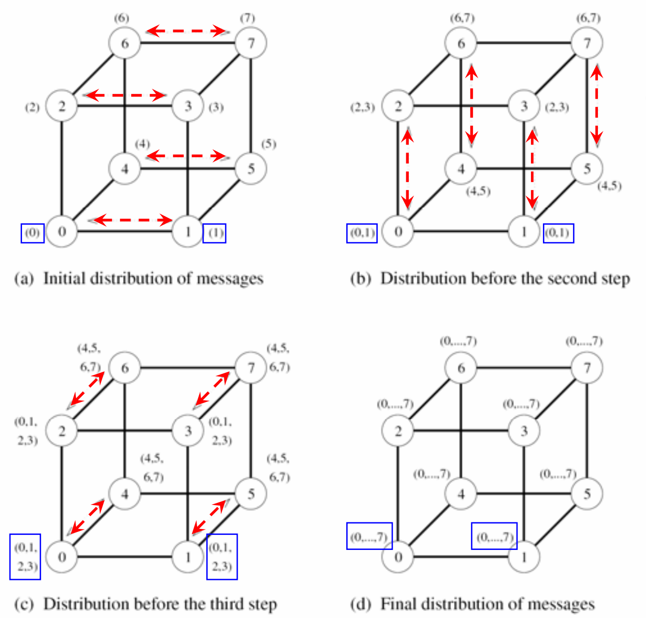
Remember the 2 extremes:

•Linear array: p nodes per (1) dimension – p^1.

•Hypercubes: 2 nodes per log*p* dimensions – $2^{\log p}$.

And in between 2-D mesh sqrt(p) nodes per (2) dimensions – $\text{sqrt}(p)^2$.

All-to-All Broadcast – Hypercubes



Algorithm

```
procedure ALL_TO_ALL_BC_HCUBE(my_id, my_msg, d, result)
begin
    result := my_msg;
    for i := 0 to d − 1 do
        partner := my_id XOR 2^i;
        send result to partner;
        receive msg from partner;
        result := result ∪ msg;
    endfor;
end ALL_TO_ALL_BC_HCUBE
```

Loop on the dimensions

Exchange messages

Forward (double size)

Algorithm 4.7 All-to-all broadcast on a d-dimensional hypercube.

At every step we have a broadcast on sub-cubes. The size of the sub-cubes doubles at every step and all the nodes exchange their messages.

All-to-All Reduction – Hypercubes

```
procedure ALL_TO_ALL_RED_HCUBE(my_id, msg, d, result)
begin
    recloc := 0;
    for i := d − 1 to 0 do
        partner := my_id XOR 2^i;
        j := my_id AND 2^i;
        k := (my_id XOR 2^i) AND 2^i;
        senloc := recloc + k;
        recloc := recloc + j;
        send msg[senloc .. senloc + 2^i − 1] to partner;
        receive temp[0 .. 2^i − 1] from partner;
        for j := 0 to 2^i − 1 do
            msg[recloc + j] := msg[recloc + j] + temp[j];
        endfor;
    endfor;
    result := msg[my_id];
end ALL_TO_ALL_RED_HCUBE
```

Similar pattern in reverse order.

Combine results

Algorithm 4.8 All-to-all broadcast on a d-dimensional hypercube. AND and XOR are bitwise logical-and and exclusive-or operations, respectively.

Cost Analysis (Time)

- Ring:
  - $T=(t_s + \boxed{t_w m)(p-1)}.$
- Mesh:
  - $T=(t_s + t_w m)(\sqrt{p}-1)+(t_s + t_w m\sqrt{p})(\sqrt{p}-1)$
    $= 2ts(\sqrt{p} - 1) + \boxed{t_w m(p-1)}.$
- Hypercube:
  - $$T = \sum_{i=1}^{\log p} (t_s + 2^{i-1} t_w m)$$

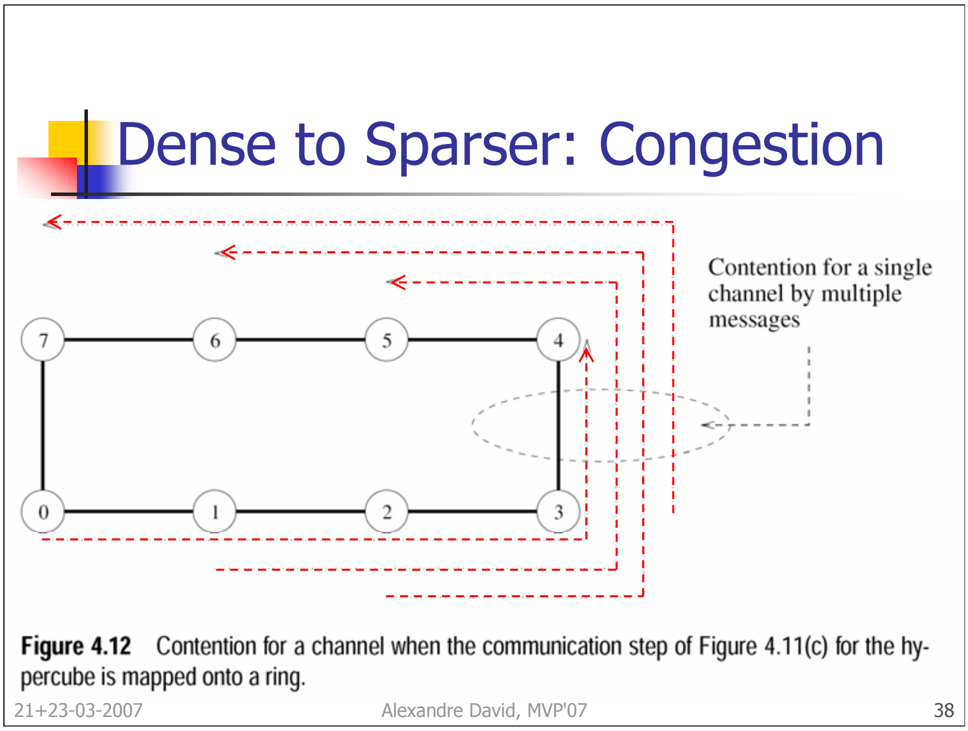
    log*p* steps
    message of size $2^{i-1}m$.

    $$= t_s \log p + \boxed{t_w m(p-1)}.$$

Lower bound for the communication time of all-to-all broadcast for parallel computers on which a node can communicate on only one of its ports at a time = $t_w m(p\text{-}1)$. Each node receives at least $m(p\text{-}1)$ words of data. That's for **any** architecture.

The straight-forward algorithm for the simple ring architecture is interesting: It is a sequence of *p* one-to-all broadcasts with different sources every time. The broadcasts are pipelined. That's common in parallel algorithms.

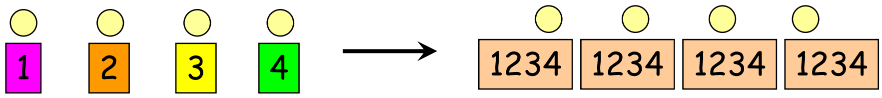
We cannot use the hypercube algorithm on smaller dimension topologies because of congestion.



Contention because communication is done on links with single ports. Contention is in the sense of the access to the link. The result is congestion on the traffic.

All-Reduce

- Each node starts with a buffer of size *m*.
- The final result is the same combination of all buffers on every node.
- Same as all-to-one reduce + one-to-all broadcast.
- Different from all-to-all reduce.



All-to-all reduce combines *p* different messages on *p* different nodes. All-reduce combines 1 message on *p* different nodes.

All-Reduce Algorithm

- Use all-to-all broadcast but
  - Combine messages instead of concatenating them.
  - The size of the messages does not grow.
  - Cost (in log*p* steps): T=$(t_s+t_w m)\log p$.

Prefix-Sum

- Given p numbers $n_0, n_1, \dots, n_{p-1}$ (one on each node), the problem is to compute the sums $s_k = \sum_{i=0}^{k} n_i$ for all k between 0 and $p-1$.
- Initially, n_k is on the node labeled k, and at the end, the same node holds S_k.

This is a reminder.

Prefix-Sum Algorithm

```
procedure PREFIX_SUMS_HCUBE(my_id, my_number, d, result)
begin
    result := my_number;
    msg := result;
    for i := 0 to d − 1 do
        partner := my_id XOR 2^i;
        send msg to partner;
        receive number from partner;
        msg := msg + number;
        if (partner < my_id) then result := result + number;
    endfor;
end PREFIX_SUMS_HCUBE
```

All-reduce

Prefix-sum

Algorithm 4.9 Prefix sums on a d-dimensional hypercube.

Prefix-Sum

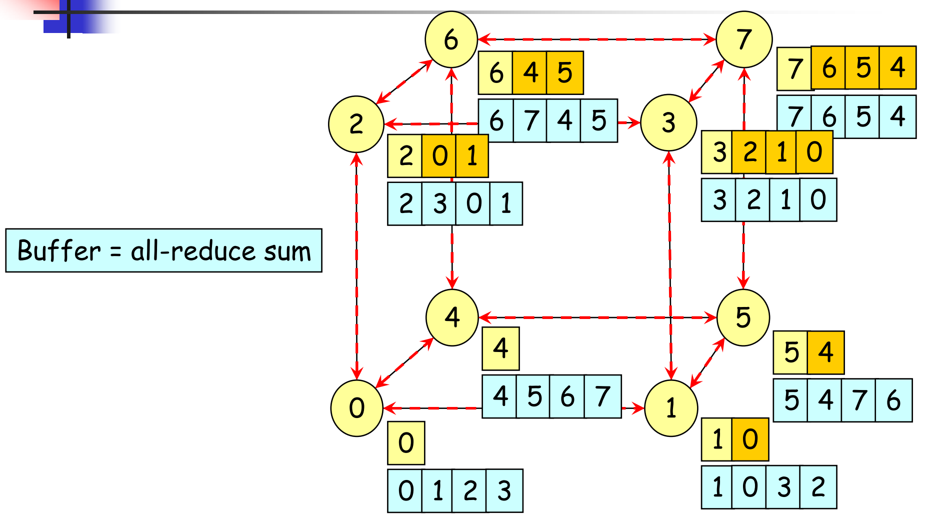
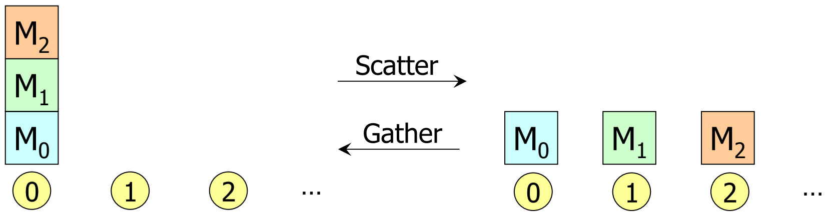


Figure in the book is messed up.

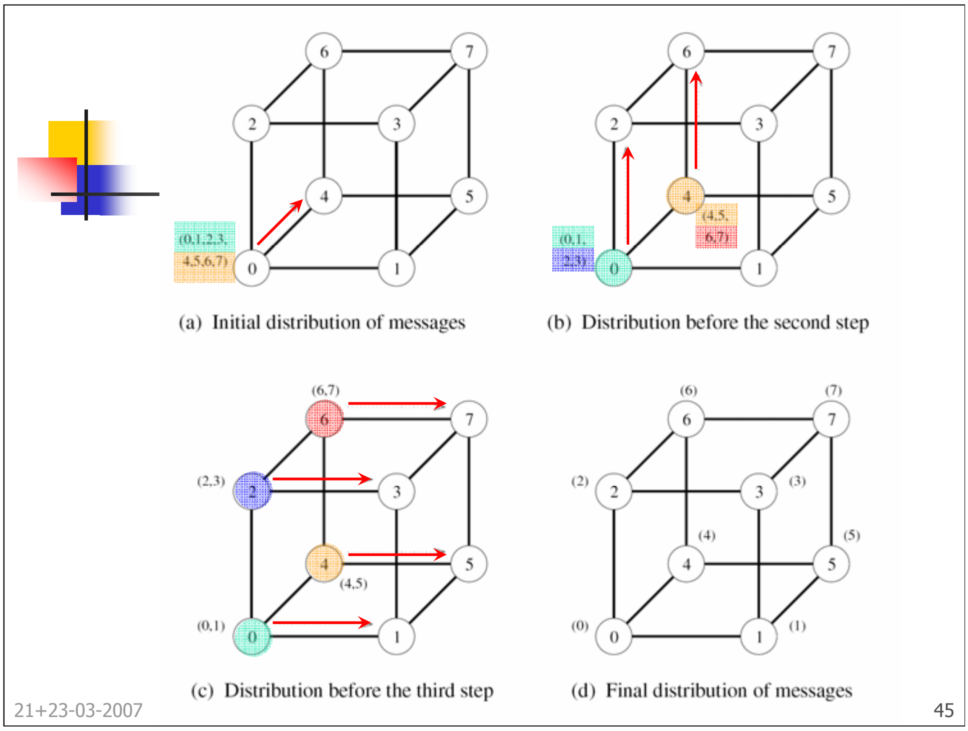
Scatter and Gather

- Scatter: A node sends a unique message to every other node – *unique per node*.
- Gather: Dual operation but the target node does not combine the messages into one.



Do you see the difference with one-to-all broadcast and all-to-one reduce? Communication pattern similar.

Scatter = one-to-all personalized communication.



The pattern of communication is identical with one-to-all broadcast but the size and the content of the messages are different. Scatter is the reverse operation. This algorithm can be applied for other topologies.

How many steps? What’s the cost?

Cost Analysis

- Number of steps: log*p*.
- Size transferred: *pm/2, pm/4,…,m*.
  - Geometric sum

$$p+\frac{p}{2}+\frac{p}{4}+\ldots+\frac{p}{2^n}=p\frac{1-\frac{1}{2^{n+1}}}{1-\frac{1}{2}}$$

$$\frac{p}{2}+\frac{p}{4}+\ldots+\frac{p}{2^n}=2p(1-\frac{1}{2^{n+1}})-p=2p(1-\frac{1}{2p})-p=p-1$$

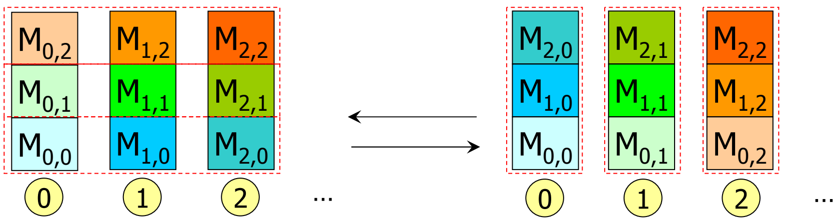
$$(2^{n+1}=2^{1+\log p}=2p)$$

- Cost $T=t_s\log p+t_w m(p-1)$.

The term $t_w m(p-1)$ is a lower bound for any topology because the message of size m has to be transmitted to p-1 nodes, which gives the lower bound of m(p-1) words of data.

All-to-All Personalized Communication

- Each node sends a *distinct* message to every other node.



See the difference with all-to-all broadcast?

All-to-all personalized communication = total exchange.

Result = transpose of the input (if seen as a matrix).

Example: Transpose

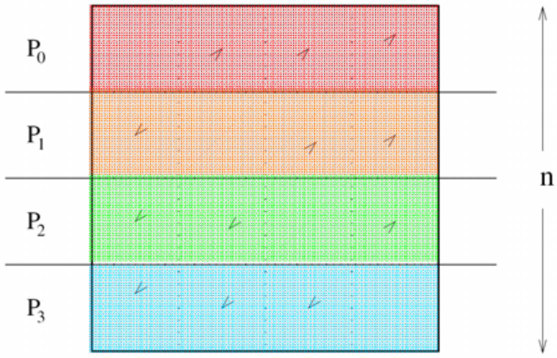
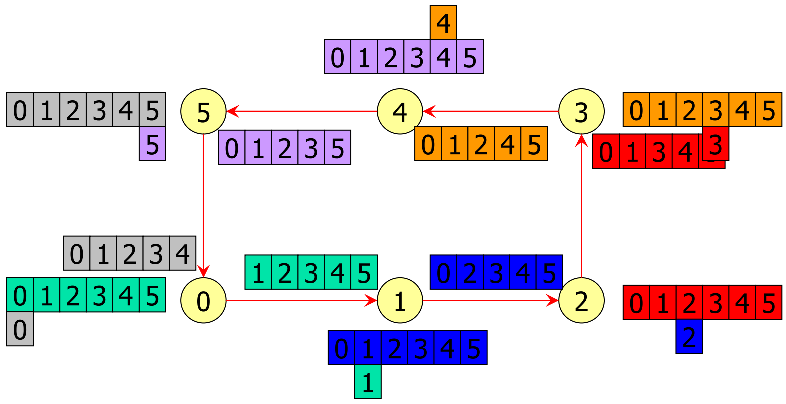
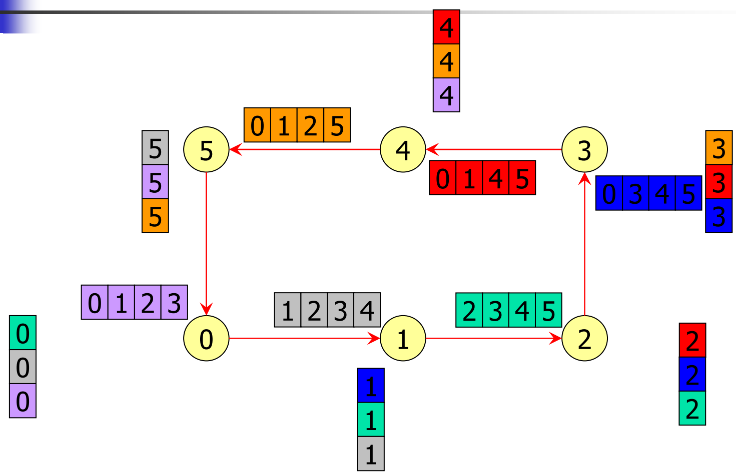


Figure 4.17 All-to-all personalized communication in transposing a 4 × 4 matrix using four processes.

Total Exchange on a Ring



Total Exchange on a Ring



Cost Analysis

- Number of steps: *p-1*.
- Size transmitted: *m(p-1),m(p-2)...,m*.

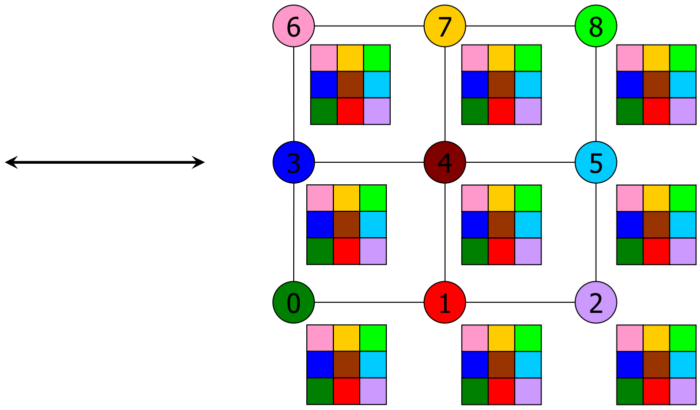
$$T = t_s(p-1) + \sum_{i=1}^{p-1} i t_w m = (t_s + t_w mp/2)(p-1)$$

$\boxed{\text{Optimal}}$

In average we transmit mp/2 words, whereas the linear all-to-all transmits m words. If we make this substitution, we have the same cost as the previous linear array procedure. To really see optimality we have to check the lowest possible needed data transmission and compare it to T.

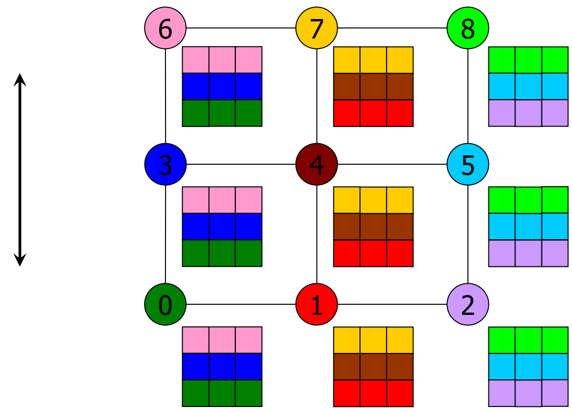
Average distance a packet travels = p/2. There are p nodes that need to transmit m(p-1) words. Total traffic = m(p-1)*p/2*p. Number of link that support the load = p, to communication time ≥ $t_w m(p-1)p/2$.

Total Exchange on a Mesh



We use the procedure of the ring/array.

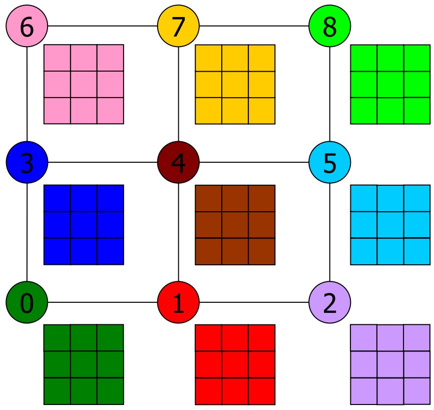
Total Exchange on a Mesh



21+23-03-2007 Alexandre David, MVP'07 53

We use the procedure of the ring/array.

Total Exchange on a Mesh



We use the procedure of the ring/array.

Cost Analysis

- Substitute *p* by $\sqrt{p}$ (number of nodes per dimension).
- Substitute message size *m* by $m\sqrt{p}$.
- Cost is the same for each dimension.
- $T=(2t_s+t_w mp)(\sqrt{p}-1)$

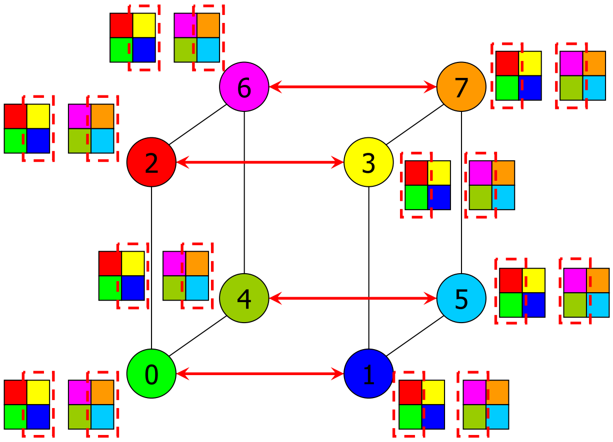
21+23-03-2007 Alexandre David, MVP'07 55

We have $p(\sqrt{p}-1)m$ words transferred, looks worse than lower bound in $(p-1)m$ but no congestion. Notice that the time for data rearrangement is not taken into account. It is almost optimal (by a factor 4), see exercise.

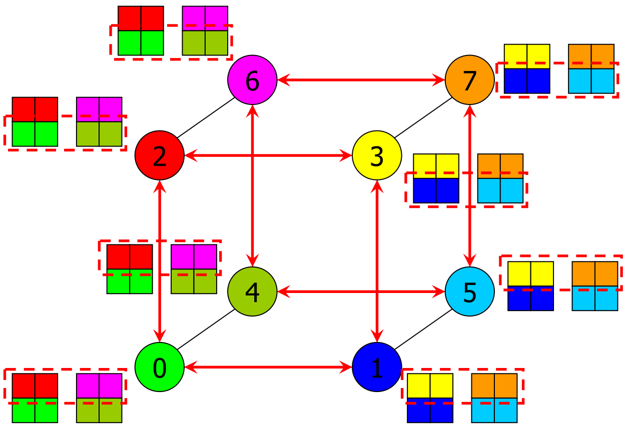
Total Exchange on a Hypercube

- Generalize the mesh algorithm to log*p* steps = number of dimensions, with 2 nodes per dimension.
- Same procedure as all-to-all broadcast.

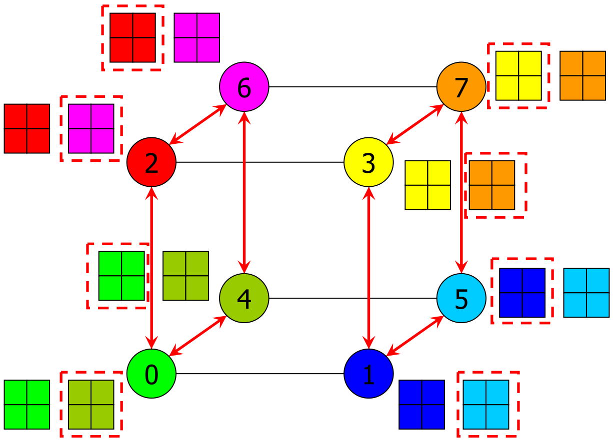
Total Exchange on a Hypercube



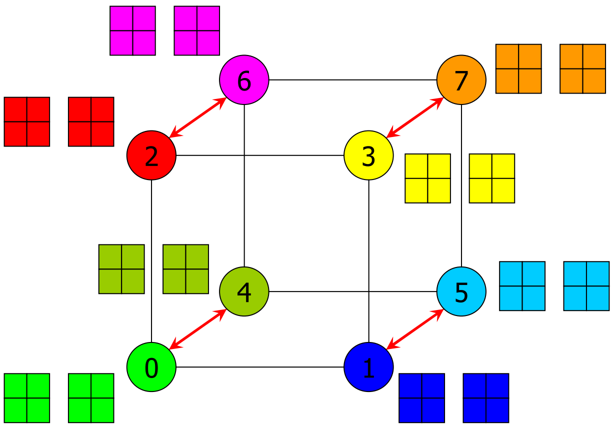
Total Exchange on a Hypercube



Total Exchange on a Hypercube



Total Exchange on a Hypercube



Cost Analysis

- Number of steps: log*p*.
- Size transmitted per step: *pm/2*.
- Cost: $T=(t_s+t_w mp/2)\log p$.
- Optimal? NO
- Each node sends and receives m(p-1) words. Average distance = $(\log p)/2$. Total traffic = $p*m(p-1)*\log p/2$.
- Number of links = $p \log p/2$.
- Time lower bound = $t_w m(p-1)$.

21+23-03-2007 Alexandre David, MVP'07 61

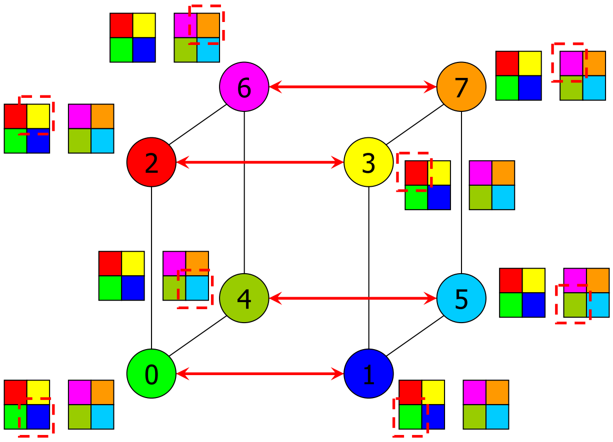
Notes:

1. No congestion.
2. Bi-directional communication.
3. How to conclude if an algorithm is optimal or not: Check the possible lowest bound and see if the algorithm reaches it.

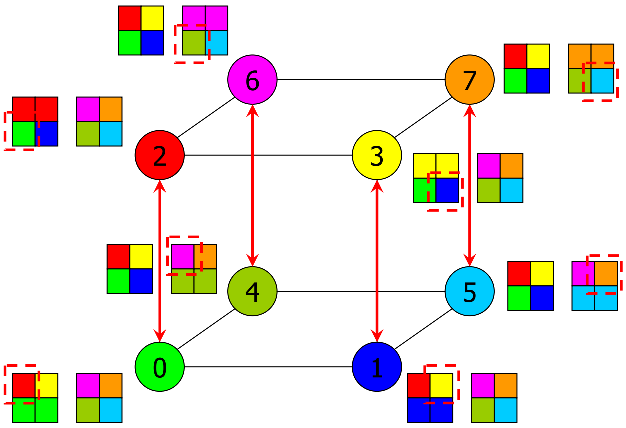
An Optimal Algorithm

- Have every pair of nodes communicate directly with each other – p-1 communication steps – but without congestion.
- At j^{th} step node i communicates with node $(i \text{ xor } j)$ with E-cube routing.

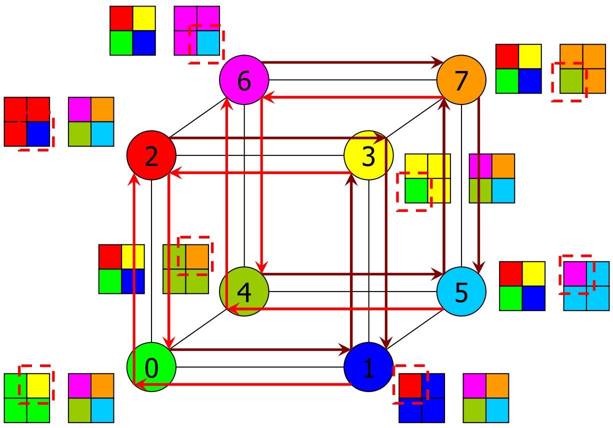
Total Exchange on a Hypercube



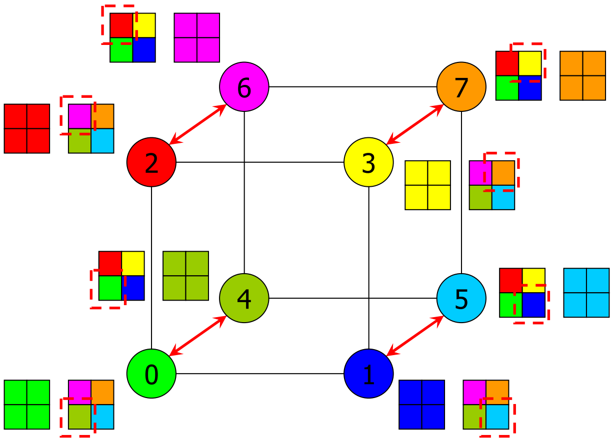
Total Exchange on a Hypercube



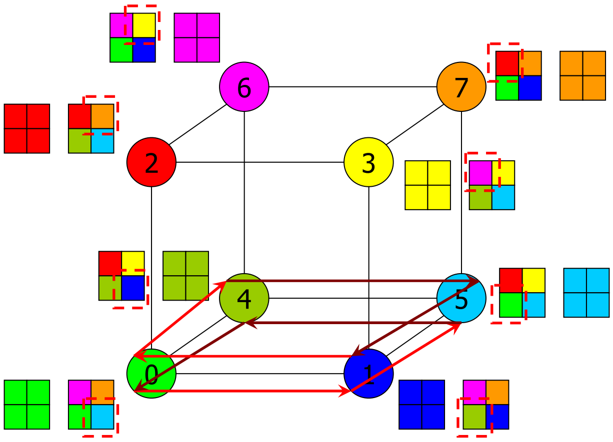
Total Exchange on a Hypercube

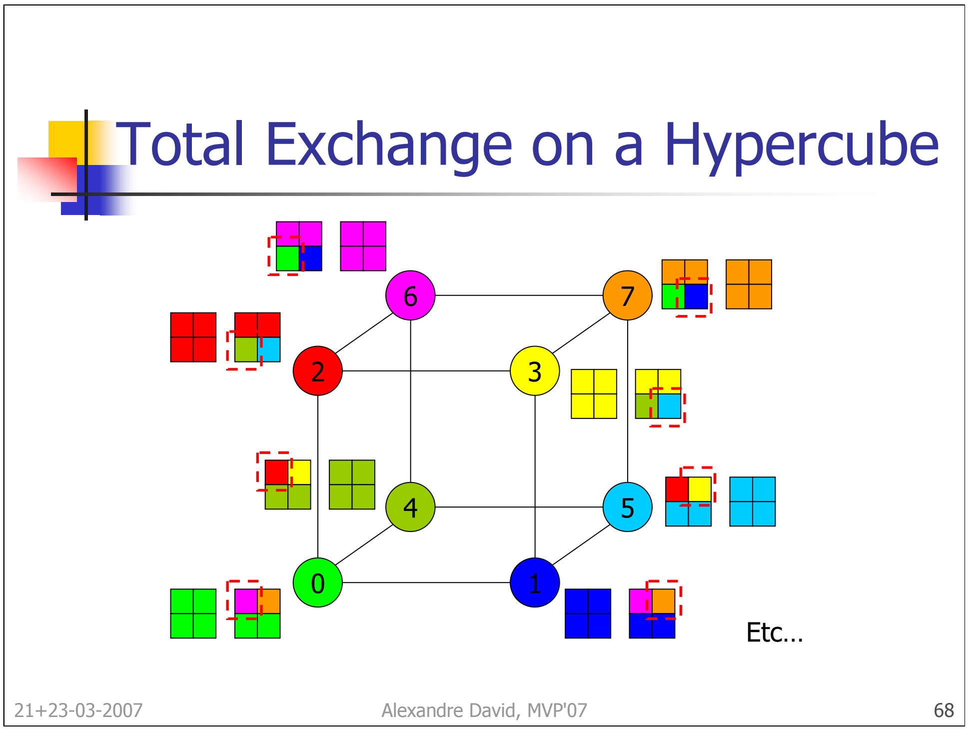


Total Exchange on a Hypercube



Total Exchange on a Hypercube





Point: Transmit less, only to the needed node, and avoid congestion with E-cube routing.

Cost Analysis

- Remark: Transmit less, only what is needed, but more steps.
- Number of steps: $p-1$.
- Transmission: size m per step.
- Cost: $T=(t_s+t_w m)(p-1)$.
- Compared with $T=(t_s+t_w mp/2)\log p$.
- Previous algorithm better for small messages.

This algorithm is now optimal: It reaches the lowest bound.

Circular Shift

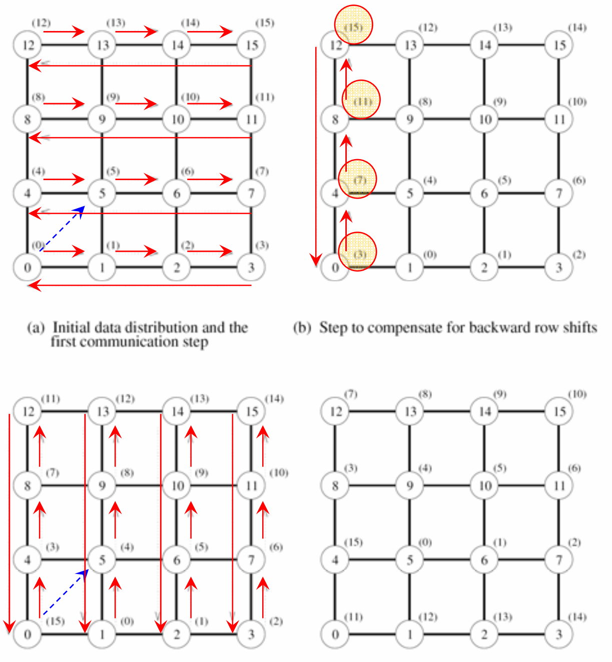
- It's a particular permutation.
- Circular q-shift: Node *i* sends data to node *(i+q) mod p* (in a set of p nodes).
- Useful in some matrix operations and pattern matching.
- Ring: intuitive algorithm in min{q,p-q} neighbor to neighbor communication steps. Why?

A permutation = a redistribution in a set.

You can call the shift a rotation in fact.

Circular 5-shift on a mesh.

q mod √p on rows
compensate
⌊q / √p⌋ on colums



(a) Initial data distribution and the first communication step

(b) Step to compensate for backward row shifts

(c) Column shifts in the third communication step

(d) Final distribution of the data

Circular Shift on a Hypercube

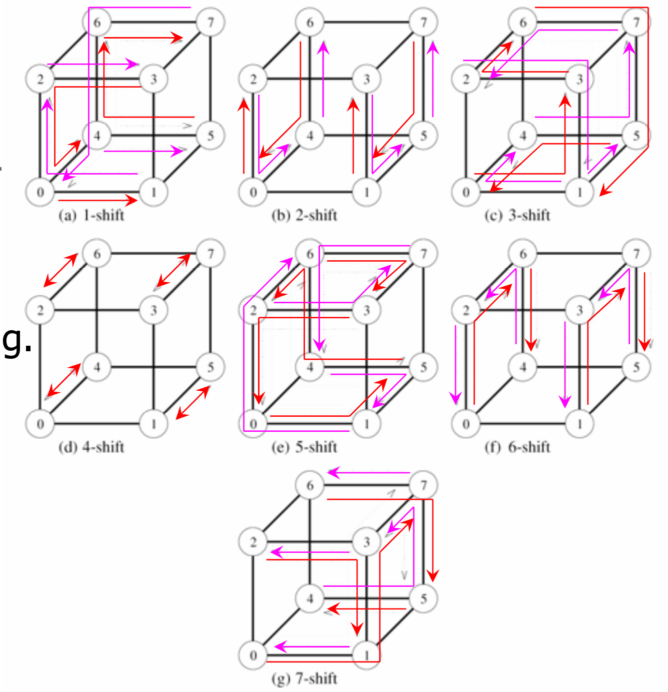
- Map a linear array with 2^d nodes onto a hypercube of dimension d.
- Expand q shift as a sum of powers of 2 (e.g. 5-shift = 2^0+2^2).
- Perform the decomposed shifts.
- Use bi-directional links for "forward" (shift itself) and "backward" (rotation part)… $\log p$ steps.

Backward and forward my be misleading in the book.

Interesting but not best solution, no idea why it's mentioned if the optimal solution is simpler.

Or better: Direct E-cube routing.

q-shifts on a 8-node hypercube.



Exercise: Check the E-cube routing and convince me that there is no congestion.

Communication time = $t_s+t_w m$ in one step.

Improving Performance

- So far messages of size *m* were not split.
- If we split them into *p* parts:
  - One-to-all broadcast = scatter + all-to-all broadcast of messages of size *m/p*.
  - All-to-one reduction = all-to-all reduce + scatter of messages of size *m/p*.
  - All-reduce = all-to-all reduction + all-to-all broadcast of messages of size *m/p*.