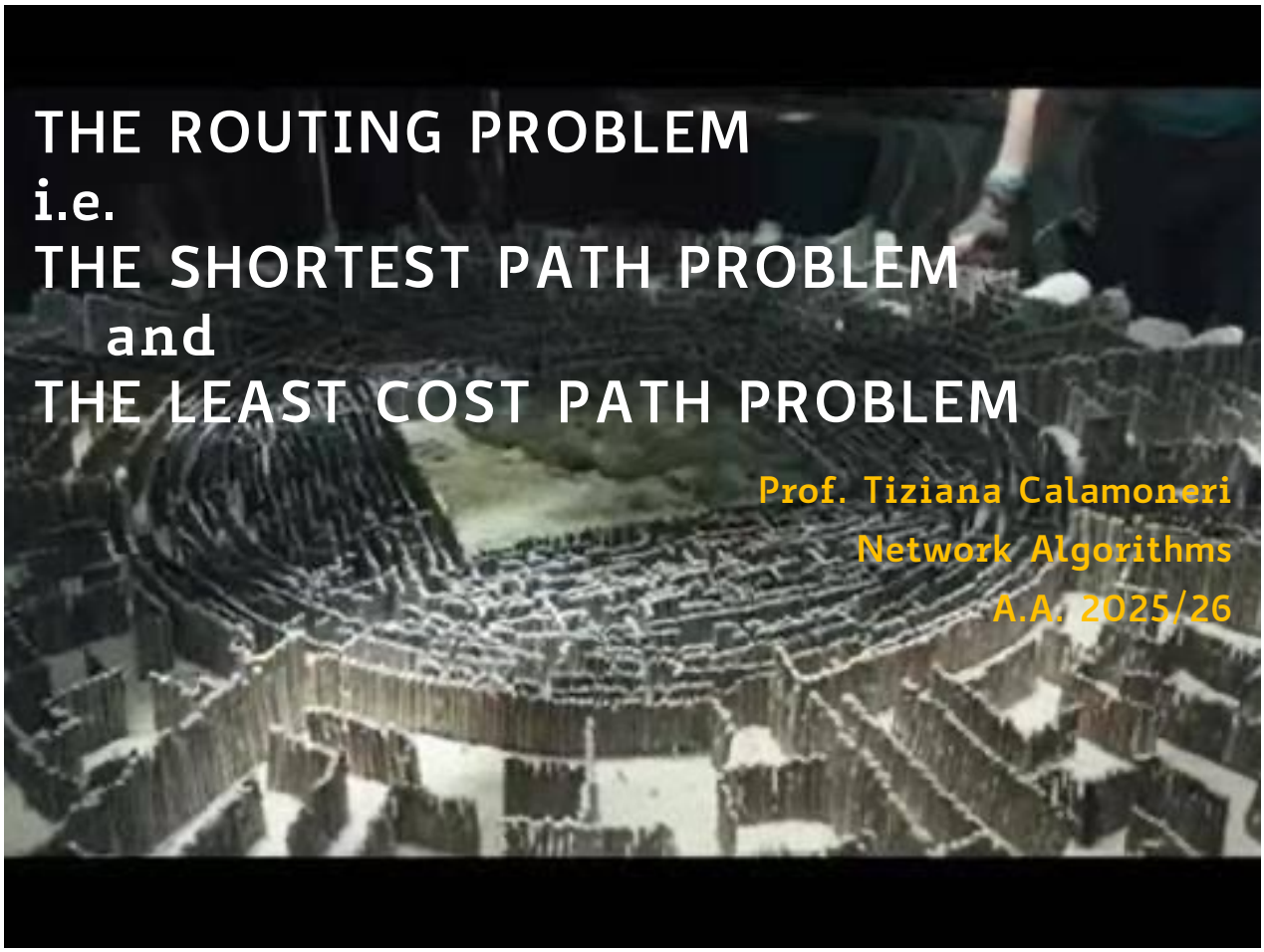
FIRST PART:
CABLE NETWORKS

1'



THE ROUTING PROBLEM
i.e.
THE SHORTEST PATH PROBLEM
and
THE LEAST COST PATH PROBLEM

Prof. Tiziana Calamoneri

Network Algorithms

A.A. 2025/26

THE PROBLEM

(already studied!)

3'

THE ROUTING PROBLEM (1)

Given a network:

- When packets are sent from a computer to another one through the network, each computer has to route data on a path passing through intermediate computers.
- This is the very general routing problem.

THE ROUTING PROBLEM (2)

Case 1. Not adaptive routing

A routing algorithm could try to send packets through a network so that the length of the used path is minimized. Such length can be measured in terms of number of hops between pairs of computers.

If the network is modeled as a graph (nodes = computers and edges = links), the problem reduces to the shortest path problem between two nodes.

THE ROUTING PROBLEM (3)

Case 2. Adaptive Routing

It takes into account the traffic conditions: in order to decide next step, the traffic is estimated, so the packet is sent toward the zones of the network not affected by traffic.

If the network is modeled by an edge-weighted graph (nodes = computers, edges = links and weights = dynamic values proportional to the traffic on the connection), the problem reduces to the (dynamic) least cost path problem.

THE ROUTING PROBLEM (4)

cases 1 and 2. Adaptive and non adaptive routings (cntd)

Non adaptive routing:

- Good results with consistent topology and traffic
- Poor performance if traffic volume or topologies change over time
- Information about the entire network has to be available
- Each packet is routed through an outgoing edge in a fixed way
- Routing tables are **used**

THE ROUTING PROBLEM (5)

cases 1 and 2. Adaptive and non adaptive routings (cntd)

Adaptive routing:

- Good results when the network's workload is high or unbalanced
- extra logic overhead in acquiring information, path arbitration and deadlock avoidance
- Decisions are based on current network state
- Packets follow dynamically computed routes
- Routers are able to communicate
- Rather often re-calculations are necessary
- Each router **creates** its own routing table

THE ROUTING PROBLEM (6)

cases 1 and 2. Adaptive and non adaptive routings (cntd)

- Half-adaptive routing:
- it switches from one mode to another, depending on the evaluation of current workload.
- Half-adaptive algorithms make significant reduction of complexity and overhead, though lose some path diversity.

THE ROUTING PROBLEM (6)

Case 3. Routing with faults

When the network is modeled as a graph, the length (edge-weight) of an edge may also represent the probability of its failing (used, for instance, in networks of thelephone lines, or broadcasting systems in computer networks or in transportation routes). In all these cases, one is looking for the route having the highest probability not to fail.

More precisely…

THE ROUTING PROBLEM (7)

case 3. Routing with faults (cntd)

- Let $p(e)$ be the probability that edge e does not fail. Under the -not always realistic- assumption that failings of edges occur independently of each other, $p(e_1)\cdot p(e_2)\cdot\ldots\cdot p(e_k)$ gives the probability that the path $P=(e_1,e_2,\ldots,e_k)$ can be used without any faults.
- We want to maximize this probability over all possible paths with starting point a and arrival point b...

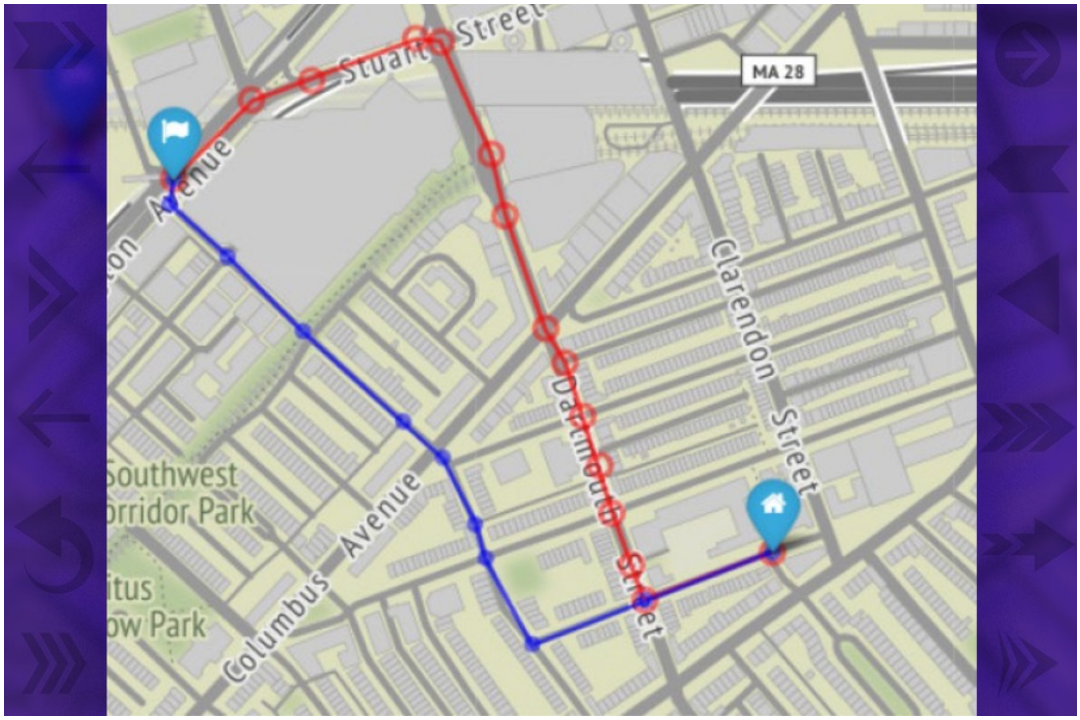
THE ROUTING PROBLEM (8)

case 3. Routing with faults (cntd)

- Note. Since function log is monotonic increasing, the maximum of the product $p(e_1)\bullet p(e_2)\bullet\ldots\bullet p(e_k)$ is reached iff the logarithm of the product is maximum, i.e. iff:

  $log\ p(e_1)+log\ p(e_2)+\ldots+ log\ p(e_k)$ is maximum.
- $log\ p(e) \le 0$ for each e because $p(e) \le 1$.
- Define $w(e)=-log\ p(e)$, then $w(e)\ge 0$ for all e; furthermore, we have to find a path from a to b for which

  $w(e_1)+w(e_2)+\ldots+w(e_k)$ becomes minimum.
- Thus, our problem is reduced again to the least cost path problem.



THE SHORTEST PATH PROBLEM and THE LEAST COST PATH PROBLEM

SHORTEST PATHS (1)

- Let *G=(V,E)* be a graph; let *w(e)* be the length of each edge *e*.
- Many versions of the shortest path problem:
  - All to all
  - One to one
  - One to all
  - All to one
- Lengths can be:
  - All equal (unit length)
  - Non negative
  - Possibly negative but without negative cycles
  - Creating possible negative cycles

SHORTEST PATHS (2)

- Algorithm designed by Moore ['59] for the one-to-all shortest path problem and unit lenghts:
  … Breadth First Search (BFS) …

- TH. G is connected iff at the end of the BFS starting from node a, $dist(a,b) < \infty$ for each node b, where *dist* is the distance in terms of number of edges.
- Note. This claim is false if G is a digraph (indeed the notion of connected graph does not exist on digraphs: strong and weak connection…)

A PARENTHESIS: CONNECTED COMPONENTS (1)

A parenthesis on BFS (and its cousin, DFS): it can be used to detect connected components.
It is of fundamental importance and can be applied to a wide range of CS fields.

A PARENTHESIS: CONNECTED COMPONENTS (2)

For example it is useful to:

- identify objects and scenes in images and videos (vision),
- analyze the structure and the evolution of evolving (social, transportation, biological, …) networks (ntw analysis),
- identify clusters and communities in large datasets (data mining),
- group together similar spam messages to detect spam campaigns (cybersecurity).

A PARENTHESIS: CONNECTED COMPONENTS (3)

In data science:

Having a (huge) set of customers using each many accounts, recognize the set of accounts related to the same user.

Construct a graph:

nodes: accounts labeled with CustomerIDs (based on the same credit card, same mobile number, etc.)

edges: between two accounts with the same CustomerID.

A PARENTHESIS: CONNECTED COMPONENTS (4)

The connected component algorithm creates individual clusters to which we can assign the same user.

We can then use these grouped accounts to provide personalized recommendations or capture fraud (if an account has done fraud in the past, it is highly probable that the connected accounts are also susceptible to fraud).

☞ student lesson on connected components

LEAST COST PATHS (1)

Let $G=(V,E)$ be a graph or a digraph and let $w: E \to \mathbb{R}$ be an edge-weight function.

- (G,w) is called network.
- $w(e)$ is called length (though including meanings such as cost, capacity, weight, probability, …)
- For each path $P=(e_1, e_2, \ldots, e_k)$ (if G is a digraph, P is a dipath), the length of P is defined as $w(P)=w(e_1)+w(e_2)+ \ldots +w(e_k)$.
- Note. If $w(e)=1$ for each edge, the least cost path problem reduces to the shortest path problem.

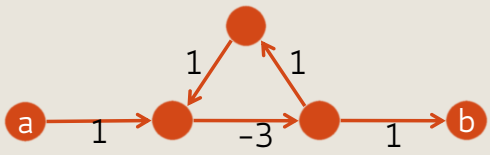
LEAST COST PATHS (2)

Given two nodes *a* and *b*, the distance *d(a,b)* is defined as the minimum, over all the paths *P* connecting *a* and *b*, of *w(P)*.

Two problems arise:

- PR.1: *b* could be unreachable from *a*
- SOL.: define *d(a,b)=∞* if *b* is unreachable from *a*

- PR.2: the minimum could not exist (cycles of negative length)
- SOL.: only networks without cycles of negative length are feasible



LEAST COST PATHS (3)

Negative lengths may occur!

Example: good shipment

- A ship travels from port *a* to port *b*, where the route (and possible intermediary ports) may be chosen freely.
- The length *w(x,y)* signifies the profit gained by going from *x* to *y*.
- For some edges, the ship might have to travel empty so that *w(e)* is negative for these edges: the profit is actually a loss.
- Replacing *w(e)* by *−w(e)* for all *e* in this network, the shortest path represents the route which yields the largest possible profit.

LEAST COST PATHS (4)

- In general, when w represents a gain, it seems natural to replace $w(e)$ by $-w(e)$ and look for least cost paths, but this could introduce cycles of negative weigth.
- There exist good algorithms that find minimum weight paths even when G contains cycles of negative weight.

LEAST COST PATHS (5)

Also negative cycles may occur and even be useful!

Example: arbitrage opportunity in finance

- Consider a market for financial transactions that is based on trading commodities.
- Imagine that you could convert 1000 USD to 950 EUR and then 950 EUR to 1020 CAD which you convert back to 1007 USD, gaining money.
- This situation is called arbitrage opportunity.

LEAST COST PATHS (6)

A table shows conversion rates among currencies:

	USD	EUR	GBP	CHF	CAD
USD	1	0.741	0.657	1.061	1.061
EUR	1.061	1	0.888	1.433	1.433
GBP	1.521	1.126	1	1.614	1.538
CHF	0.942	0.698	0.619	1	0.953
CAD	0.995	0.732	0.650	1.049	1

This data can be modeled as a graph problem:

graph: complete directed graph

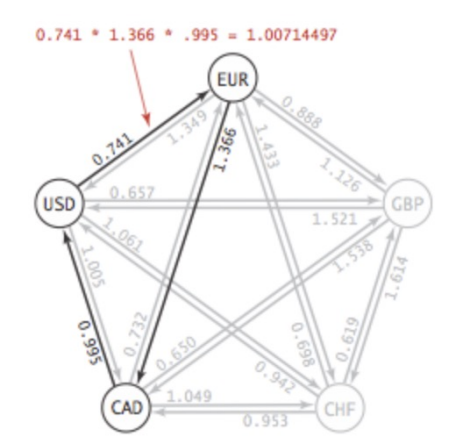
nodes: currencies

edge-weight: conversion rate

LEAST COST PATHS (7)

An arbitrage opportunity is a directed cycle such that the product of the exchange rates is greater than one.

	USD	EUR	GBP	CHF	CAD
USD	1	0.741	0.657	1.061	1.061
EUR	1.061	1	0.888	1.433	1.433
GBP	1.521	1.126	1	1.614	1.538
CHF	0.942	0.698	0.619	1	0.953
CAD	0.995	0.732	0.650	1.049	1



LEAST COST PATHS (8)

To formulate the arbitrage problem as a negative-cycle detection problem, replace each weight by its logarithm, negated (as we have already done). With this change, computing path weights by multiplying edge weights in the original problem corresponds to adding them in the transformed problem.

LEAST COST PATHS (9)

OBSERVATIONS CONCERNING THE SOLUTION

In any solution:

- Cycles having negative length cannot exist (we avoided them by hypthesis)
- Cycles having positive length cannot exist (by contradiction: if one of them is in the solution, the new solution without it has a lower cost)
- Cycles having null length do not exist without loss of generality: if one of them is in the solution, the new solution without it has the same cost and so is feasible, too
- So: our solution does not contain any cycles and hence it passes through at most $n-1$ edges.

LEAST COST PATHS (10)
OBSERVATIONS CONCERNING THE SOLUTION (CNT.D)

- In order to univocally determine a path from *a* to *b* it is enough, for each node in such path, starting from *b* and coming back, to store its predecessor on the path.
- To do it: for each node *v* in *G* define a pointer *pt(v)*, initially equal to NULL; at the end, it points at the predecessor of *v* on the path.

SOME CLASSICAL ALGORITHMS
(already studied!)

BELLMAN-FORD ALGORITHM ['58]

- *G=(V,E)* directed with edge-weights possibly negative
- It solves the problem of the shortest path from single source, hence it outputs the distances from the (single) source to each node
- It assumes that *G* does not contain any cycles of negative length
- It is based on the principle of relaxation
- Time complexity: *O(nm)*

DIJKSTRA ALGORITHM ['59]

- *G=(V,E)* directed with non negative edge-weights
- It solves the problem of the shortest path from single source, hence it outputs the distances from the (single) source to each node
- It is based on the principle of relaxation
- Time complexity: either $O(n^2)$ or $O(m \log n)$
- The time complexity of the Dijkstra Algorithm is better than the time complexity of the Bellman-Ford Algorithm, but it is less versatile, as it requires not negative edge weight edges.

FLOYD-WARSHALL ALGORITHM ['62]

- $G=(V,E)$ directed with edge-weights possibly negative
- It solves the problem of the all pairs shortest path, hence it outputs a matrix with the distances from each node to each other node
- Repeatedly applying the algs treated before, varying the source over all nodes in V:
  - Bellman-Ford: $n\ O(nm)=O(n^2m)$
  - Dijkstra: $n\ O(n^2)=O(n^3)$ o $n\ O(m\ log\ n)=O(mn\ log\ n)$
- Time complexity: $O(n^3)$ and negative edge weights are allowed

THE RELAXATION

- For each node v, let $d(v)$ be a function representing an estimate of the weight of the shortest path from s to v.
- At the beginning $d(v)=\infty$ for each v
- One relaxation step is performed as follows:
  - Given an edge (u,v)
  - If $d(u)+w(u,v)<d(v)$

    $d(v)=d(u)+w(u,w)$

    $pt(v)=u$
- Time complexity of one relaxation step: $O(1)$

BELLMAN-FORD ALGORITHM (1)

- Assume that G does not contain any cycles of negative length

  For each v initialize $d(v)$ and $p(v)$ — $\Theta(n)+$

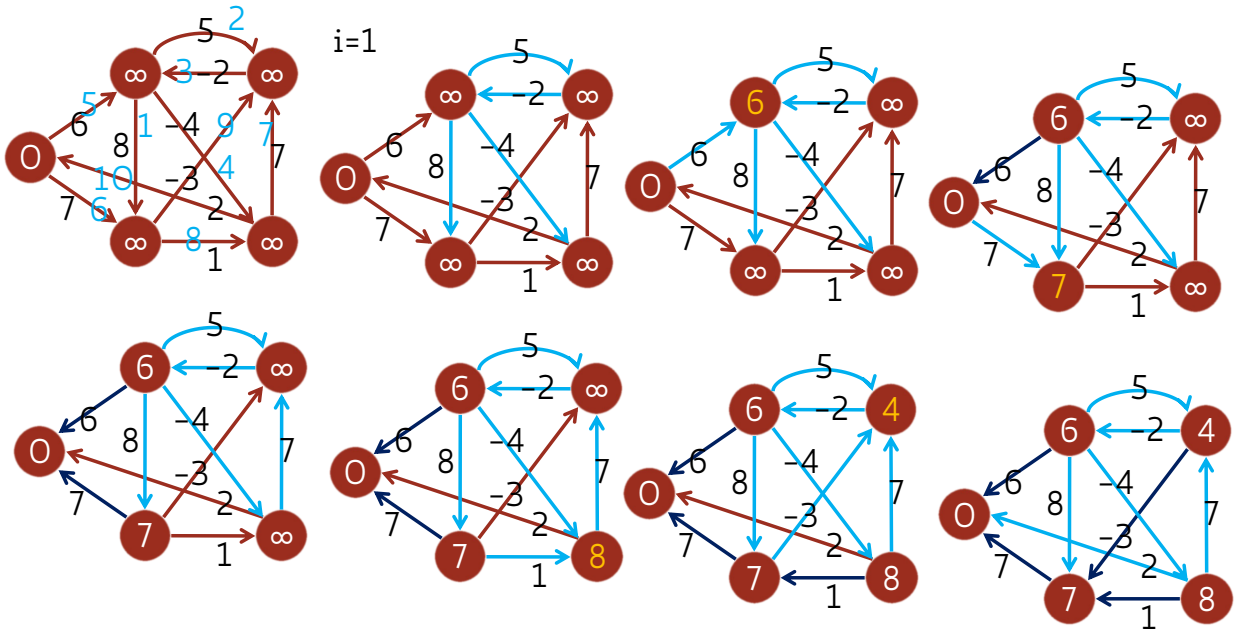
  For $i=1$ to $n-1$ do — $|n-1$ times

  For each (u,v) relax v w.r.t. (u,v) — $| m\ \Theta(1)$

- Time Complexity: $\Theta(nm)$

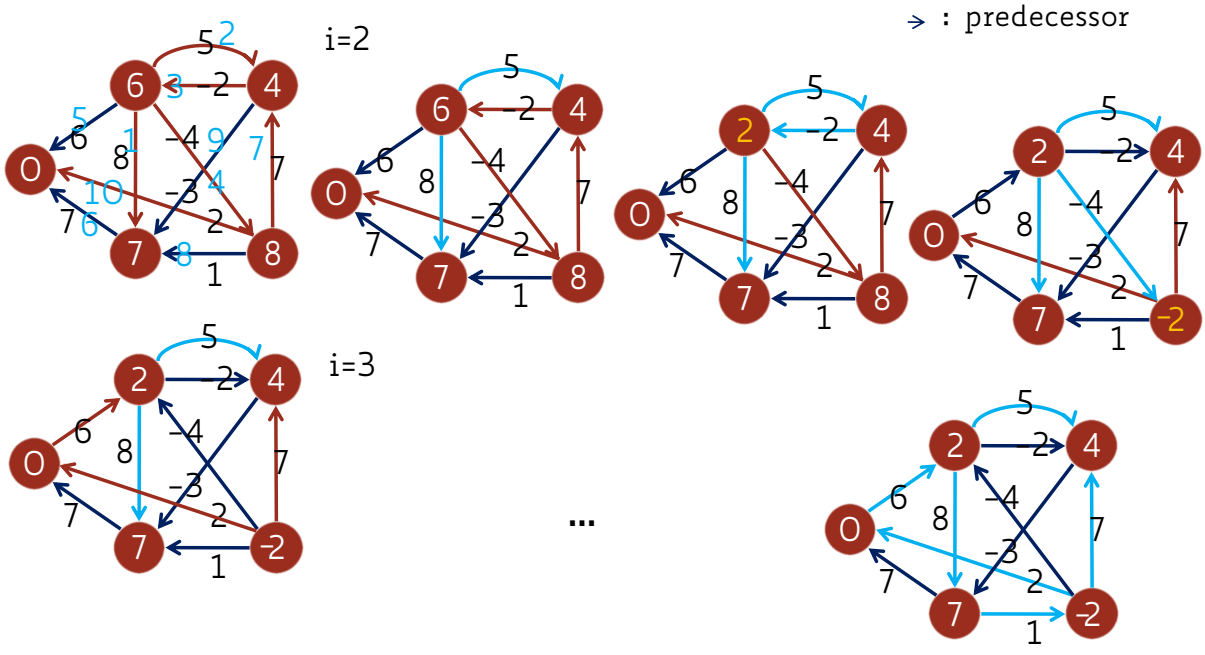
BELLMAN-FORD ALG. (2)

x: order of the edges
→ : visited edges
→ : predecessor

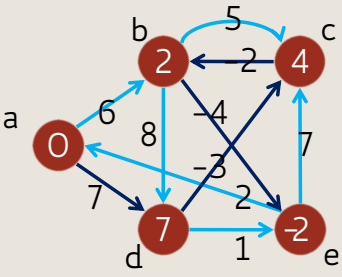


BELLMAN-FORD ALG. (3)

x: order of the edges
→ : visited edges
→ : predecessor



BELLMAN-FORD ALGORITHM (4)



(Let us consider predecessors in the same direction than original edges)

Forwarding table of *a*:

b *(a,d)* as the shortest path is *a-d-c-b*
c *(a,d)* as the shortest path is *a-d-c*
d *(a,d)*
e *(a,d)* as the shortest path is *a-d-c-b-e*

BELLMAN-FORD ALGORITHM (5)

Bellman-Ford Algorithm is used for *Distance Vector routing*, an iterative, asynchronous and distributed protocol:

- $c(x,v)$ = cost for (directed) link from x to v
- $D_x(y)$ = estimate of least cost from x to y; x mantains distance vector D_x = [$D_x(y)$ for each y neighbor of x]; for each neighbor y, x mantains also D_y
- …

BELLMAN-FORD ALGORITHM (6)

- …
- Each node x periodically sends D_x to its neighbors
- Neighbors update their own distance vector:
  $D_x(y)=min\{ D_x(y), c(x,v)+D_v(y) \}$
- x notifies neighbors when its distance vector changes
- Over the time, D_x converges

An example: Routing Information Protocol (RIP)

DIJKSTRA ALGORITHM (1)

- $G=(V,E)$ directed with non negative edge-weights
- It partitions the nodes of G: nodes whose shortest path from s has already been found (S) and all the other nodes (V-S)
- Greedy algorithm
- At each step, let u be the node in V-S with minimum value of d; add u to S and relax all the edges outcoming from u
- Keep V-S in a priority queue (e.g. min heap)

DIJKSTRA ALGORITHM (2)

- For each v initialize $d(v)$ and $pt(v)$
- S=empty set
- Q= V
- While Q is not empty
  - u=ExtractMin(Q)
  - S=S $\cup$ $\{u\}$
  - For each edge (u,v) outcoming from u
    relax v w.r.t. (u,v)
    Update Q

The time complexity depends on the data structure used to implement Q:

Queue: $O(n^2)$
Heap: $O(m \log n)$
Fibonacci Heap: $O(m+n \log n)$

DIJKSTRA ALGORITHM (3)

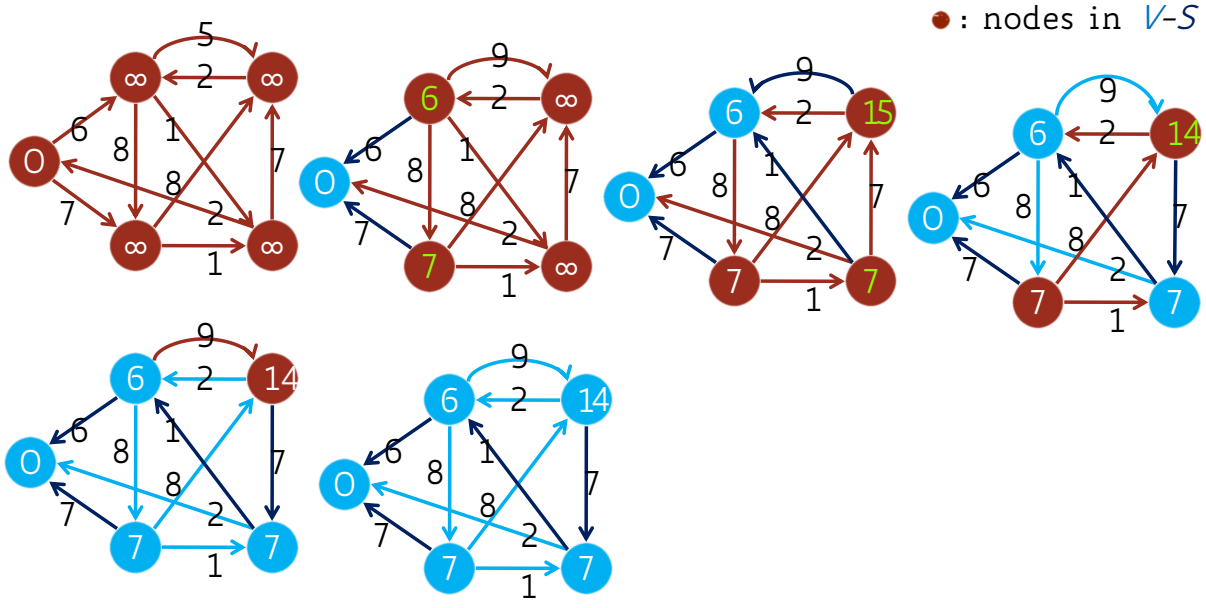
Using heap:

- For each v initialize $d(v)$ and $pt(v)$ — $\Theta(n)$
- S=empty set — $\Theta(1)$
- Q=V — $\Theta(n)$
- While Q is not empty — n times
  - u=ExtractMin(Q) — | $O(\log n)$
  - S=S U $\{u\}$ — | $\Theta(1)$
  - For each edge (u,v) outcoming from u — | $\Theta(deg\ u)$ times
    - relax v w.r.t. (u,v) — | | $\Theta(1)$
    - Update Q — | | $O(\log n)$

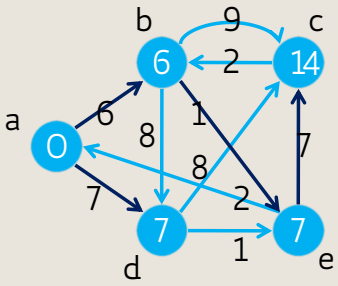
tot. $O(n \log n + m \log n) = O(m \log n)$

DIJKSTRA ALG. (4)

: visited edges
: predecessor
: nodes in S
: nodes in V-S



DIJKSTRA ALGORITHM (5)



(Let us consider drawings of predecessors in the same direction than original edges)

Forwarding table of a:

b *(a,b)*
c *(a,b)* as the shortest path is *a-b-e-c*
d *(a,d)*
e *(a,b)* as the shortest path is *a-b-e*

DIJKSTRA ALGORITHM (6)
APPLICATION

In Dijkstra's own words:
«What is the shortest way to travel from Rotterdam to Groningen, in general: from given city to given city?
It is the algorithm for the shortest path.»
Nowadays, variations of the Dijkstra algorithm are used extensively in GoogleMaps to find the shortest routes.

DIJKSTRA ALGORITHM (7)
APPLICATION

Also used for dynamic routing protocols.

Each router:

- Keeps trace of its incident links
  - Whether the link is up or down
  - The cost on the link (varying in time)
- Broadcasts the link state (flooding)
  - So, every router has a complete view of the graph
- Runs Dijkstra Algorithm
  - To compute the shortest paths…
  - …and construct the forwarding table

An example: Open Shortest Path First (OSPF) used in the networks with Internet Protocol (IP)

FLOYD-WARSHALL ALGOR. (1)

Sometimes, it is not enough to calculate the distances w.r.t. a certain node s: we need to know the distances between all pairs of nodes.

- $G=(V,E)$ directed with any edge-weight.
- Algorithm for the all-to-all shortest path problem
- **Trick 1:** all edges are in; the non-existing ones have $w=\infty$
- **Trick 2:** in order to go from i to j you can either go directly or passing through a third node k
- dynamic programming

FLOYD-WARSHALL ALGOR. (2)

Algorithm:

- For each node i, initialize $dist(i,i)=0$ $\Theta(n)$
- For each edge (i,j) initialize $dist(i,j)=w(i,j)$ $\Theta(n^2)$
- For each node k n times
  - For each node i n times
    - For each node j n times

    $dist(i,j)=min\{dist(i,j), dist(i,k)+dist(k,j)\}$ $\Theta(1)$

- Time Complexity $\Theta(n^3)$

OTHER APPLICATIONS (1)

Difference Constraints:

- Let be given some tasks with precedence constraints and running lengths, and an unlimited (or limited by n=number of tasks) number of processors:
  - Each task i has:
    - Starting time s_i
    - Time to complete $b_i>0$
    - Constraint $s_j+b_j \leq s_i$ if task i can be started after that task j has been completed
  - First task can start at time 0
  - When can we finish last task?

OTHER APPLICATIONS (2)

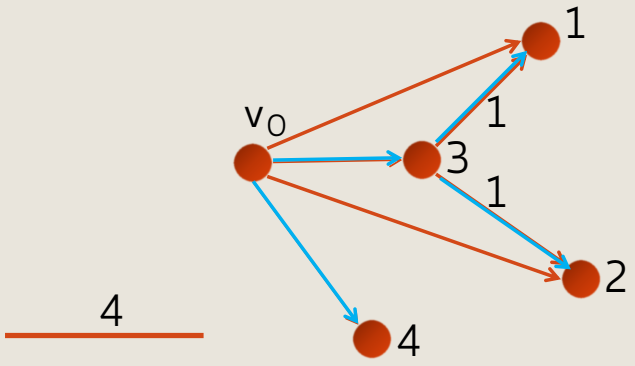
This is the Shortest Path Problem on directed acyclic graphs:

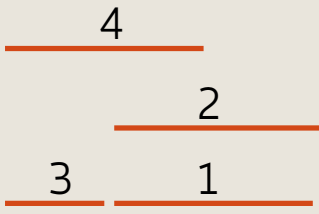
- Define a graph having a node for each task
- Insert a dummy node v_0 (that models time 0)
- Insert an arc (v_0,i) for each task node i and let 0 be its weight
- For each precedence constraint $s_j+b_j \leq s_i$ insert an arc (j,i) with weight b_j
- Optimal Solution: start each task i at time equal to the length L_i of the longest path from v_0 to i. All the tasks are surely completed within time $max_i\ (L_i+b_i)$

OTHER APPLICATIONS (3)

To transform to the shortest path problem, multiply all lengths and times by -1:

$b_1=2$
$b_2=3$
$b_3=1$
$b_4=2$
$s_3+b_3 \leq s_2$
$s_3+b_3 \leq s_1$





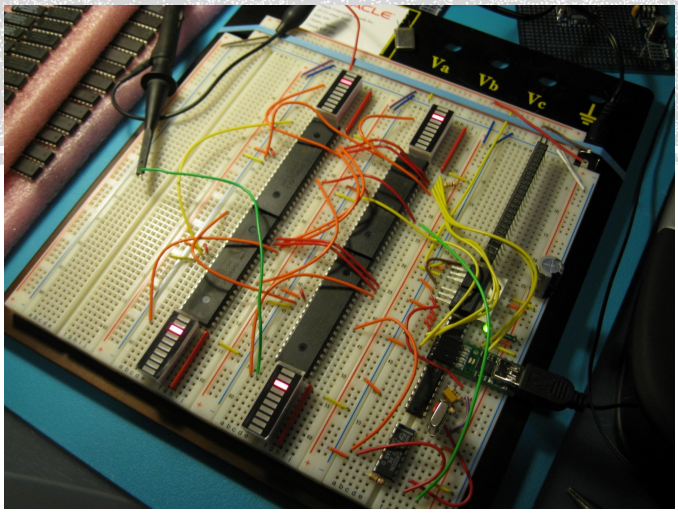
OTHER APPLICATIONS (4)

Social networks' friendship:

Often, social networks suggest the list of friends that a particular user may know.

The Dijkstra algorithm is usually applied using the shortest path between users measured through handshakes or connections among them.

PACKET-ROUTING ALGORITHMS ON INTERCONNECTION TOPOLOGIES



PACKET-ROUTING ON INTERCONNECTION TOPOLOGIES (1)

- Up to now, in the routing problem we have considered the network as a graph unknown to the nodes and variable in time (faults, varying traffic, etc.)
- Nevertheless, when the network is an interconnection topology (and connects, for example, processors), it is known and fixed in time. Furthermore, efficiency is a primary issue.
- Solutions having stronger properties than the simple shortest path algorithms are required.

A PARENTHESIS ON INTERCONNECTION TOPOLOGIES (1)

- With the development of semiconductor industry, the number of cores integrated in a single machine increases quickly and Network-on-Chip (NoC) is proposed and is gradually replacing the traditional on-chip interconnections such as sharing buses and crossbars.

A PARENTHESIS ON INTERCONNECTION TOPOLOGIES (2)

- An interconnection topology describes how the processors are connected inside a multicore machine.
- The routing algorithm, which decides the paths of packets, has significant impact on the latency and throughput of the network, so it plays a vital role in a well-performed network.

PACKET-ROUTING ON INTERCONNECTION TOPOLOGIES (2)

Many different types of routing models.
Here, we will focus on the store-and-forward model (also known as the packet-switching model):

- Data are divided into packets
- Each packet is maintained as an entity that is passed from node to node as it moves through the network
- A single packet can cross each edge during each step of the routing
- …

PACKET-ROUTING ON INTERCONNECTION TOPOLOGIES (3)

- Depending on the algorithm, packets may or may not be piled up in queues located at each node. When queues are allowed: effort to keep them short.
- Global controller to precompute routing paths not allowed: problem handled using only local control
- …

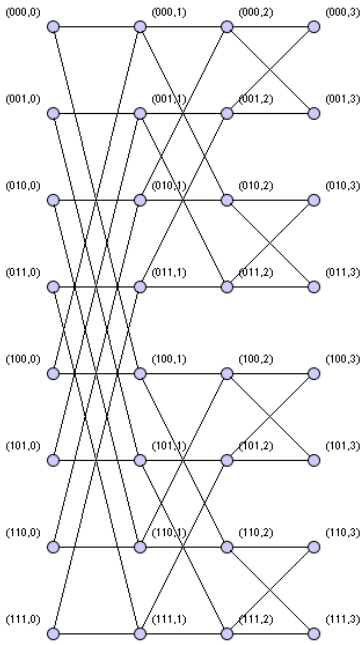
PACKET-ROUTING ON INTERCONNECTION TOPOLOGIES (4)

A routing problem is called one-to-one if at most one packet must be addressed to every node and each packet has a different destination.
In contrast, one-to-many and many-to-one

BUTTERFLY NETWORK (1)

Def. Let $N=2^n$ (hence $n=\log N$);
the *n*-dimensional Butterfly is a layered graph with:

- $N(n+1)$ nodes ($n+1$ layers with 2^n nodes each) and
- $2Nn$ edges.



Nodes:

nodes correspond to pairs *(w, i)*, where:

- *i* is the *layer* of the node
- *w* is an *n*-bit *binary number* that denotes the *row* of the node.

...

BUTTERFLY NETWORK (2)

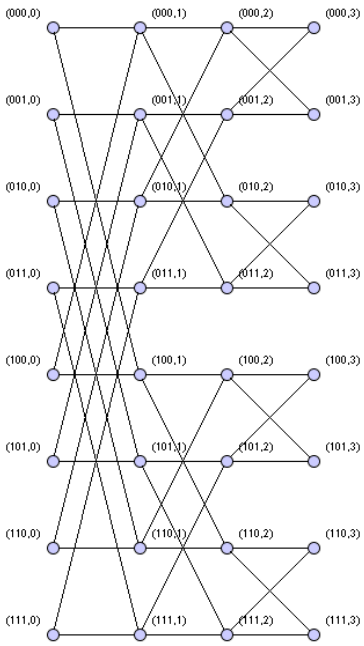
def. of *n*-dimensional butterfly (cntd)

...

Edges:

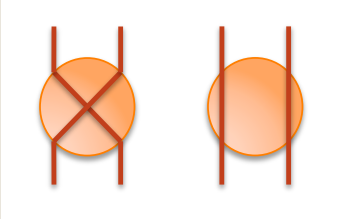
Two nodes *(w, i)* e *(w', i')* are linked by an edge iff $i'=i+1$ and either:

- $w=w'$ (*straight edge*) or
- *w* and *w'* differ in precisely the *i*-th bit (*cross edge*)



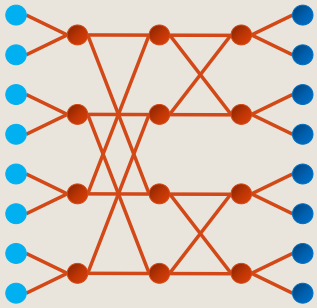
BUTTERFLY NETWORK (3)

- The nodes of the Butterfly are *crossbar switches*, i.e. switches with two input and two output values and can assume two states, *cross* and *bar*.
- Hence, the butterfly can be seen as a switching network connecting $2N$ ($N=2^n$) input units to $2N$ output units trough a $\log N+1$ layered network, having N nodes each.
- ...



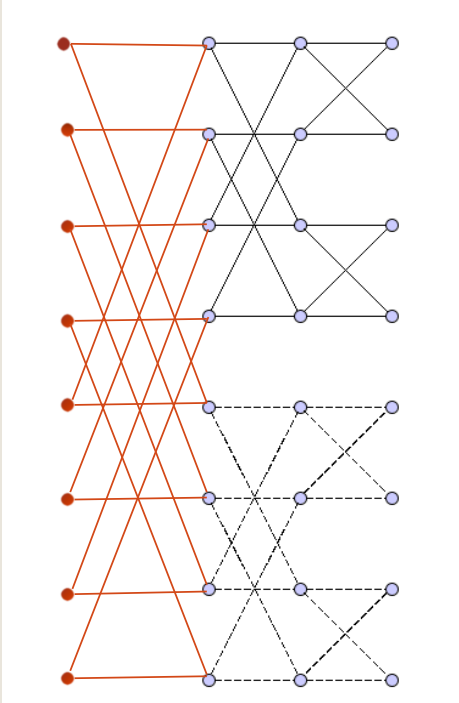
BUTTERFLY NETWORK (4)

- Input and output devices are usually processors and are often omitted in the graphical representations for the sake of simplicity.



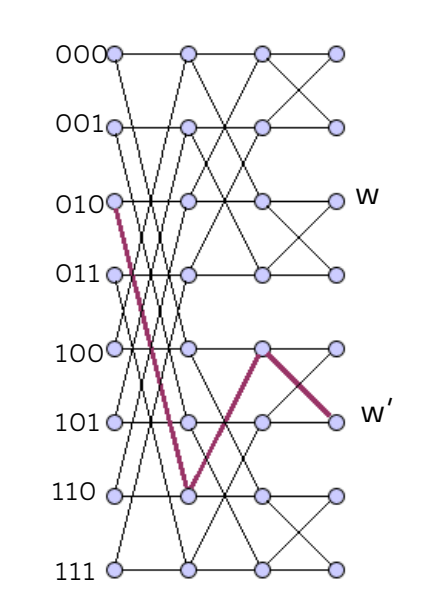
BUTTERFLY NETWORK (5)

The butterfly has a simple recursive structure:
the one n-dim. butterfly contains two (n-1)-dim. butterflies as subgraphs (just remove either the layer O nodes or the layer n nodes of the n-dim. butterfly to get two $(n-1)$-dimensional butterflies).



BUTTERFLY NETWORK (6)

For each pair of rows w and w', there exists a unique path of length n (known as greedy path) from (w,O) to (w', n);
this path passes through each layer exactly once, using a cross-edge from layer i to layer i+1 (i=0,…,n) iff w and w' differ in the i-th bit and using a straight-edge otherwise.



ROUTING ON THE BUTTERFLY (1)

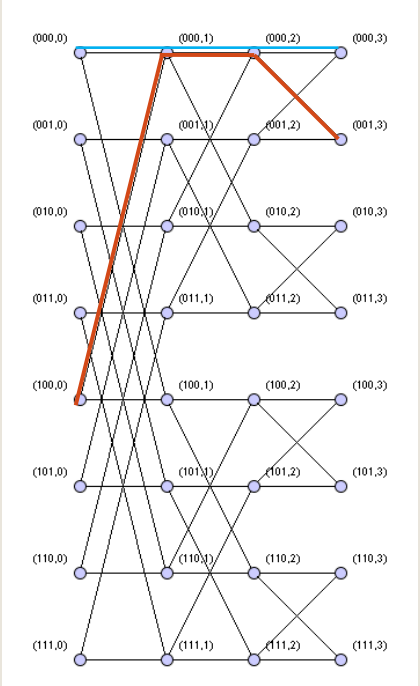
Problem of routing N packets from layer 0 to layer n in an n-dimensional butterfly:

- Each node $(u,0)$ on layer 0 of the butterfly contains a packet that is destined for node $(\pi(u), n)$ on layer n, where $\pi:[1, N] \rightarrow [1,N]$ is a permutation.
- In the greedy routing algorithm, each packet is constrained to follow its greedy path.
- When there is only one packet to route, the greedy algorithm performs very well.
- Trouble can arise when many packets have to be routed in parallel…

ROUTING ON THE BUTTERFLY (2)

Many greedy paths might pass through a single node or edge. Since only one packet can use the edge at a time, the other ones must be delayed before crossing the edge.

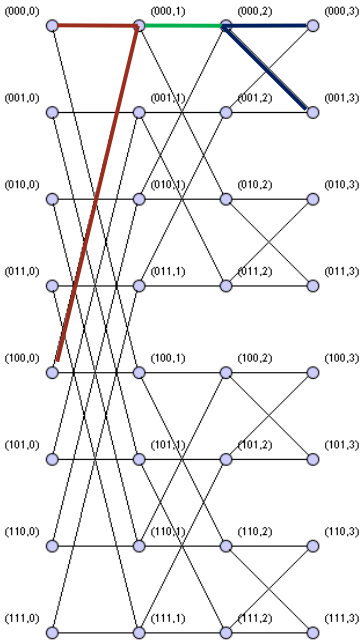
The butterfly is not able to route each permutation without delays, i.e. is a blocking network.
The arising congestion problem can be serious. In fact…



ROUTING ON THE BUTTERFLY (3)

Assume for simplicity that n is odd (but similar results hold when n is even), and consider edge e=((OO…O, (n-1)/2), (OO…O,(n+1)/2))

Node (OO…O, (n-1)/2) is the root of a complete binary tree extending to the left having $2^{(n-1)/2}$ leaves
Analogously to the right
…



ROUTING ON THE BUTTERFLY (4)

The permutation can be such that each greedy path from a leaf of the left tree arrives to a leaf of the right tree traverses e.

There are $2^{(n-1)/2} = \sqrt{N/2}$ possible such paths, and thus $2^{(n-1)/2} = \sqrt{N/2}$ packets may traverse e. So at least one of them may be delayed by $\sqrt{N/2} - 1$ steps.
It takes at least $n = \log N$ steps to traverse the whole network and to route a packet to its destination.
In this case, the greedy algorithm can take $\sqrt{N/2} + \log N - 1$ steps to route a permutation.

ROUTING ON THE BUTTERFLY (5)

In general:

Th. *Given any routing problem on an n-dimensional butterfly for which at most one packet starts at each layer-0 node and at most one packet is destined for each layer-n node, the greedy algorithm will route all the packets to their destinations in $O(\sqrt{N})$ steps.*

Proof. …

ROUTING ON THE BUTTERFLY (6)

Proof. For simplicity, assume that n is odd (but the case n even is similar).

Let e be any edge in layer i, $0<i\leq n$, and define n_i to be the number of greedy paths that traverse e

$n_i \leq 2^{i-1}$ (left tree) and, similarly, $n_i \leq 2^{n-i}$ (right tree) so $n_i \leq \min\{2^{i-1}, 2^{n-i}\}$

Any packet crossing e can be delayed by at most the other n_i-1 packets that want to cross the edge.

…

ROUTING ON THE BUTTERFLY (7)

…

As this packet traverses layers 1, 2, …, *n*, the total delay encountered can be at most:

$$\sum_{i=1}^{n}(n_i-1)=\sum_{i=1}^{(n+1)/2}(n_i-1)+\sum_{i=(n+3)/2}^{n}(n_i-1)\le\sum_{i=1}^{(n+1)/2}(2^{i-1}-1)+\sum_{i=(n+3)/2}^{n}(2^{n-i}-1)\le$$

recalling that $\sum_{j=0}^{k}2^j=2^{k+1}-1$

=(n+1)/2+1

$=\sum_{j=0}^{(n+1)/2-1}(2^j-1)$

$=\sum_{j=0}^{(n-3)/2}(2^j-1)$

$$\le 2^{(n+1)/2}+2^{(n-1)/2}-n=O(\sqrt{N})-n=O(\sqrt{N})$$ ■

ROUTING ON THE BUTTERFLY (8)

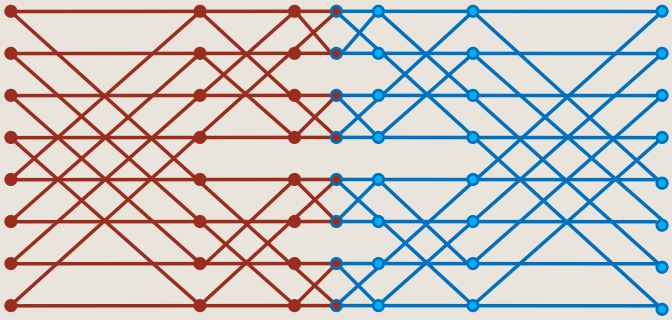
Despite the fact that the greedy routing algorithm performs poorly in the worst case, the greedy algorithm is very useful in practice.

For many useful classes of permutations, the greedy algorithm runs in *n* steps, which is optimal and, for most permutations, the greedy algorithm runs in $n + o(n)$ steps.

As a consequence, the greedy algorithm is widely used in practice.

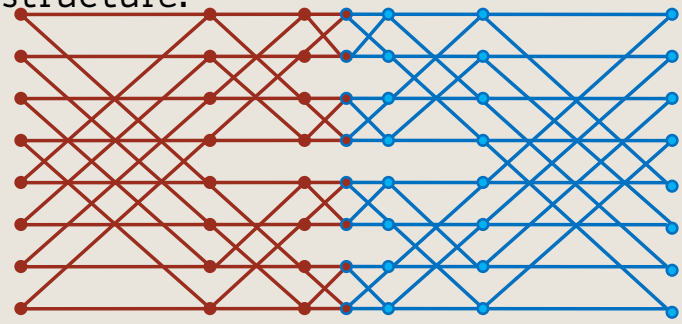
BENEŠ NETWORK (1)

A possibility to avoid a routing with delays is providing a non blocking topology.
Beneš network has this property.
It consists of two back-to-back butterflies…



BENEŠ NETWORK (2)

The n-dimensional Beneš network has $2n+1$ layers, each with 2^n nodes.
The first and last $n+1$ layers form an n-dimensional Butterfly (the middle layer is shared).
Not surprisingly, the Beneš network is very similar to the Butterfly, in terms of both its computational power and its network structure.



BENEŠ NETWORK (3)

The reason for defining the Beneš network is that it is an excellent example of a rearrangeable network.

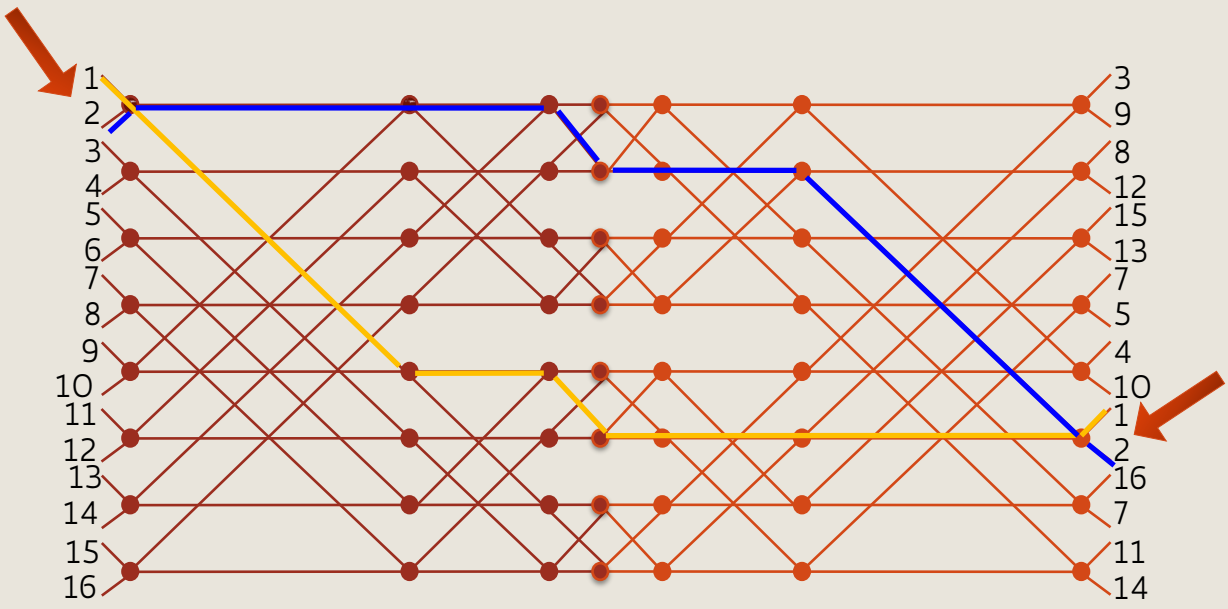
Def. A network with N inputs and N outputs is said to be rearrangeable if for any one-to-one mapping π of the inputs to the outputs (i.e. for any permutation), we can construct edge-disjoint paths in the network linking the i-th input to the $\pi(i)$-th output for $1 \le i \le N$.

BENEŠ NETWORK (4)

In the case of the n-dimensional Beneš network, we can have *two* inputs for each node at layer 0 and *two* outputs for each node at layer $2n$, and still connect every permutation of inputs to outputs with edge-disjoint paths.

Hence, in this case, # of inputs $= 2^{n+1}$.

BENEŠ NETWORK (5)



BENEŠ NETWORK (6)

It seems extraordinary that we can find edge-disjoint paths for any permutation. Nevertheless, the result is true, and it is even fairly easy to prove, as we show in the following:

Th. *Given any one-to-one mapping π of 2^{n+1} inputs to 2^{n+1} outputs on an n-dimensional Beneš network, there is a set of edge-disjoint paths from the inputs to the outputs connecting input i to output $\pi(i)$ for $1 \le i \le 2^{n+1}$.*

Proof. …

BENEŠ NETWORK (7)

PROOF OF THE REARRANGEABILITY OF THE BENEŠ NETWORK (CNTD)

Proof. By induction on *n*.

Basis: if *n=0*, the Beneš network consists of a single node (i.e. a single 2x2 switch) and the result is obvious.

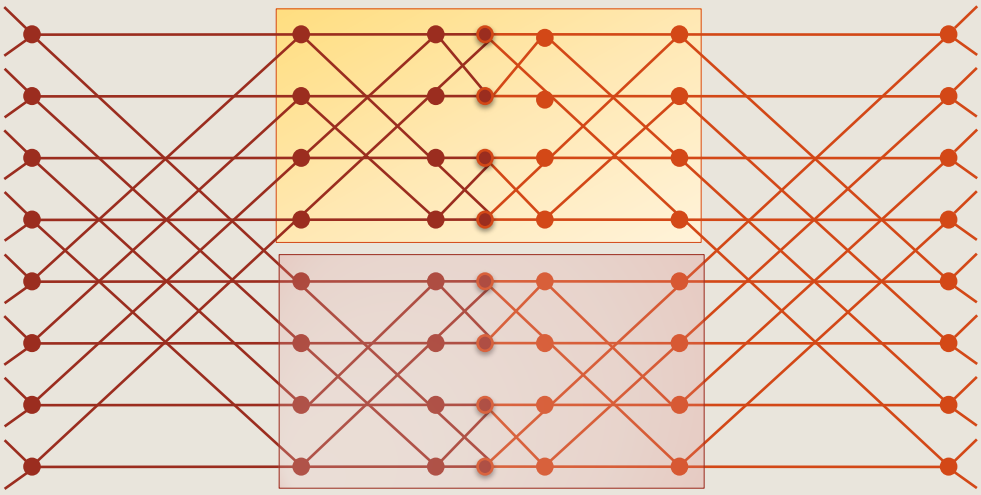
Induction: assume that the result is true for an *(n-1)*-dim. Beneš network

Key observation: the middle *2n-1* layers of an *n*-dim. Beneš network comprise two *(n-1)*-dim. Beneš networks:

…

BENEŠ NETWORK (8)

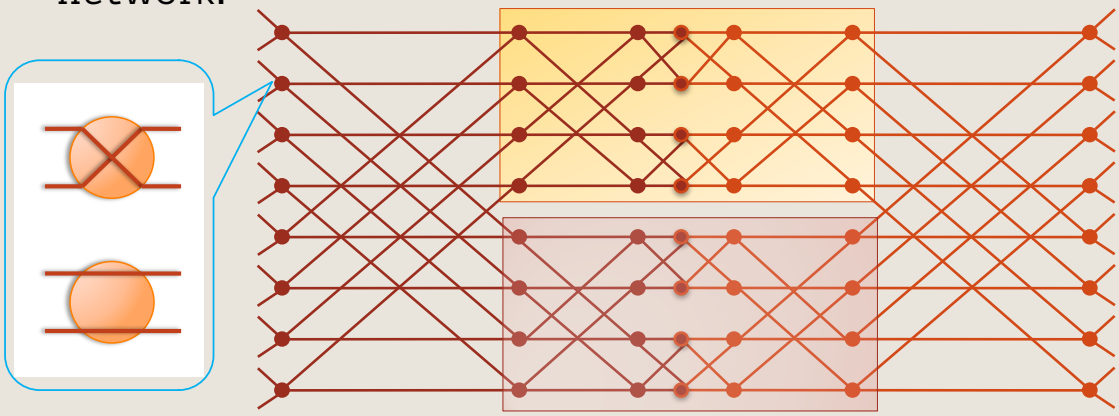
PROOF OF THE REARRANGEABILITY OF THE BENEŠ NETWORK (CNTD)



BENEŠ NETWORK (9)

PROOF OF THE REARRANGEABILITY OF THE BENEŠ NETWORK (CNTD)

Hence, for each path, it will be sufficient to decide whether it is to be routed through the upper sub-Beneš network or through the lower sub-Beneš network.

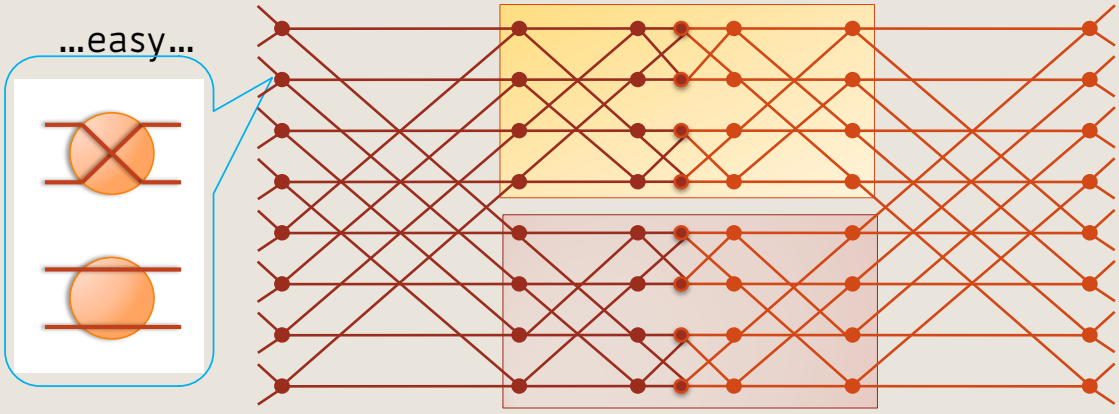


BENEŠ NETWORK (10)

PROOF OF THE REARRANGEABILITY OF THE BENEŠ NETWORK (CNTD)

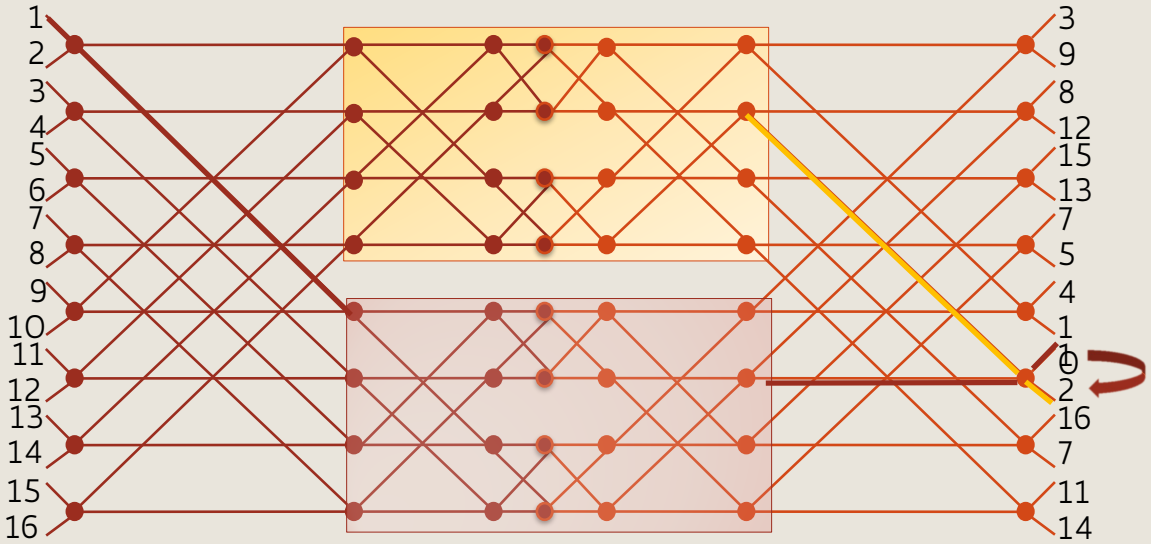
The only constraints we have to consider to decide whether paths use the upper or lower subnetworks are that paths from inputs $2i-1$ and $2i$ must use different subnetworks for $1 \leq i \leq 2n$, and that paths to outputs $2i-1$ and $2i$ must use different sub-networks.

...easy...



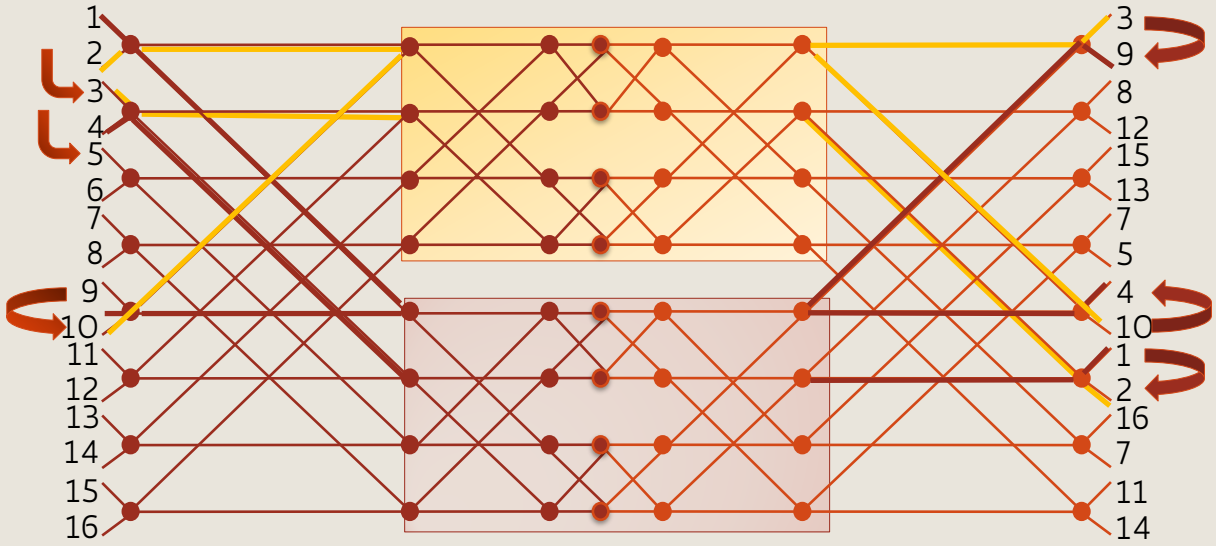
BENEŠ NETWORK (11)

PROOF OF THE REARRANGEABILITY OF THE BENEŠ NETWORK (CNTD)



BENEŠ NETWORK (12)

PROOF OF THE REARRANGEABILITY OF THE BENEŠ NETWORK (CNTD)



And so on…

BENEŠ NETWORK (13)
PROOF OF THE REARRANGEABILITY OF THE BENEŠ NETWORK (CNTD)

Summary of the steps:

- We start by routing the first path through the upper sub-network.
- We next satisfy the constraint generated at the output by routing the corresponding path through the lower sub-network.
- …

BENEŠ NETWORK (14)
PROOF OF THE REARRANGEABILITY OF THE BENEŠ NETWORK (CNTD)

- …
- We keep on going back and forth through the network, satisfying constraints at the inputs by routing through the upper sub-network and satisfying constraints at the outputs by routing through the lower sub-network.
- Eventually, we will close the loop by routing a path through the lower sub-network (in response to an output constraint) that shares an input switch with the first path that was routed.
- …

BENEŠ NETWORK (15)
PROOF OF THE REARRANGEABILITY OF THE BENEŠ NETWORK (CNTD)

- …
- If any additional paths needs to be routed, we con- tinue as before, starting over again with an arbitrary unrouted path.
- In this way, all paths can be assigned to the upper or lower sub-networks without conflict.

BENEŠ NETWORK (16)
PROOF OF THE REARRANGEABILITY OF THE BENEŠ NETWORK (CNTD)

- This algorithm is called looping algorithm.
- It is easy to see that all paths can be assigned to the upper or lower sub-networks without conflict:
- By construction, if we start going to the upper sub-network, we will arrive to the corresponding output in the upper sub-network and we will leave it to the lower sub-network, and so on.
- …

BENEŠ NETWORK (17)

PROOF OF THE REARRANGEABILITY OF THE BENEŠ NETWORK (CNTD)

- …
- For parity reason, when a loop is close, we will correctly arrive from the right sub-network.
- The remainder of the path routing and switch setting is handled by induction in the sub-networks.

■

BENEŠ NETWORK (18)

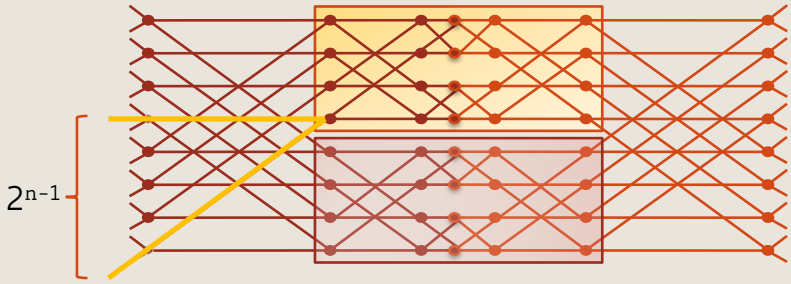
In the case that each layer 0 node of the *n*-dimensional Beneš network has just one input and each layer *2n* node has just one output, then the paths from the inputs to the outputs can be constructed so as to be node-disjoint (instead of only edge-disjoint):

…

BENEŠ NETWORK (19)

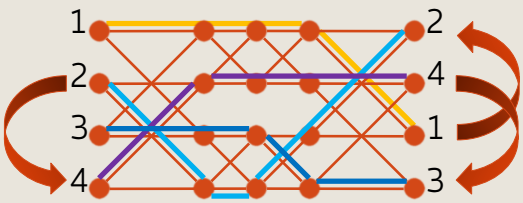
Th. *Given any one-to-one mapping of π of 2^n inputs to 2^n outputs in an n-dim. Beneš network, there is a set of node-disjoint paths from the inputs to the outputs connecting input i to output $\pi(i)$ for $1 \le i \le 2^n$.*

Proof. Identical to the previous one, but the paths needing to use different Beneš networks are now *i* and $i+2^{n-1}$, $1 \le i \le 2^{n-1}$ (and not *2i-1* and *2i*). ■



BENEŠ NETWORK (20)

- Example: n=2, hence 2^{n-1}=2



BENEŠ NETWORK (21)

Drawbacks of the looping algorithms (both versions):

- we do not know how to set the switches on-line. In other words, each switch needs to be told what to do by a **global control** that has knowledge of the permutation being routed
- there exist numerous methods for overcoming this difficulty (not studied here).
- every time a new permutation must be routed, $\Theta(N \log N)$ time is necessary to re-set switches.

MESH NETWORK

Another important and widely used interconnection topology is the mesh:

- For integers m and n, the $m \times n$ mesh $M_{m,n}$ has node set $\{1,2,\ldots,m\}\times\{1,2,\ldots,n\}$.
- The edges of $M_{m,n}$ connect nodes (i, j) and (i', j') just when $|i - i'| + |j - j'| = 1$.
- The path induced by the set of nodes $\{i\} \times \{1,2,\ldots,n\}$ (resp., the set $\{1,2,\ldots,m\} \times \{j\}$) is the i-th row (resp., the j-th column) of $M_{m,n}$

For the convenience of physical layout, mesh is the most used topology in Network-on-Chip design.

Routing algorithms on mesh: ☞ student lesson