Combinatorial Optimization and Applications in VLSI Design

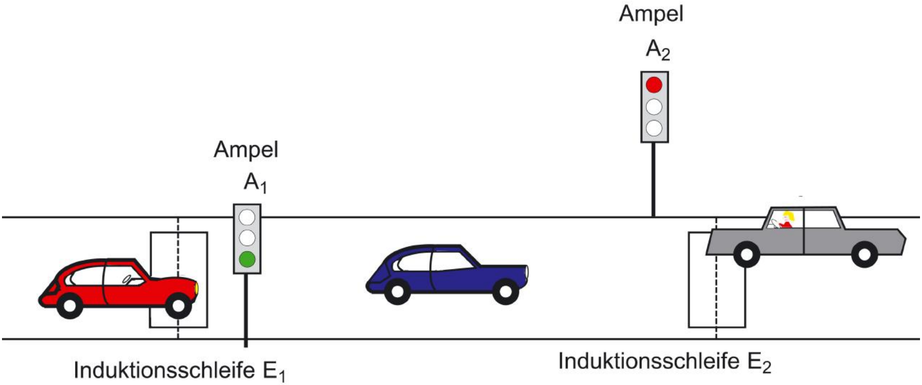
Jens Vygen

University of Bonn

Outline

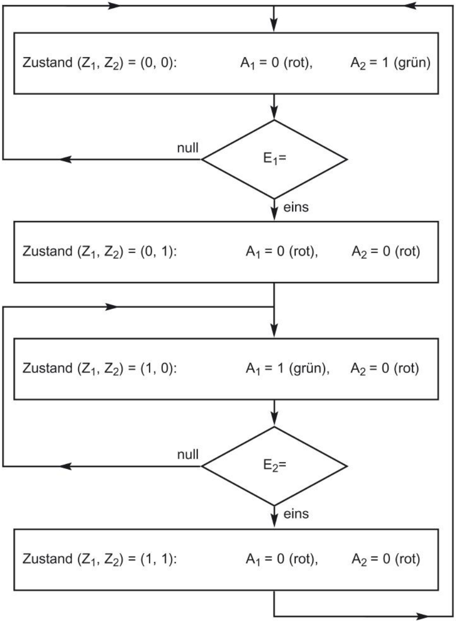
- Introduction
- Placement
- Routing
- Timing Optimization
- Clocktree Design

Example

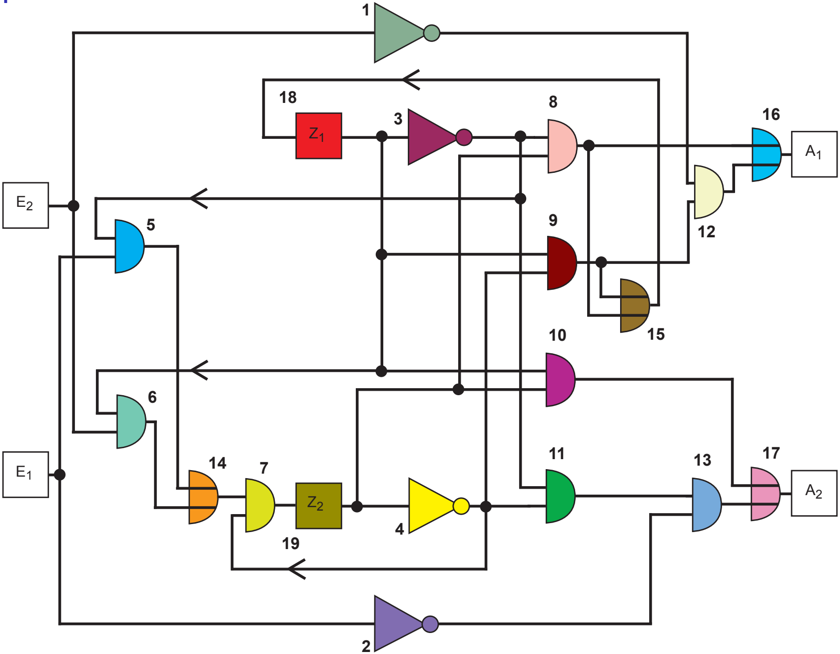


Ampel
A_2
Ampel
A_1
Induktionsschleife E_1
Induktionsschleife E_2

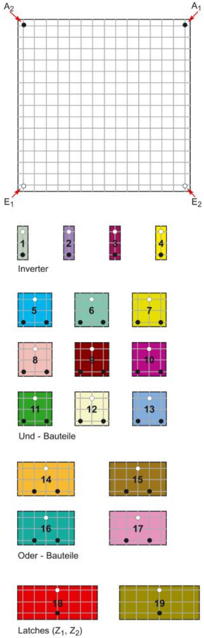
Example


Zustand (Z_1, Z_2) = (0, 0): A_1 = 0 (rot), A_2 = 1 (grün)
E_1=
null
eins
Zustand (Z_1, Z_2) = (0, 1): A_1 = 0 (rot), A_2 = 0 (rot)
Zustand (Z_1, Z_2) = (1, 0): A_1 = 1 (grün), A_2 = 0 (rot)
E_2=
null
eins
Zustand (Z_1, Z_2) = (1, 1): A_1 = 0 (rot), A_2 = 0 (rot)

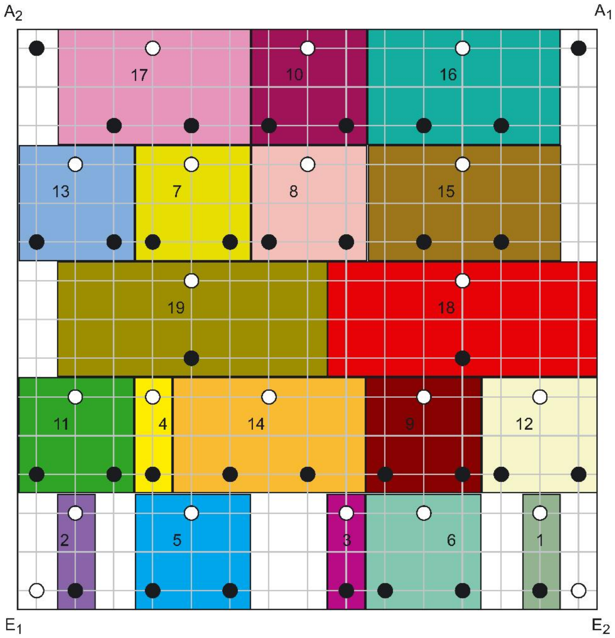
Example



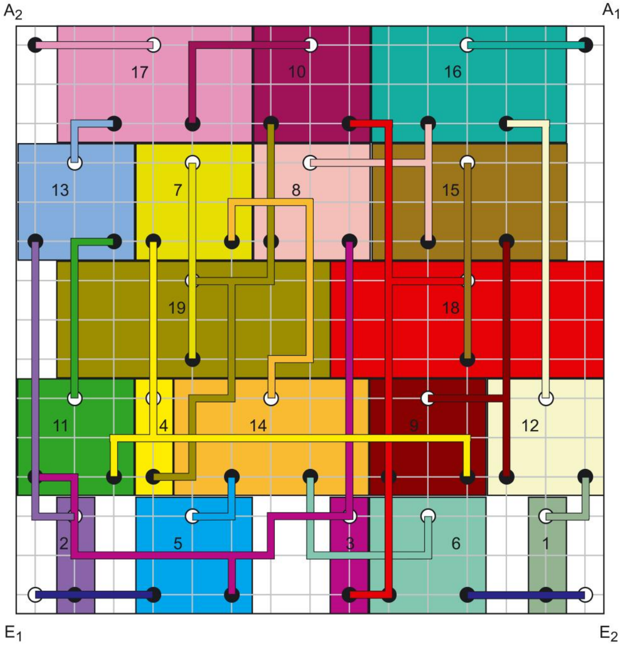
Example


A$_2$
A$_1$
E$_1$
E$_2$
1
2
3
4
Inverter
5
6
7
8
9
10
11
12
13
Und - Bauteile
14
15
16
17
Oder - Bauteile
18
19
Latches (Z$_1$, Z$_2$)

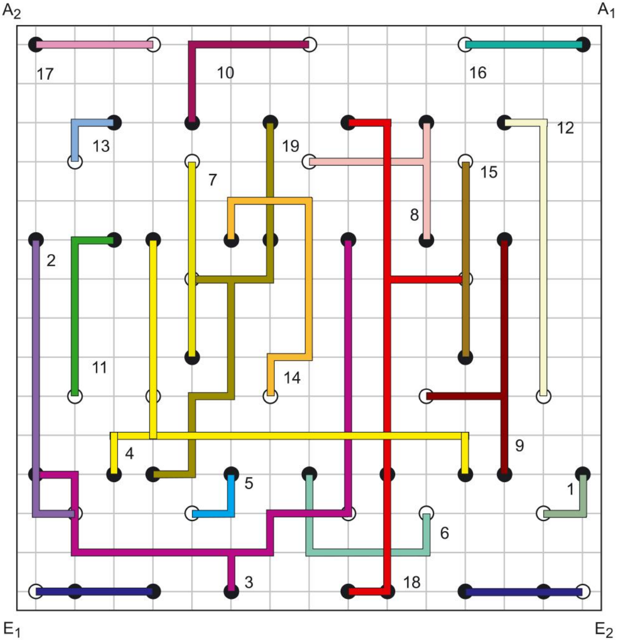
Example



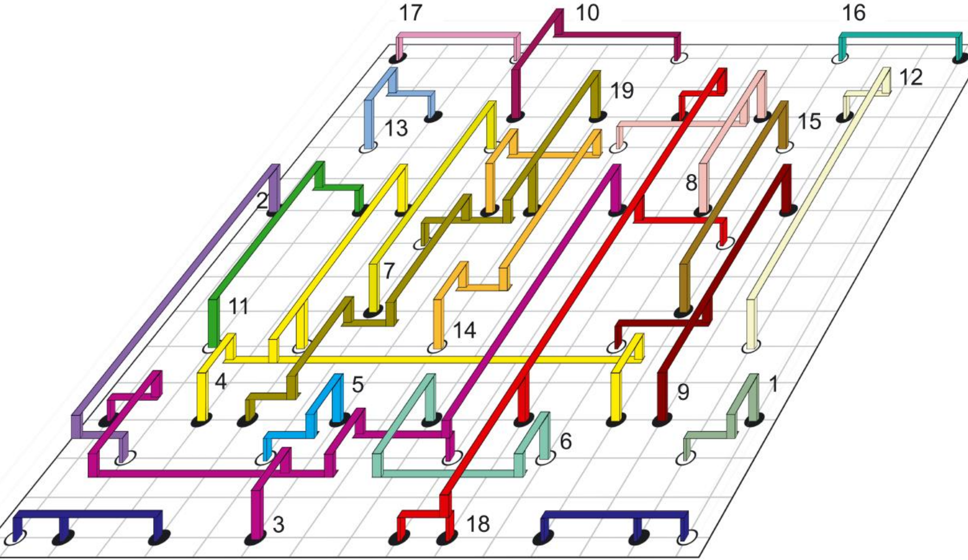
Example



Example



Example



VLSI design: overall task

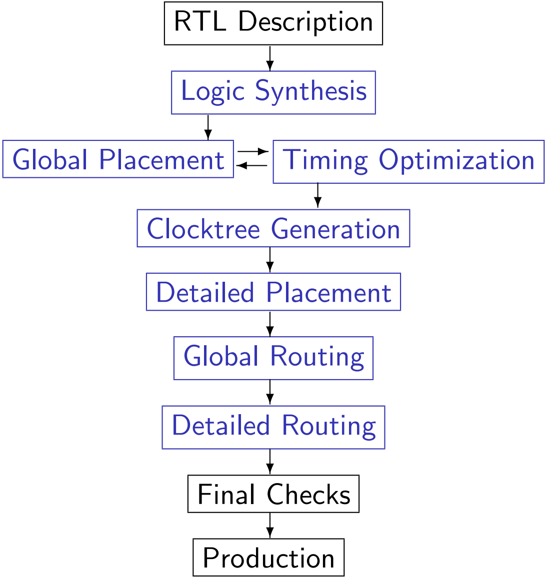
Given a netlist, with primary inputs, registers, primary outputs, complex logic cores, combinational logic, and constraints for placement, routing, and timing, the task is to

- compute an equivalent netlist, where registers and parts of the combinational logic can be replaced if the Boolean function $\Phi : \{0,1\}^I \rightarrow \{0,1\}^O$ representing the netlist does not change. I contains the primary inputs and output pins of registers and cores, and O contains the primary outputs and input pins of registers and cores.
- place the components of this netlist without overlaps on the chip area,
- and route all nets i.e. find node-disjoint Steiner trees, each connecting the pins of a net, in a given 3-dimensional grid graph.

Combinatorial optimization in VLSI design automation

- shortest paths
- network design, in particular Steiner trees
- maximum flows, discrete time-cost tradeoff problems
- transportation and minimum cost flows
- multicommodity flows, disjoint paths and trees
- minimum mean cycles, parametric shortest paths
- facility location
- ... and others: minimum spanning trees, knapsack problem, bin packing, traveling salesman problem, Huffman codes, ...
- ... also used: advanced data structures, computational geometry, nonlinear programming, parallelization ...

Design flow


RTL Description

Logic Synthesis

Global Placement

Timing Optimization

Clocktree Generation

Detailed Placement

Global Routing

Detailed Routing

Final Checks

Production

Main objectives

- minimize cycle time / meet timing constraints
  (all signal arrival times within prescribed time intervals)
- minimize power consumption
  (depending on transistor sizes, length and widths of wires, coupling, leakage)
- minimize cost
  (area, number of masks, yield, design effort)

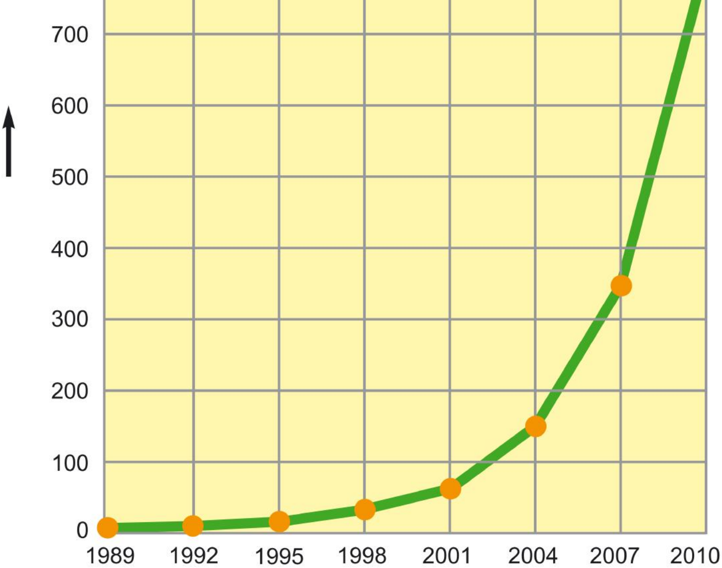
Main objectives in the design flow

Logic Synthesis	minimize area
Global Placement	minimize wirelength
Timing Optimization	minimize delay and power

Now, assuming an optimum Steiner tree for each net, all signals must arrive in time.

Clocktree Generation	minimize power subject to timing
Detailed Placement	minimize changes
Global Routing	minimize power subject to timing
Detailed Routing	minimize changes

Moore’s law

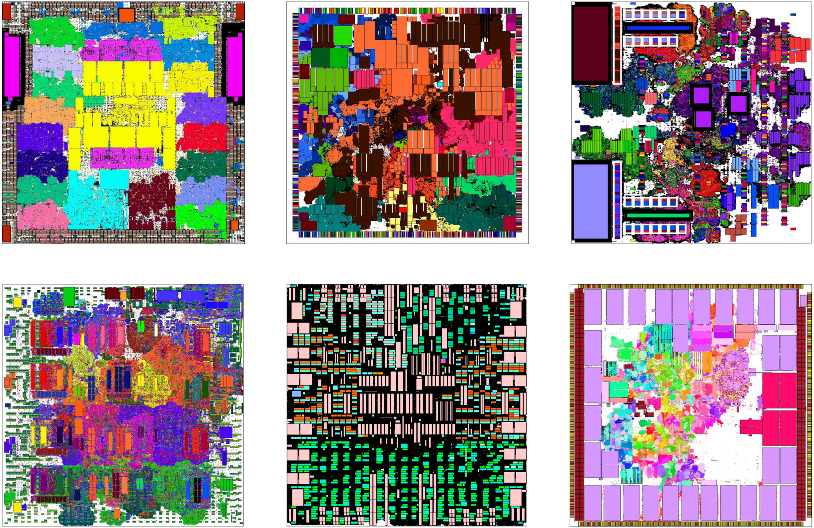


million transistors on a chip

The Bonn Tools

- are developped by the Research Institute for Discrete Mathematics at the University of Bonn,
- cover all major areas of layout and timing optimization,
- include libraries for combinatorial optimization, advanced, data structures, computational geometry, etc.,
- have more than one million lines of code in C and C++,
- are used by IBM and its customers for almost 20 years,
- are now also used by Magma Design Automation and its customers,
- have been used for the design of more than 1000 chips,
- including several complete microprocessor series,
- approximately hundred ASICs every year,
- and the most complex chips of major technology companies.

Some recent chips



The Bonn group

currently consists of:

Christoph Bartoschek, Florian Berger, Ulrich Brenner, Alexander von Dambrowski, Laura Geisen, Michael Gester, Stephan Held, Günther Hutzl, Fritz Jahns, Johannes Klauser, Alexander Kleff, Bernhard Korte, Immor Krupke, Jens Maßberg, Andreas Menge, Dirk Müller, Karsten Muuss, Christian Panten, Manuel Peelen, Sven Peyer, Dieter Rautenbach, Rüdiger Schmedding, Jan Schneider, Christian Schulte, Matthias Schwamborn, Markus Struzyna, Jens Vygen, Jürgen Werber

Thanks to all of them.

Thanks also to our cooperation partners at IBM and Magma

Introduction

Placement
- General theory
- Analytical placement
- Multisection
- Detailed Placement

Routing

Timing Optimization

Clocktree Design

How to measure interconnect length?

Let N be a finite set of points in the plane. Define net models:

- $\textsc{Steiner}(N)$ is the length of an optimum rectilinear Steiner tree for N.
- $\textsc{BB}(N) := \max_{p \in N} x(p) - \min_{p \in N} x(p) + \max_{p \in N} y(p) - \min_{p \in N} y(p)$.
- $\textsc{MST}(N)$ is the length of a minimum spanning tree for N, where edge weights are rectilinear distances.
- $\textsc{Clique}(N) := \dfrac{1}{|N|-1} \displaystyle\sum_{p,p' \in N} (|x(p) - x(p')| + |y(p) - y(p')|)$.
- $\textsc{Star}(N) := \min_{(x',y') \in \mathbb{R}^2} \displaystyle\sum_{p \in N} (|x(p) - x'| + |y(p) - y'|)$.

Worst case ratios of various net models

Entry (r, c) is $\sup \frac{c(N)}{r(N)}$ over all point sets N with $|N| = n$.

	BB	STEINER	MST	CLIQUE	STAR
BB	1	1	1	1	1
STEINER	$\frac{n-1}{\lceil\sqrt{n}\rceil + \left\lceil \frac{n}{\lceil\sqrt{n}\rceil} \right\rceil - 2}$ $\cdots$ $\frac{\lceil\sqrt{n-2}\rceil}{2} + \frac{3}{4}$	1	1	$\begin{cases} \frac{9}{8} & (n = 4) \\ 1 & (n \neq 4) \end{cases}$	1
MST	$\left\lfloor \frac{\sqrt{2n-1}+1}{2} \right\rfloor$ $\cdots$ $\frac{\sqrt{n}}{\sqrt{2}} + \frac{3}{2}$	$\frac{3}{2}$	1	$1 + \Theta\left(\frac{1}{n}\right)$ $\cdots$ $\frac{3}{2}$	$\begin{cases} \frac{4}{3} & (n = 3) \\ \frac{3}{2} & (n = 4) \\ \frac{6}{5} & (n = 5) \\ 1 & (n > 5) \end{cases}$
CLIQUE	$\frac{\lceil \frac{n}{2} \rceil \lfloor \frac{n}{2} \rfloor}{n-1}$	$\frac{\lceil \frac{n}{2} \rceil \lfloor \frac{n}{2} \rfloor}{n-1}$	$\frac{\lceil \frac{n}{2} \rceil \lfloor \frac{n}{2} \rfloor}{n-1}$	1	1
STAR	$\lfloor \frac{n}{2} \rfloor$	$\lfloor \frac{n}{2} \rfloor$	$\lfloor \frac{n}{2} \rfloor$	$\frac{n-1}{\lceil \frac{n}{2} \rceil}$	1

(Hwang [1976], Brenner and Vygen [2001], Rautenbach [2004])

Net models in placement

- STEINER is best, but *NP*-hard to compute
- all others can be computed in $O(n)$ time (BB, STAR) or in $O(n \log n)$ time (MST, CLIQUE).
- in quadratic placement (see below), CLIQUE and STAR are used
- BB is often used as a simple measure. As most nets have few pins, this is not too bad.

Clique is the best topology-independent net model

Theorem

For $n \geq 2$*, a connected graph* G *with* $\{1,\dots,n\} \subseteq V(G)$*,*
$c : E(G) \to \mathbb{R}_{>0}$*, and* $p : \{1,\dots,n\} \to \mathbb{R}^2$ *let*
$\mathcal{M}_{(G,c)}(p) :=$

$$\min\left\{ \sum_{e=\{v,w\}\in E(G)} c(e)||p(v)-p(w)||_1 \;\middle|\; p : V(G)\setminus\{1,\dots,n\} \to \mathbb{R}^2 \right\}.$$

Then the ratio of supremum and infimum of

$$\left\{ \mathcal{M}_{(G,c)}(p) \middle| p : \{1,\dots,n\} \to \mathbb{R}^2,\ \textsc{steiner}(\{p(1),\dots,p(n)\}) = 1 \right\}$$

is minimum for the complete graph K_n *with unit weights.*

(Brenner, Vygen [2001])

Placement: simplified problem formulation

Input:

- a rectangular chip area, and a set of rectangular blockages
- a finite set C of (rectangular) cells
- a finite set P of pins, and a partition $\mathcal{N}$ of P into nets
- a weight $w(N) > 0$ for each net N
- an assignment $\gamma : P \to C \cup \{\Box\}$ of the pins to cells
  [pins p with $\gamma(p) = \Box$ are fixed; we set $x(\Box) := y(\Box) := 0$]
- offsets $x(p), y(p) \in \mathbb{R}$ of each pin p

Task:

Find a position $(x(c), y(c)) \in \mathbb{R}^2$ of each cell c such that

- each cell is contained in the chip area,
- no cell overlaps with another cell or a blockage,

and the weighted netlength

$$\sum_{N \in \mathcal{N}} w(N)\,\mathrm{BB}\left(\{(x(\gamma(p)) + x(p), y(\gamma(p)) + y(p)) : p \in N\}\right)$$

is minimum.

Why minimize netlength?

- Netlength is a good estimate for power consumption
- Short nets have small delay (net weights for critical nets!)
- Nets have to be packed in routing, and long nets take more resources (but we must also avoid local congestion!)
- Experience shows: a good algorithm for the simplified placement problem can be extended to a good algorithm for real placement problems.
- Bounding box netlentgh is the main measure in benchmarks
- It's simple. But not easy...

Special case: Quadratic Assignment Problem (QAP)

Instance: A graph G. Weights $w : E(G) \to \mathbb{R}_+$. A set U with $|U| \geq |V(G)|$. Distances $d(\{u, v\}) \geq 0$ for all $u, v \in U$. Weights $c : V(G) \times U \to \mathbb{R}_+$.

Task: Find an injective mapping $f : V(G) \to U$ such that

$$\sum_{e=\{x,y\}\in E(G)} w(e)d(\{f(x), f(y)\}) + \sum_{x\in V(G),u\in U} c(x, u)d(\{f(x), u\})$$

is minimum.

Theorem

Unless $P = NP$ there is no constant-factor approximation algorithm for the special case of the Quadratic Assignment Problem *where $w(e) = 1$ for all $e \in E(G)$, c is identically zero, U is a finite subset of $\mathbb{Z}$ and $d(\{u, v\}) = |u - v|$ for all $u, v \in U$.*

(Queyranne [1986])

Proof of non-approximability

- $\textsc{Bin-Packing}$ is strongly *NP*-hard. More precisely, it is *NP*-hard to decide whether for given $n \in \mathbb{N}$ and $s_1, \ldots, s_{4n}, B \in \{1, \ldots, 10^{10}n^4\}$ there is a mapping $p : \{1, \ldots, 4n\} \to \{1, \ldots, n\}$ with $\sum_{i \in p^{-1}(j)} s_i \le B$ for $j = 1, \ldots, n$.
- Define an instance of QAP by $V(G) = \{1, \ldots, \sum_{i=1}^{n} s_i\}$; $E(G) := \{\{x, x+1\} : x \in V(G) \setminus \{\sum_{i=1}^{j} s_i : j = 1, \ldots, n\}\}$; $U := \{(kn+1)Bj + z : z = 1, \ldots, B,\ j = 1, \ldots, n\}$.
- If there exists a mapping p as above, there is a placement $f : V(G) \to U$ defined by $f(\sum_{i=1}^{j-1} s_i + z) := (kn+1)Bp(j) + z$ for $j = 1, \ldots, 4n$ and $z = 1, \ldots, s_j$, such that $\sum_{e=\{x,y\} \in E(G)} |f(x) - f(y)| = |E(G)|$.
- Otherwise, for any injective mapping $f : V(G) \to U$ there is an edge $\{x, y\} \in E(G)$ with $|f(x) - f(y)| \ge (kn+1)B + 1 - \max_{i=1}^{4n} s_i \ge knB > k|E(G)|$.
- Hence a k-factor approximation algorithm for such instances of QAP can distinguish between these two cases. $\square$

Positive results

There are only few, and these are not very useful.

Special case: OPTIMUM LINEAR ARRANGEMENT PROBLEM
Given a graph G with $n := |V(G)|$, we ask for a bijection $f : V(G) \to \{1, \ldots, n\}$ minimizing $\sum_{\{x,y\} \in E(G)} |f(x) - f(y)|$.

- Even this problem is *NP*-hard. (Garey, Johnson [1976])
- There is an $O(\sqrt{\log n} \log \log n)$-factor approximation algorithm. (Charikar, Hajiaghayi, Karloff, Rao [2006])
- However, the problem is not known to be *MAXSNP*-hard!

Polylogarithmic approximation algorithm also for some two-dimensional problems
(Even, Naor, Rao, Schieber [2000], Even, Guha, Schieber [2003], Vempala [1998])

Placement approaches in practice

- simulated annealing: start with any placement and try to improve it
  (mostly used in the 80s)
- min-cut: successive bisection, with simple exchange heuristics
  (mostly used in the 90s)
- analytical placement: minimize either linear or quadratic netlength estimate, then work towards disjointness
  (dominant strategy today)

Here we discuss analytical placement only.

Minimizing weighted netlength

$$\min \sum_{N \in \mathcal{N}} w(N)(X_N + Y_N)$$

where

$$X_N := \max\{x(\gamma(p))+x(p) : p \in N\} - \min\{x(\gamma(p))+x(p) : p \in N\}$$

$$Y_N := \max\{y(\gamma(p))+y(p) : p \in N\} - \min\{y(\gamma(p))+y(p) : p \in N\}$$

Net weights $w(N)$ reflect timing criticality (slack, Lagrange multipliers).

Minimizing weighted netlength

Equivalent formulation:

$$\min \sum_{N \in \mathcal{N}} w(N)(r(N) - l(N) + t(N) - b(N))$$

subject to

$$\begin{aligned} l(N) \leq x(\gamma(p)) + x(p) \leq r(N) \qquad (p \in N \in \mathcal{N}) \\ b(N) \leq y(\gamma(p)) + y(p) \leq t(N) \qquad (p \in N \in \mathcal{N}) \end{aligned}$$

This is the dual of a minimum cost flow problem.

Quadratic placement (QP)

$$\min \sum_{N \in \mathcal{N}} \frac{w(N)}{|N|-1} \sum_{p,q \in N} (X_{p,q} + Y_{p,q})$$

where

$$X_{p,q} := |x(\gamma(p)) + x(p) - x(\gamma(q)) - x(q)|^2$$

and

$$Y_{p,q} := |y(\gamma(p)) + y(p) - y(\gamma(q)) - y(q)|^2$$

The placement where this minimum is attained is unique if the netlist is connected. It is called quadratic placement.

Why using Quadratic placement?

- QP can be solved very fast (conjugate gradient method)
- Delay along unbuffered wires grows quadratically with length
- QP gives a lot of information on relative positions
- QP is stable:

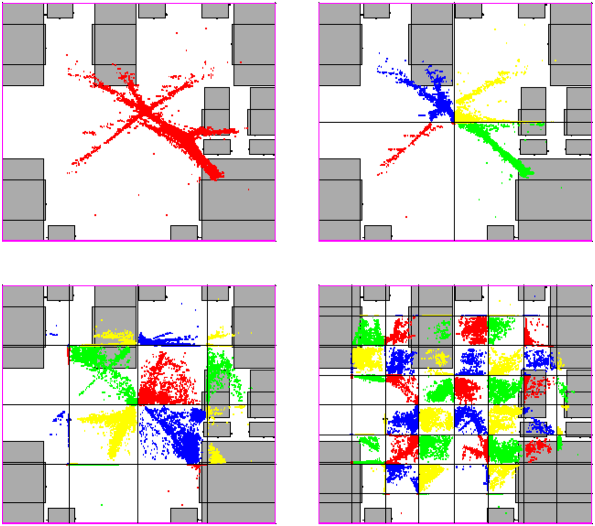
Theorem

Small netlist changes imply small changes of QP solution.
In contrast, min-cut and local search are unstable.

(Vygen [2002])

Global Placement by successive partitioning

Remove overlaps by successive quadrisection.



Successively distribute the set C of cells to regions R.

Quadratic placement with an array of regions

$$\min \sum_{N \in \mathcal{N}} \frac{w(N)}{|N|-1} \sum_{p,q \in N} (X_{p,q} + Y_{p,q})$$

where $X_{p,q}$ is

- $|x(\gamma(p)) + x(p) - x(\gamma(q)) - x(q)|^2$ if $\gamma(p)$ and $\gamma(q)$ are cells assigned to regions in the same column.
- $|x(\gamma(p)) + x(p) - b|^2 + |x(\gamma(q)) + x(q) - a'|^2$ if $\gamma(p)$ is a cell assigned to a region with x-range $[a, b]$, $\gamma(q)$ is a cell assigned to a region with x-range $[a', b']$ and $b \le a'$.
- $|x(\gamma(p)) + x(p) - v|^2$ if $\gamma(p)$ is a cell assigned to a region with x-range $[a, b]$, q is fixed, and $v = \max\{a, \min\{b, x(q)\}\}$.
- 0 if p and q are both fixed.

$Y_{p,q}$ is defined analogously, but with respect to y-coordinates, and with rows playing the role of columns.
(Vygen [1997])

Replace large cliques by stars

The running time of the conjugate gradient method depends on

- the number of variables (cells) and
- the number of connected pin pairs

Therefore, one should replace large cliques, i.e. sets of pins belonging to cells in the same column (row) by a stars: introduce a new variable (representing the center of the star) and connect it to each of the pins.

With appropriate weights, this does not change the result.

A single partitioning step

Let C be a set of cells, each with a size,
and R a set of (sub)regions, each with a capacity.

Task: Find an assignment $f : C \to R$
meeting the capacity constraints

$$\sum_{c \in C : f(c) = r} \mathit{size}(c) \le \mathit{cap}(r) \text{ for all } r \in R$$

such that the total movement

$$\sum_{c \in C} d(c, f(c))$$

is minimum.
Here d denotes, e.g., the ℓ_1-distance.

Fractional relaxation: Hitchcock (transportation) problem

Find $g : C \times R \rightarrow \mathbb{R}_{\geq 0}$

with

$$\sum_{r \in R} g(c, r) = size(c) \text{ for all } c \in C$$

and

$$\sum_{c \in C} g(c, r) \leq cap(r) \text{ for all } r \in R$$

such that

$$\sum_{c \in C} \sum_{r \in R} g(c, r) d(c, r)$$

is minimum.

Note: $|R| \ll |C|$.

Solving the fractional relaxation is sufficient

Proposition

From any optimum solution we can obtain another one in $O(|C||R|^2)$ *time that is integral up to* $|R|-1$ *cells.*

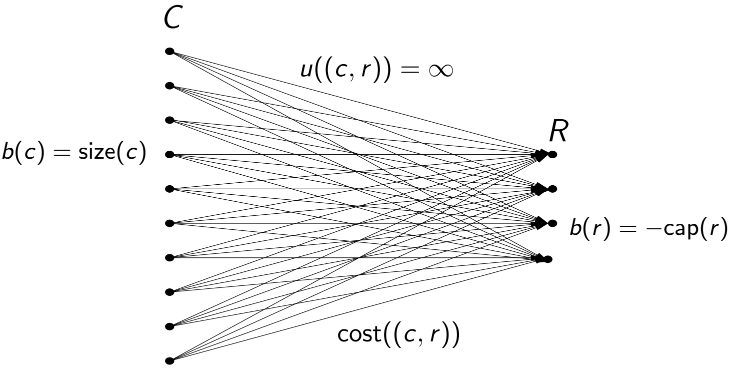
Proof.

Define G by $V(G) := R$ and $E(G) := \{\{r, r'\} : c \in C, g(r) > 0, g(r') > 0, g(r'') = 0 \text{ for } r'' \in \{1, \ldots, \max\{r, r'\}\} \setminus \{r, r'\}\}$
While G contains a cycle, move flow around. ☐

(Vygen [1996,2005])

Hitchcock Problem

Let G be the digraph with $V(G) := C \dot{\cup} R$ and $E(G) := C \times R$.
Let $b(c) := \text{size}(c)$ for $c \in C$ and $b(r) := -\text{cap}(r)$ for $r \in R$.
Let $\text{cost}((c,r)) := \frac{d((c,r))}{\text{size}(c)}$ $(c \in C,\ r \in R)$



Task: Find an uncapacitated b-flow in G of minimum cost.

Algorithms for the Hitchcock problem

Let $n := |C|$ and $k := |R|$. We assume $n \geq k$.

- $O(n \log n(n \log n + kn))$: general transshipment algorithm (Orlin [1993])
- $O(n\,f(k))$ with exponential functions f, inefficient already for very small k: Dyer [1984], Zemel [1984], Tokuyama, Nakano [1991], Meggido, Tamir [1993], Matsui [1993]
- First application to VLSI placement and very efficient $O(n)$-algorithm for $k = 4$ and $cost((c, r_1)) + cost((c, r_3)) = cost((c, r_2)) + cost((c, r_4))$ for all $c \in C$: Vygen [1996,2005]
- $O(nk^2 \log^2 n)$ (fastest previous strongly polynomial algorithm for unbalanced instances): Tokuyama, Nakano [1992, 1995]
- New algorithm: $O(nk^2(\log n + k \log k))$: Brenner [2005]

Residual graph

Given $f : E(G) \to \mathbb{R}_{\geq 0}$, we define the residual graph G_f as follows.

- $V(G_f) := V(G) \cup \{t\}$.
- $E(G_f)$ contains all arcs $e \in E(G)$, with $u_f(e) := \infty$.
- For each arc $(c, r) \in E(G)$ with $f((c, r)) > 0$, $E(G_f)$ contains the backward arc (r, c) with $u_f((r, c)) := f((c, r))$.
- For each $r \in R$ with $f(\delta^-(r)) + b(r) < 0$, $E(G_f)$ contains the arc (r, t) with $u_f((r, t)) := -b(r) - f(\delta^-(r))$.

Successive shortest path algorithm

Input: An instance (G, b, cost) of the HITCHCOCK PROBLEM.
Output: A minimum cost b-flow f in G.

```
① $f(e) := 0$ for $e \in E(G)$.
② Let $C = \{c_1, \ldots, c_n\}$.
③ For $i := 1$ to $n$:
        While $f(\delta^+(c_i)) < b(c_i)$:
            Find a shortest $c_i$-$t$-path $P$ in $G_f$.
            $\gamma := \min\left\{\min_{e \in E(P)} u_f(e), b(c_i) - f(\delta^+(c_i))\right\}$.
            Augment $f$ along $(s, c_i)$ and $P$ by $\gamma$.
```

Idea: Replace each phase (iteration of the outer loop) by one min-cost flow computation in a graph whose size depends on k only.

Lemma on almost integral solutions

Definition: Let f be a solution of the HITCHCOCK PROBLEM.
For $c \in C$ let $\tau_f(c) := |\{r \in R : f((c,r)) > 0\}|$.
Let $F_f := \{c \in C : \tau_f(c) > 1\}$.

Lemma
Given an instance $(G, b, u, cost)$ of the HITCHCOCK PROBLEM, *an optimum solution f, and the set F_f, we can transform f in $O(k \cdot \sum_{c \in F_f} \tau_f(c))$ time into an optimum solution g such that:*

- $|F_g| \leq k - 1$, *and*
- $\sum\limits_{c \in F_g} \tau_g(c) \leq 2k - 2$.

General strategy

- Sort the cells such that $\text{size}(c_1) > \text{size}(c_2) > \cdots > \text{size}(c_n)$.
- We will show: In each phase we have to change the flow only on $O(k^2)$ arcs.

Notation

- Let f_{i-1} be the flow at the beginning of phase i.
- For $r \in R$ let $M_r^i := \{c \in C : f_{i-1}((c,r)) = \mathrm{size}(c)\}$.
- If $M_r^i \neq \emptyset$ for $r \in R$, then choose for each $q \in R \setminus \{r\}$ an arbitrary $c_{r,q}^i \in M_r^i$ with

$$\mathrm{cost}((c_{r,q}^i, q)) - \mathrm{cost}((c_{r,q}^i, r)) =$$

$$\min\left\{\mathrm{cost}((c', q)) - \mathrm{cost}((c', r)) : c' \in M_r^i\right\}.$$

The subgraph G_i

$$\begin{aligned} V(G_i) \; := \; & R \cup \{t\} \\ & \cup \{c_i\} \cup F_{f_{i-1}} \\ & \cup \{c^i_{r,q} : r, q \in R,\ r \neq q,\ M^i_r \neq \emptyset\} \end{aligned}$$

$$\begin{aligned} E(G_i) \; := \; & (R \times \{t\}) \\ & \cup (\{c_i\} \times R) \\ & \cup (F_{f_{i-1}} \times R) \\ & \cup \{(c^i_{r,q}, r) : r, q \in R,\ M^i_r \neq \emptyset\} \\ & \cup \{(c^i_{r,q}, q) : r, q \in R,\ M^i_r \neq \emptyset\} \end{aligned}$$

$\Rightarrow$ G_i has $O(k^2)$ vertices and $O(k^2)$ arcs.

Main lemma

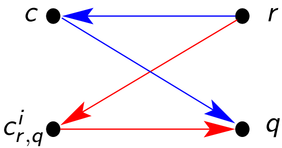
In phase i we can choose augmenting paths such that there are no two subsequent arcs $(r, c), (c, q)$ with $c \in C \setminus (F_{f_{i-1}} \cup \{c^i_{r,q}, c^i_{q,r}\})$.

Proof (Sketch).

Consider a sequence of shortest augmenting paths in $G_{f_{i-1}}$.
Consider the first path that contains arcs (r, c), (c, q) with $c \in C \setminus (F_{f_{i-1}} \cup \{c^i_{r,q}, c^i_{q,r}\})$.

Then $c \in M^i_p$ for some $p \in R$.

Case 1: $p = r$. Replace c by $c^i_{r,q}$ in the augmenting path.



cost of new subpath

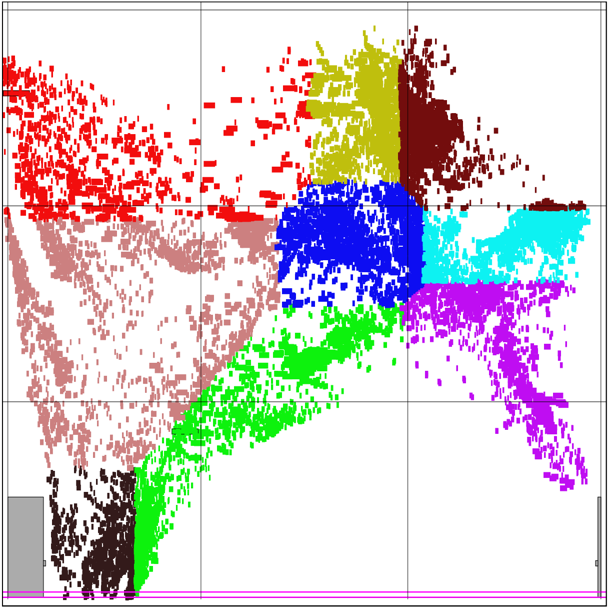
$$\begin{aligned} &= \text{cost}((c^i_{r,q}, q) - \text{cost}((c^i_{r,q}, r)) \\ &\leq \text{cost}((c, q)) - \text{cost}((c, r)) \\ &= \text{cost of old subpath} \end{aligned}$$

Case 2: $p \neq r$. Replace (c, q) by $(c, p, c^i_{p,q}, q)$ in the augmenting path. □

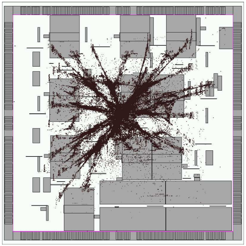
Main proof

- After phase i, we adjust the flow f_i such that $|F_{f_i}| \leq k-1$ and $\sum_{c \in F_{f_i}} \tau_{f_i}(c) \leq 2k-2$.
- Two more modifications:
    - Replace each $c^i_{r,q}$ (with its two incident arcs) by one uncapacitated arc from r to q. $\Rightarrow$ Only $O(k)$ vertices remain.
    - Only $O(k)$ arcs (entering the elements of $F_{f_{i-1}}$) have finite capacity. Replace each of them equivalently by two uncapacitated arcs. $\Rightarrow$ All arcs are uncapacitated.
- Thus, if G_i is given, a phase can be computed in $O(k \log k(k^2 + k \log k)) = O(k^3 \log k)$ time (Orlin[1993])
- By storing the sets M^i_r in heaps, G_i can be computed from G_{i-1} in $O(k^2 \log n)$ time.
- In each phase: $O(k^2)$ insert and remove operations suffice.
- Time to adjust the flow after a phase: $O(k^3)$.
- Total running time: $O(nk^2(\log n + k \log k))$. $\square$

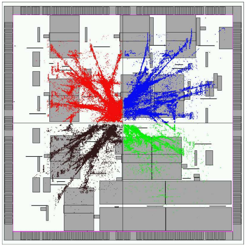
Multisection example



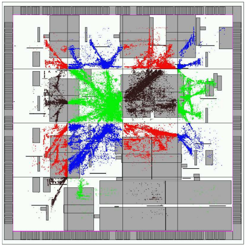
QP and quadrisection in BonnPlace



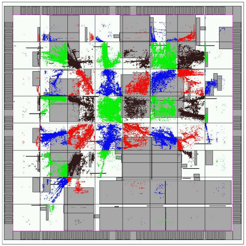
Level 0



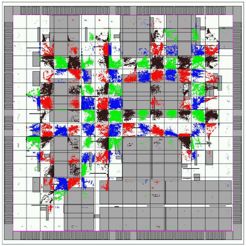
Level 1



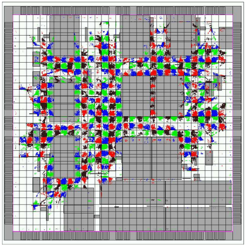
Level 2



Level 3



Level 4



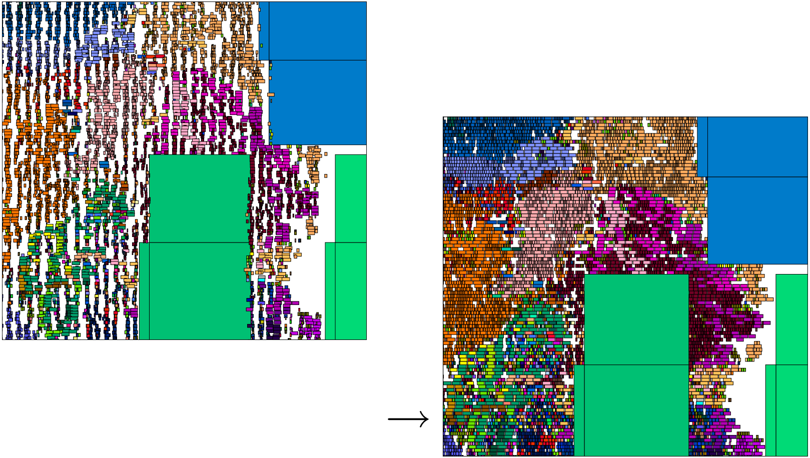
Level 5

Further components of BonnPlace

- Repartitioning
- Congestion-driven placement
- Macro placement

Detailed placement

After global placement we have an optimized but illegal placement.



We want to legalize it without changing it too much.

The legalization problem

Input:

- a rectangular chip area
- a set of rectangular blockages
- a set C of rectangular cells with unit height
- a width $w(c)$ and a position $(x(c), y(c)) \in \mathbb{R}^2$ of each cell $c \in C$.

Task:

Find new positions $(x'(c), y'(c)) \in \mathbb{Z}^2$ of the cells such that

- each cell is contained in the chip area,
- no two cells overlap,
- no cell overlaps with any blockage,

and $\sum_{c \in C} ((x(c) - x'(c))^2 + (y(c) - y'(c))^2)$ is minimum.

The problem is *NP*-hard.

Three-step approach:

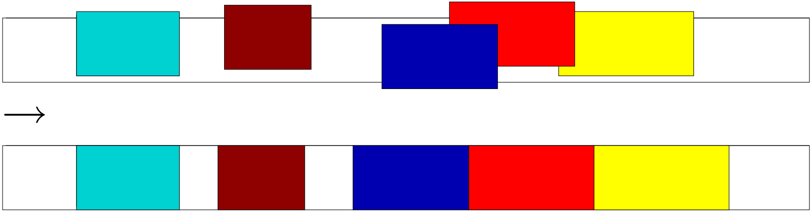
A zone is a maximal part of a row that is completely blocked or completely free.

Step 1: Make sure that no zone contains more cells than fit into it.

Step 2: Place the cells legally within their zones, keeping their horizontal order.

Step 3: Postoptimzation heuristics

Step 2: legalizing within zones



An optimal placement of n rectangles in a row in a given order can be found in

- $O(n \log n)$ time (for linear movement)
- $O(n)$ time (for quadratic movement)

(Garey, Tarjan, Wilfong [1988], Brenner, Vygen [2004])

Algorithm for a single zone

Input: $n \in \mathbb{N}$. Convex functions $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$. Widths $w_1, \ldots, w_n > 0$ and bounds $x_{\min}, x_{\max} \in \mathbb{R}$ with $x_{\max} - x_{\min} \geq w_1 + \ldots + w_n$.

Output: $x_1, \ldots, x_n$ with $x_{\min} \leq x_1$, $x_i + w_i \leq x_{i+1}$ for $i = 1, \ldots, n-1$, $x_n \leq x_{\max}$, and $\sum_{i=1}^{n} f_i(x_i)$ minimum.

```
① $x_0 := x_{\min}$.
  $W_0 := 0$, $W_i := w_i$ for $i = 1, \ldots, n$.
  Let $\mathcal{L}$ be the list consisting of $0, 1, \ldots, n$.
  $i := 1$.
```

Algorithm for a single zone (2)

```
② Let $h$ be the predecessor of $i$ in $\mathcal{L}$.
   If $h = 0$ or
   $x_h + W_h \leq \min\{x_{\max} - W_i, \max\{x : f_i(x) \text{ minimum}\}\}$
   then go to ③ else go to ④.

③ $x_i := \max\{x_h + W_h, \min\{x_{\max} - W_i, \min\{x : f_i(x) \text{ minimum}\}\}\}$.
   If there is a successor $j$ of $i$ in $\mathcal{L}$
   then set $i := j$ and go to ② else go to ⑤.

④ Redefine $f_h$ by $f_h : x \mapsto f_h(x) + f_i(x + W_h)$.
   $W_h := W_h + W_i$.
   Remove $i$ from $\mathcal{L}$.
   $i := h$
   Go to ②.

⑤ For $i \in \{1, \dots, n\} \setminus \mathcal{L}$ do: $x_i := x_h + \sum_{j=h}^{i-1} w_j$, where $h$ is the
   maximum index smaller than $i$ that belongs to $\mathcal{L}$.
```

Algorithm for a single zone: result

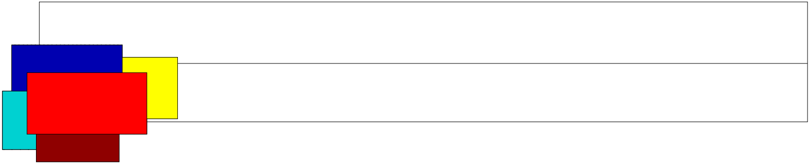
Theorem

This algorithm finds an optimum placement in linear time.

(Brenner, Vygen [2004], based on
Kahng, Tucker, Zelikovsky [1999])

Problem: zones can be very wide

Example:

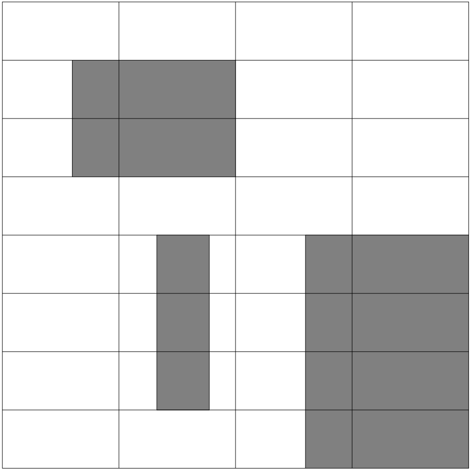


Even if all cells can be placed within the lower zone it is much better to move some of them to the upper zone.

Idea: partition into columns

Subdivide zones into regions.

Example:



An area with 22 zones and 44 regions.

Which cells should be moved where?

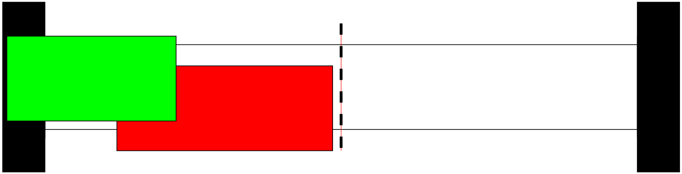
Idea:

Formulation as a minimum cost flow problem, where

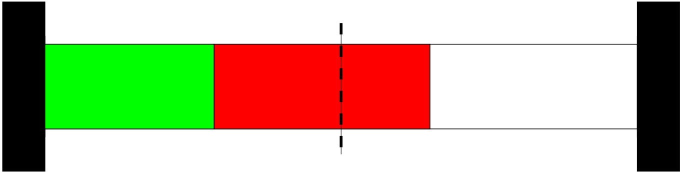
- the vertices are the regions,
- edges connect adjacent regions,
- regions with overload are sources, and
- regions with free capacity are sinks.

But this causes unnecessary movements

Example:



The left region would be a supply region although the two cells could be placed legally with their centers in this region:



Even for a legal placement it is often impossible to assign cells to regions such that no regions is overloaded!

Relaxing constraints

Ideas:

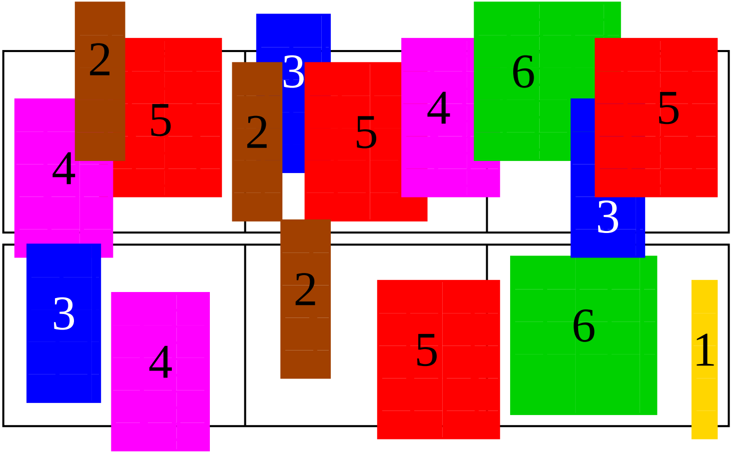
- Only require that at least half of a cell is placed within its region.
- Consider sequences of regions instead of single regions.

Notation:

An interval is a sequence of consecutive regions in the same zone.

- Let $\{A_1, \ldots, A_l\}$ be a set of regions that form a usable zone (ordered from left to right).
- Let $C^i = \{c^i_1, \ldots, c^i_{k_i}\}$ be the set of cells assigned to region A_i, ordered from left to right (for $i \in \{1, \ldots, l\}$).
- Let w denote (total) width.

Example



Supply intervals

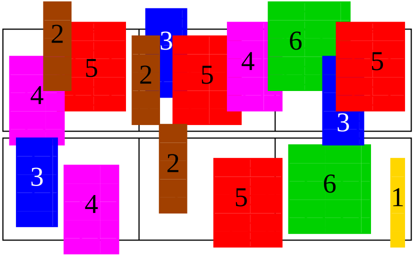
To compute the size of cells that have to be removed from an interval $A_{\mu,\nu}$ we define for $1 \leq \mu \leq \nu \leq l$:

$$s_{\mu,\nu} := \max\left\{0,\ \sum_{i=\mu}^{\nu}\Big(w(C^i) - w(A_i)\Big) - \frac{1}{2}\left(w(c_1^{\mu}) + w(c_{k_\nu}^{\nu})\right)\right\}.$$

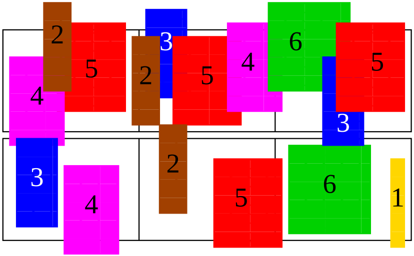
Using these numbers, we define recursively (for $1 \leq \mu \leq \nu \leq l$):

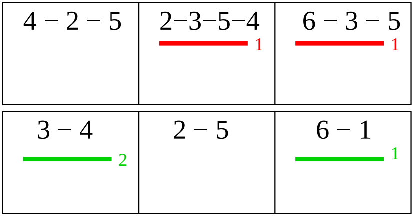
$$\mathrm{supp}(A_{\mu,\nu}) := \max\left\{0,\ s_{\mu,\nu} - \sum_{\substack{\mu \leq \mu' \leq \nu' \leq \nu \\ (\mu,\nu) \neq (\mu',\nu')}} \mathrm{supp}(A_{\mu',\nu'})\right\}.$$

Example: initial placement



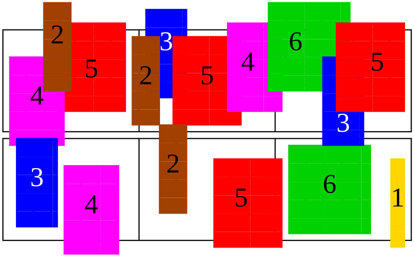
Example: supply and demand regions

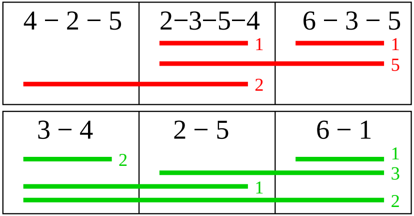




4 − 2 − 5	2−3−5−4 1	6 − 3 − 5 1
3 − 4 2	2 − 5	6 − 1 1

Example: supply and demand intervals





Demand intervals

To compute the size of cells that can be moved into an interval $A_{\mu,\nu}$ we define for $1 \leq \mu \leq \nu \leq l$:

$$t_{\mu,\nu} := \min \left\{ 0,\ \sum_{i=\mu}^{\nu} \Big(w(C^i) - w(A_i) \Big) + \frac{1}{2} \Big(w(c^{\mu-1}_{k_{\mu-1}}) + w(c^{\nu+1}_1) \Big) \right\}.$$

Using these numbers, we define recursively (for $1 \leq \mu \leq \nu \leq l$):

$$\mathrm{dem}(A_{\mu,\nu}) := \min \left\{ 0,\ t_{\mu,\nu} - \sum_{\substack{\mu \leq \mu' \leq \nu' \leq \nu \\ (\mu,\nu) \neq (\mu',\nu')}} \mathrm{dem}(A_{\mu',\nu'}) \right\}.$$

Theorem

- No region can be both part of a demand interval and part of a supply interval.
- For $\mu < \kappa \leq \lambda < \nu$ with supp$(A_{\kappa,\lambda}) > 0$ we have supp$(A_{\mu,\nu}) = 0$.
- For $\mu < \kappa \leq \lambda < \nu$ with dem$(A_{\kappa,\lambda}) < 0$ we have dem$(A_{\mu,\nu}) = 0$.
- The number of supply and demand intervals is at most twice the number of regions.
- They can be computed in linear time.

(Brenner, Vygen [2004])

The minimum cost flow instance

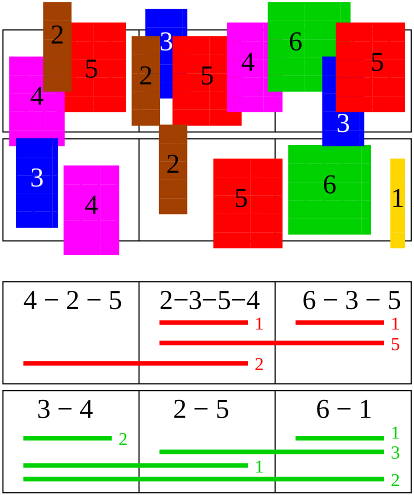
$$
\begin{aligned}
V(G) &:= \{\text{regions, supply intervals, demand intervals}, s, t\} \\
E(G) &:= \{(A, A') : A, A' \text{ adjacent regions}\} \\
&\quad \cup \{(A, A') : A \text{ supply interval}, A' \text{ maximal proper subset of } A\} \\
&\quad \cup \{(A, A') : A' \text{ demand interval}, A \text{ maximal proper subset of } A'\}
\end{aligned}
$$

- For two adjacent regions A and A', let $c(A, A')$ be the expected cost of moving a cell of width 1 from A to A'.
- All other arcs have zero cost. All arcs have infinite capacity.

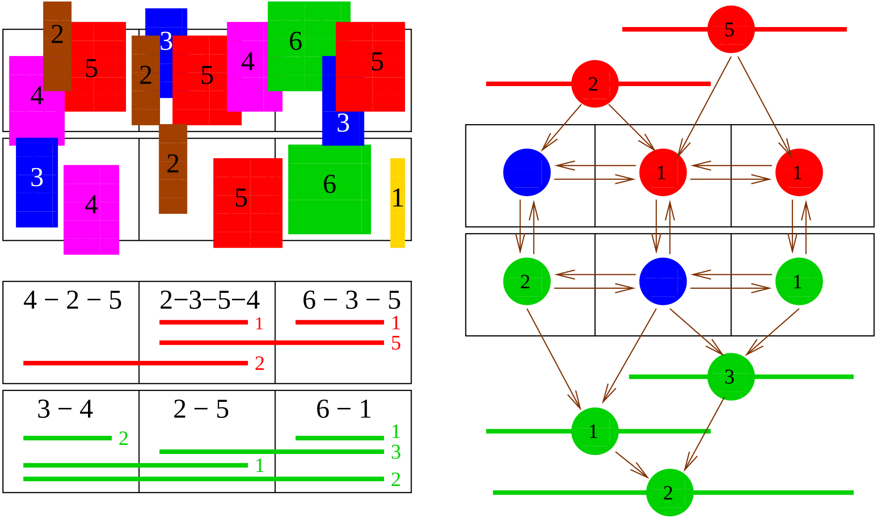
We look for a minimum cost flow f with $f(\delta^+(v)) - f(\delta^-(v)) \geq \mathrm{supp(v)} + \mathrm{dem(v)}$ for all $v \in V(G)$.

This can be done in $O(n^2 \log^2 n)$ time by standard min-cost flow algorithms (Orlin [1993], Vygen [2002])

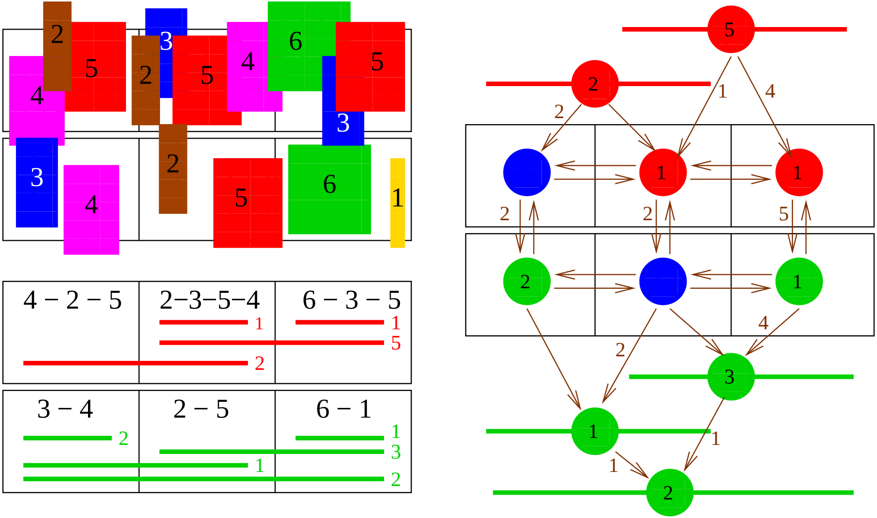
Example: supply and demand intervals



Example: minimum cost flow instance



Example: minimum cost flow



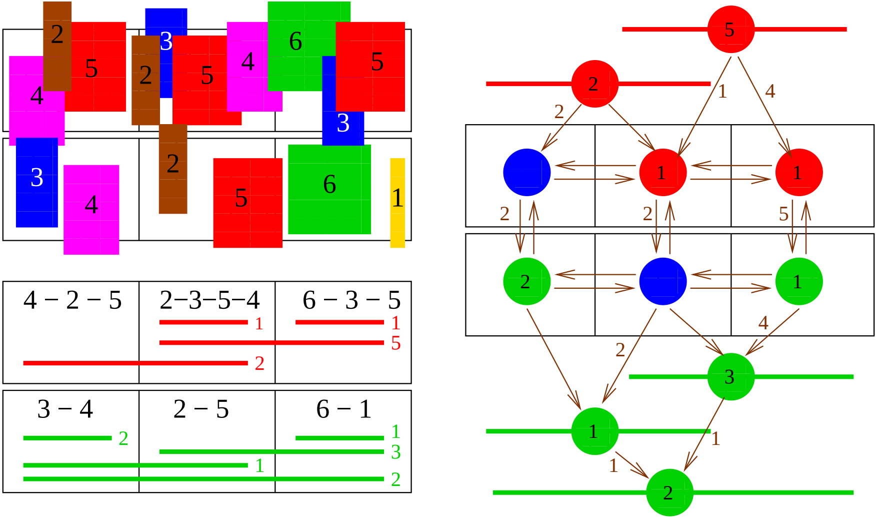
Realization of the flow

By realizing a flow f we mean moving cells of total size $f(A, A')$ from region A to region A' for each pair of neighbours (A, A').

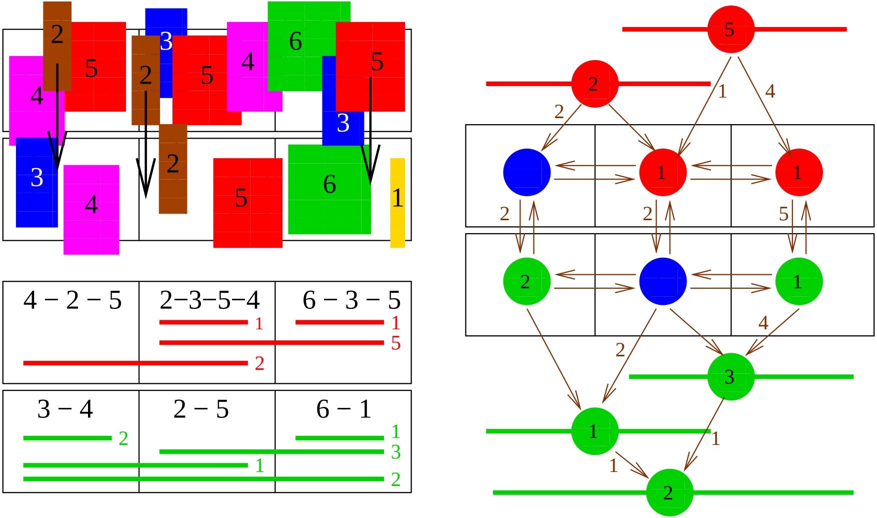
Theorem

- *Let f be a solution to the minimum cost flow instance. Then a realization of f that does not move any leftmost or rightmost cell of a region yields a feasible assignment of the cells.*
- *On non-trivial instances, we cannot decrease the supply- or increase the demand-values without losing this property.*

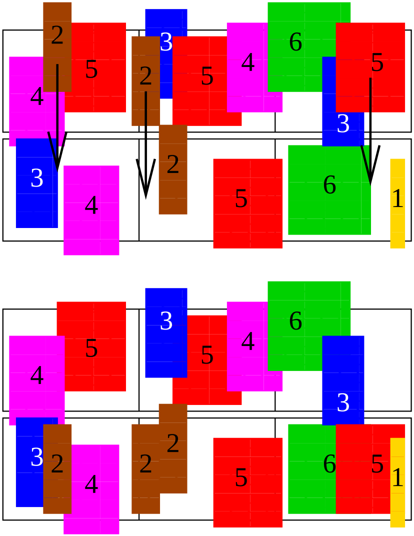
Example: minimum cost flow

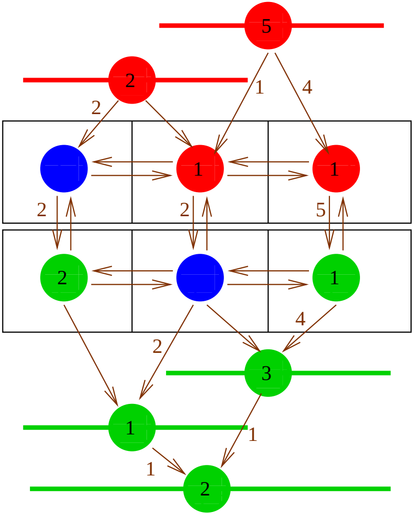


Example: realizing the flow

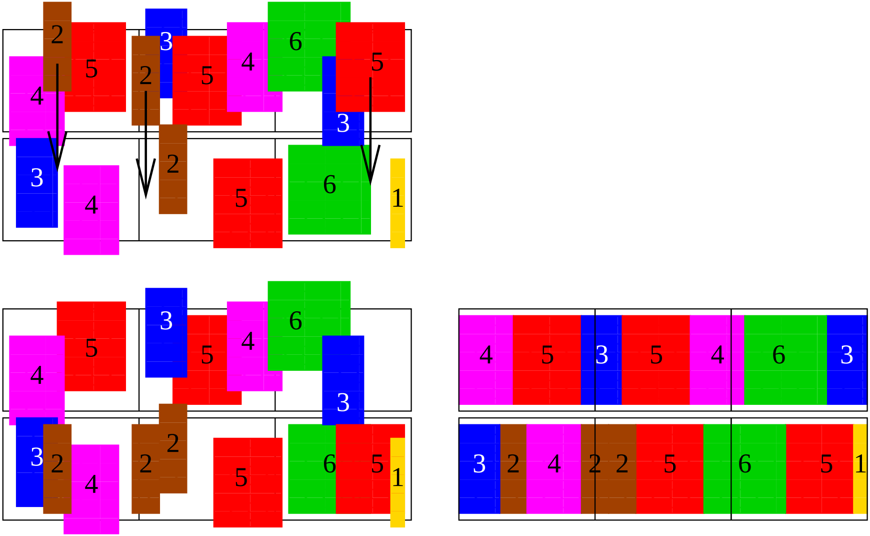


Example: realizing the flow





Example: legal placement



Realization of the flow

Exact realization is in general impossible. We consider approximate realizations:

Theorem

Moving cells between regions such that the total size of cells that leave $A_{\mu,\nu}$ minus the total size of cells that are moved into $A_{\mu,\nu}$ is at least

$$\sum_{i=\mu}^{\nu}\Big(w(C^i) - w(A_i)\Big) - \frac{1}{2}\left(w(c_1^{\mu}) + w(c_{k_\nu}^{\nu})\right)$$

for each interval $A_{\mu,\nu}$ leads to an assignment of the cells to the regions for which there is a legal placement such that each cell is placed within the region it is assigned to or within a horizontally adjacent region.

Realization of the flow

- The arcs carrying flow form an acyclic subgraph. Consider the vertices in topological order w.r.t. this subgraph.
- The cells to be moved are chosen according to the solution of a MULTI-KNAPSACK PROBLEM (dynamic programming), trying to maintain feasibility.
- We cannot always find cells of appropriate total size
  $\Rightarrow$ There can still be overloads after the realization.

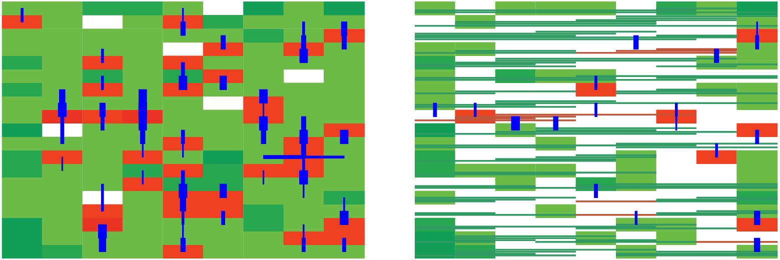
Overall algorithm

- Compute the min-cost flow instance.
- Find a minimum cost flow f.
- Realize f by moving cells along the flow edges.
- Repeat these steps as long as there are overloaded zones. (If necessary, increase column width, decrease demand values.)
- Step 2: Legalize the cells within their zones.
- Step 3: Postoptimization: each step consists of a legal sequence of moves

  $$c_0 \longrightarrow c_1 \longrightarrow \cdots \longrightarrow c_k \longrightarrow (\text{place of } c_0 \text{ or free place})$$

  reducing total (squared) movement. Dynamic programming.

Detailed Placement: old and new approach



moving between regions
(old approach)

moving between intervals
(new approach)

rectangles = regions

horizontal lines = intervals

green = demand regions/intervals

red = supply regions/intervals

blue = edges with flow, width proportional to amount of flow

Lower bound: integer linear programming formulation

$$\text{minimize} \sum_{k=1}^{|C|} \sum_{i=1}^{W} \sum_{j=1}^{H} d_{i,j,k} \cdot x_{i,j,k}$$

subject to

$$x_{i,j,k} \in \{0,1\} \qquad \forall\ i = 1,\ldots,W,\ j = 1,\ldots,H,\ k = 1,\ldots,|C|$$

$$\sum_{i=1}^{W} \sum_{j=1}^{H} x_{i,j,k} = 1 \qquad \forall\ k = 1,\ldots,|C|$$

$$\sum_{k=1}^{|C|} \sum_{i'=i-w(c_k)+1}^{i} x_{i',j,k} \leq 1 \qquad \forall\ i = 2,\ldots,W,\ j = 1,\ldots,H$$

where $d_{i,j,k} := (x(c_k) - i)^2 + (y(c_k) - j)^2$

Lower bound: LP relaxation

Let $\delta > 0$ be a usually sufficient radius.

$$\text{minimize} \sum_{k=1}^{|C|} \left(\sum_{i=1}^{W} \sum_{j=1}^{H} d_{i,j,k} \cdot x_{i,j,k} + \delta \cdot x_{\delta,k} \right)$$

subject to

$$0 \le x_{i,j,k} \le 1 \qquad \forall\ i = 1, \dots, W,\ j = 1, \dots, H,\ k = 1, \dots, |C|$$

$$\left(\sum_{i=1}^{W} \sum_{j=1}^{H} x_{i,j,k} \right) + x_{\delta,k} = 1 \qquad \forall\ k = 1, \dots, |C|$$

$$\sum_{k=1}^{|C|} \sum_{i'=i-w(c_k)+1}^{i} x_{i',j,k} \le 1 \qquad \forall\ i = 2, \dots, W,\ j = 1, \dots, H$$

$\Rightarrow$ We can skip all variables $x_{i,j,k}$ with $d_{i,j,k} \ge \delta$.

Integrality gap

- We do not know the integrality gap of this LP.
- However, a simple example shows that it is at least $\frac{12}{10}$ (for $\delta = \infty$).

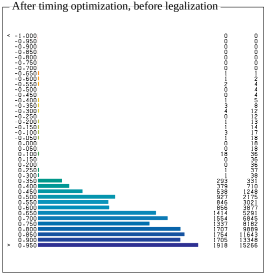
Detailed placement: experimental results

weighted average of squared Euclidean distances in μm:

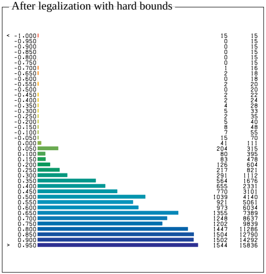
number of objects	old	new	difference (%)	lower bound	gap (%)
72 447	18.06	13.65	24.4	12.37	10.3
72 794	18.67	7.57	59.5	7.34	3.1
284 705	75.18	6.95	90.8	6.25	11.2
411 926	17.32	10.92	37.0	9.85	10.9
1 301 795	8.44	6.08	28.0	5.84	4.1
1 645 691	9.72	5.41	44.3	5.01	7.9
2 395 218	14.95	3.40	77.3	3.08	10.4

lower bound: LP relaxation, solved by CPLEX
HB = hard boundaries between regions (old approach)
SB = soft boundaries between regions (new approach)
maximum total runtime: 40 minutes, 8.5 GB memory

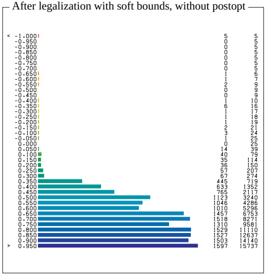
Timing experiments: legalization does not hurt



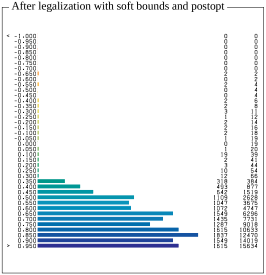
(a) before



(b) HB (old)



(c) SB (new)



(d) after postopt

Introduction

Placement

Routing
    Problem formulation, general approach
    Detailed Routing
    Global Routing

Timing Optimization

Clocktree Design

VLSI routing: task

Instance:

- a number of routing planes
- a set of nets, where each net is a set of pins (terminals)
- a set of shapes for each pin, each of which is a rectangle in a routing plane
- a set of blockage shapes
- rules that tell when two shapes are connected and when they are separated
- timing constraints, information on power, crosstalk, yield, ...

Task:

Compute a feasible routing, i.e. a set of wire shapes for each net, connecting the pins, and separate from blockages and shapes of other nets

- such that all timing constraints are met
- and the (estimated) power consumption is minimized.

VLSI routing: simplified view

Find vertex-disjoint Steiner trees connecting given terminal sets in a 3-dimensional grid graph.

Order of magnitude: 5 million Steiner trees in a graph with 100 billion vertices!

⟶ Even linear-time algorithms are too slow!

Global and detailed routing

VLSI routing is usually performed in three phases:

- Global routing: Eliminates congestion and timing problems on a global level, performs global optimization, and determines corridors for each net to reduce search space in detailed routing
- Detailed routing: Actually constructs wires connecting each net within the corridors obtained from global routing, respecting all design rules necessary for the lithographic processes in fabrication
- Postoptimization: Improve the wiring by spreading and do some postprocessing for more robust manufacturing

Today's designs are huge: 100,000,000,000 vertices in detailed routing, 10,000,000 vertices in global routing. In fact even more, as the underlying grid is an abstraction that does not work anymore.

Key features of global and detailed routing

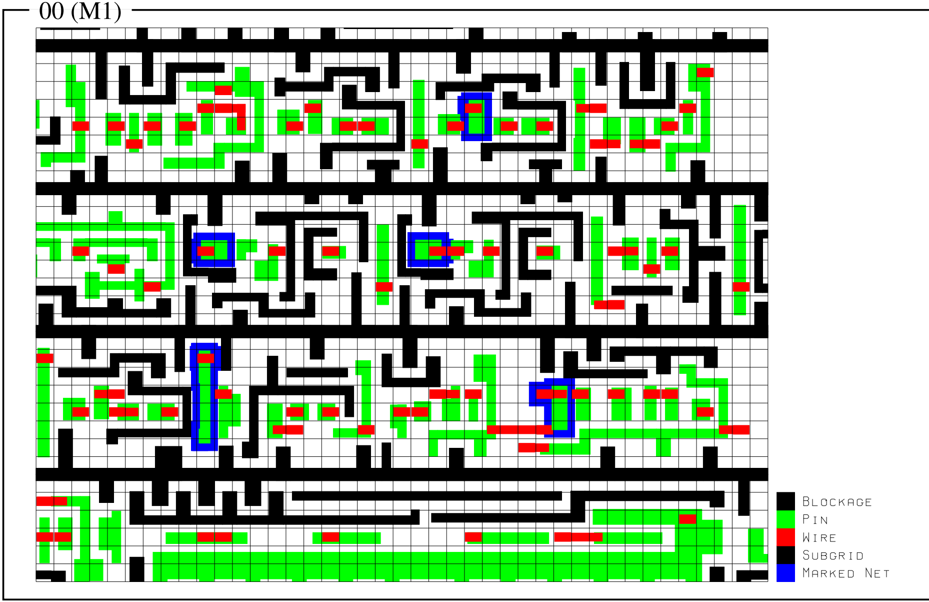
Global routing

- contract regions of approx. 100x100 points to a single vertex
- compute capacities of edges between adjacent regions
- pack Steiner trees with respect to these edge capacities
- do global optimization
- define a detailed routing area for each net according to its Steiner tree

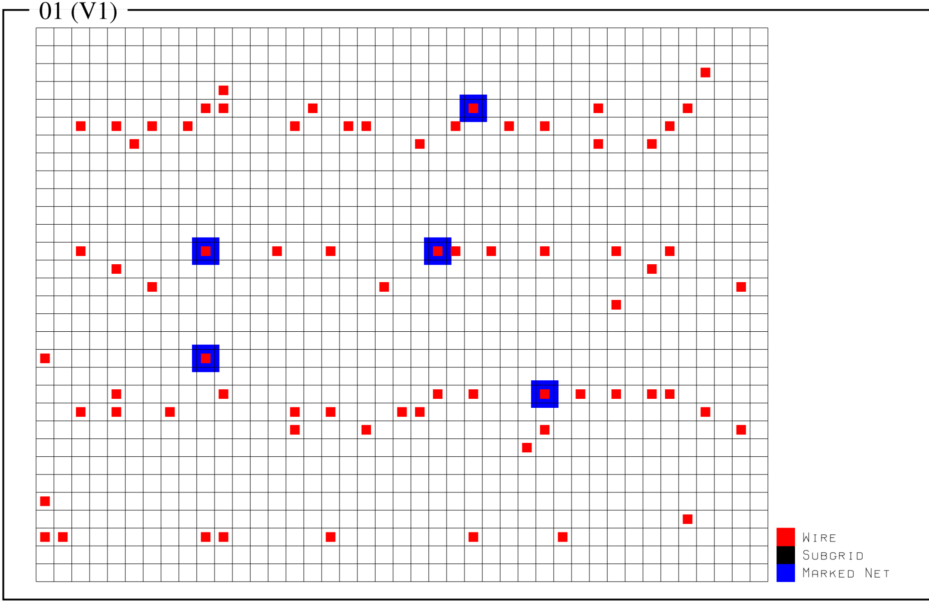
Detailed routing

- route nets sequentially, mainly by shortest path algorithms
- goal-oriented shortest path algorithms
- label intervals rather than single points
- restrict path search to small areas

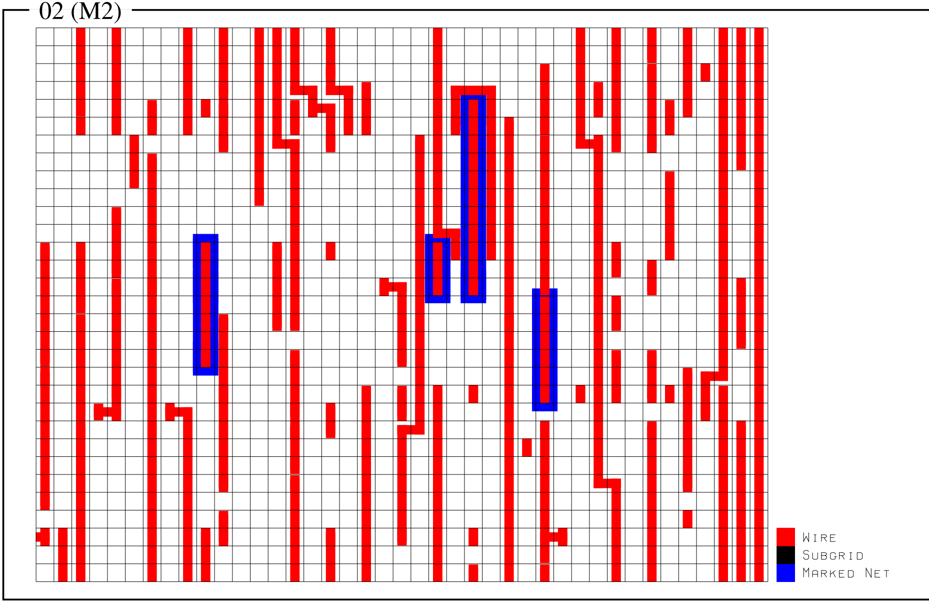
Detailed routing: example



Detailed routing: example

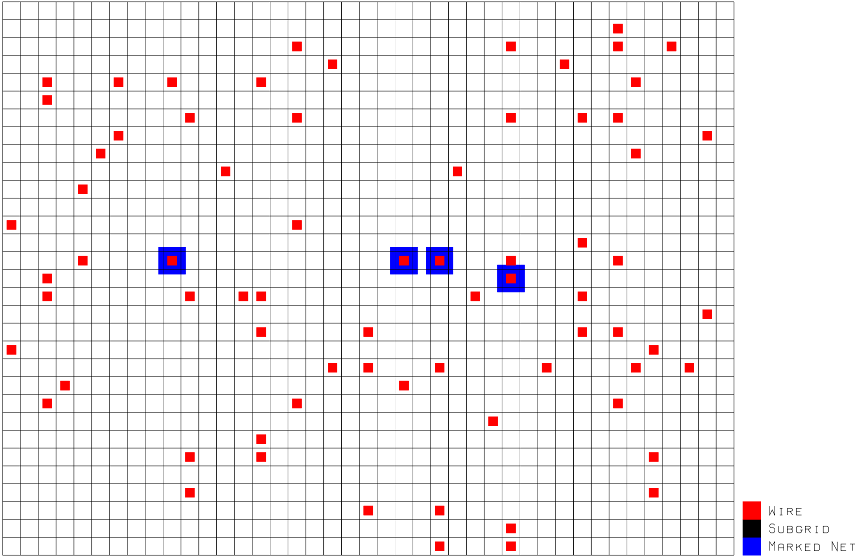


Detailed routing: example

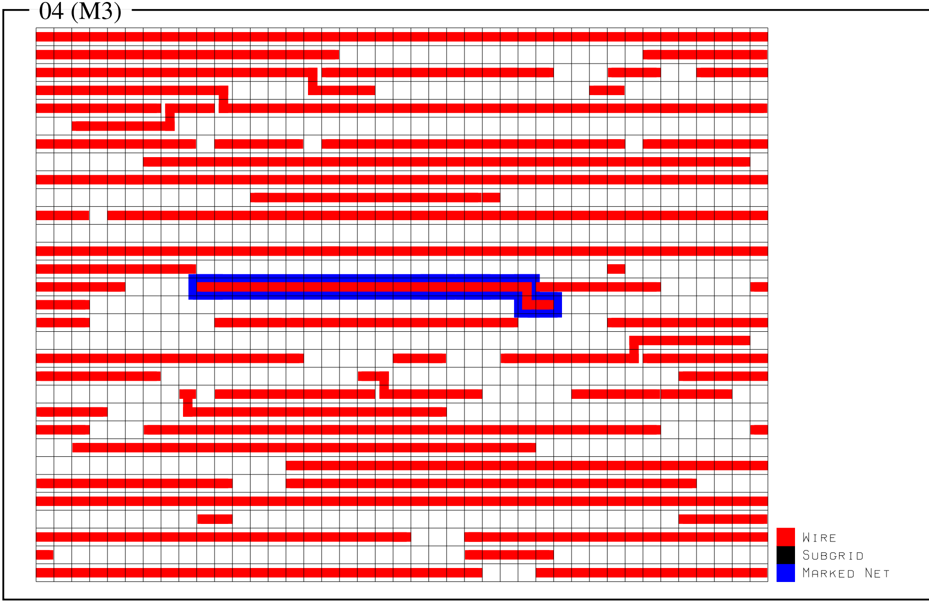


Detailed routing: example

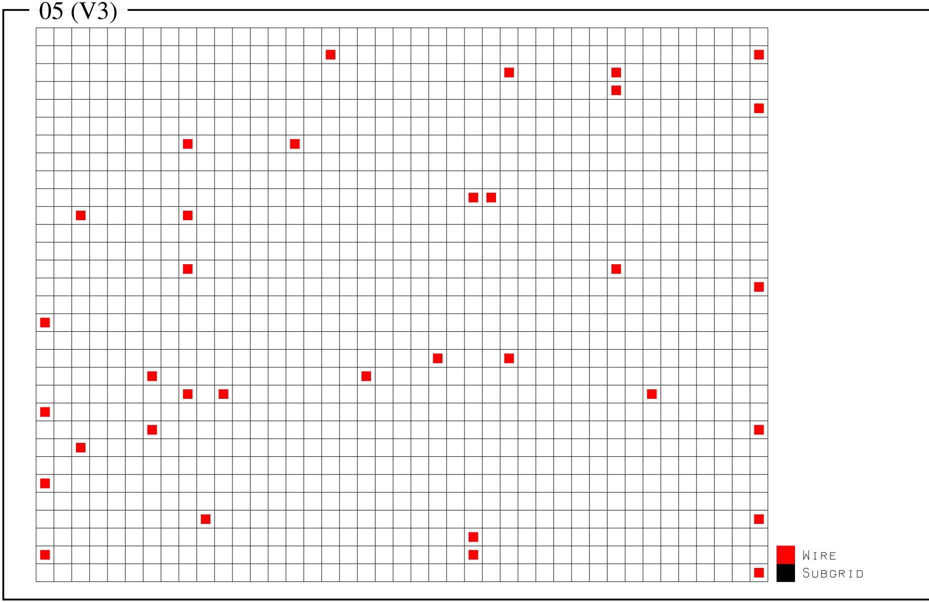
03 (V2)



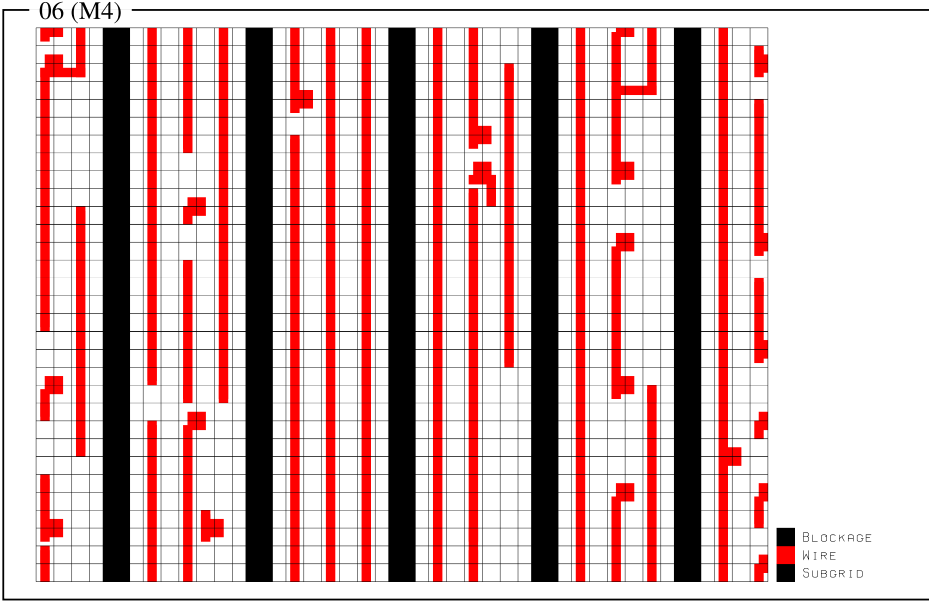
Detailed routing: example



Detailed routing: example

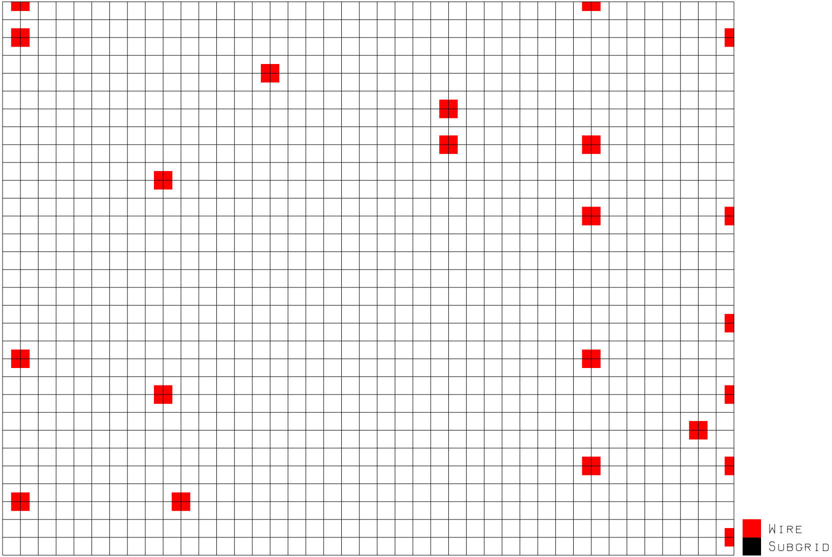


Detailed routing: example

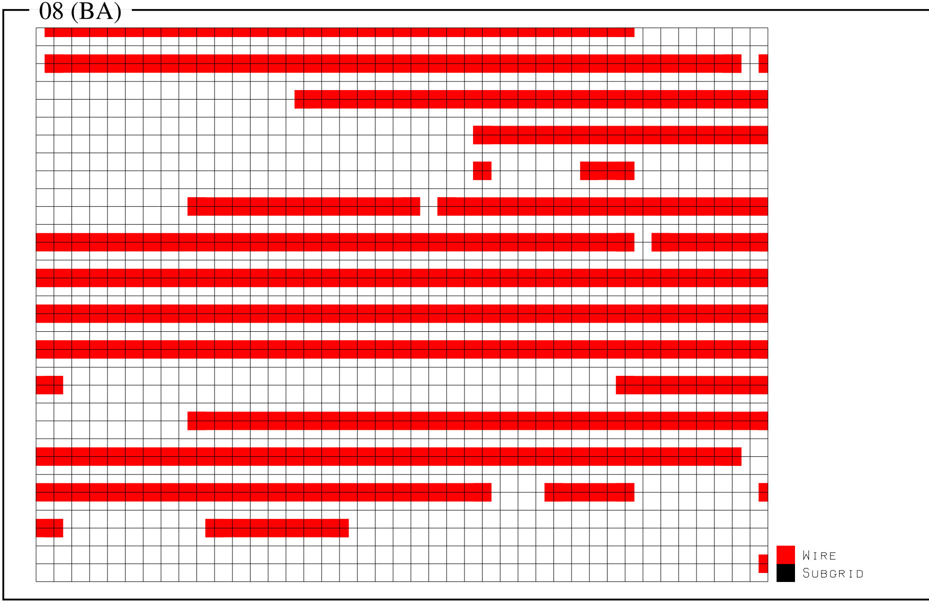


Detailed routing: example

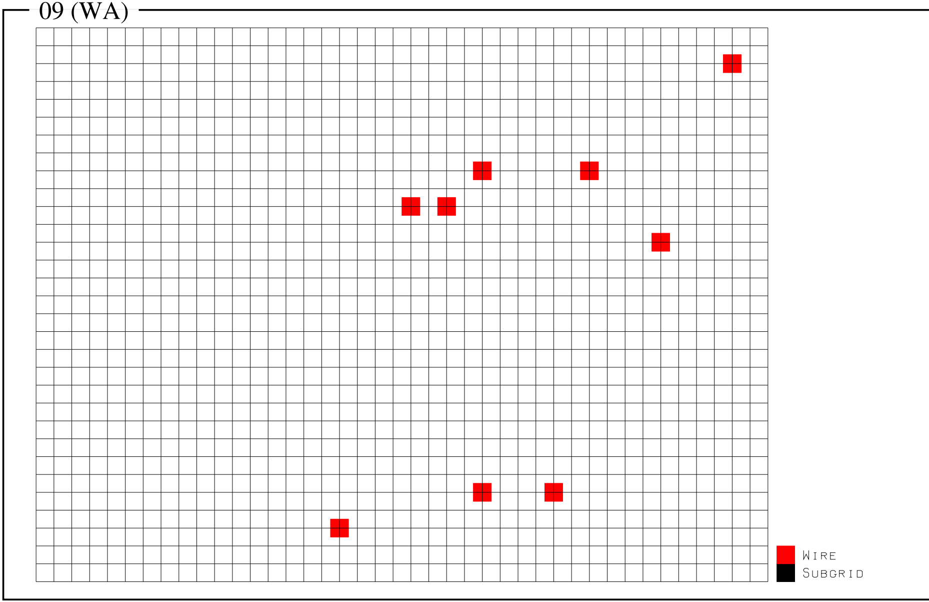
07 (WT)



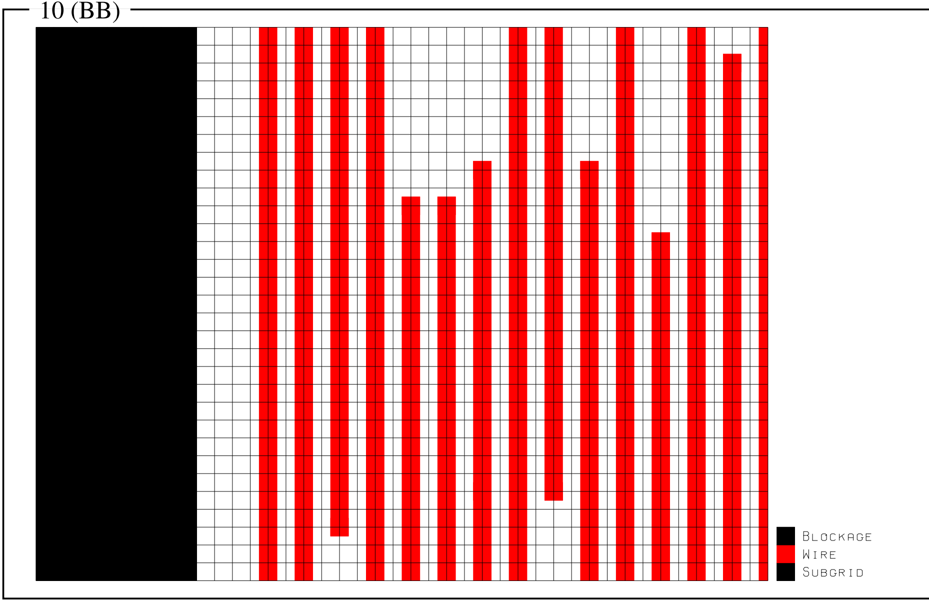
Detailed routing: example



Detailed routing: example



Detailed routing: example



Future cost — feasible potentials

Given a digraph G with arc costs $c : E(G) \to \mathbb{R}_+$.

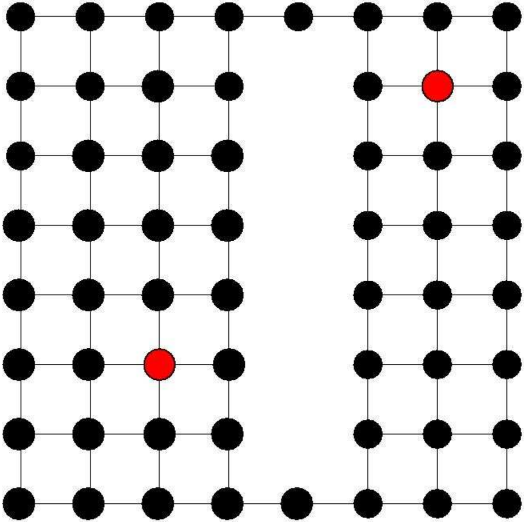
A function $\pi : V(G) \to \mathbb{R}$ is called a feasible potential if the reduced cost $c_\pi(e) := c(e) + \pi(v) - \pi(w)$ is nonnegative for each $e = (v, w) \in E(G)$.

Let $s, t \in V(G)$. We look for a shortest s-t-path w.r.t. c.

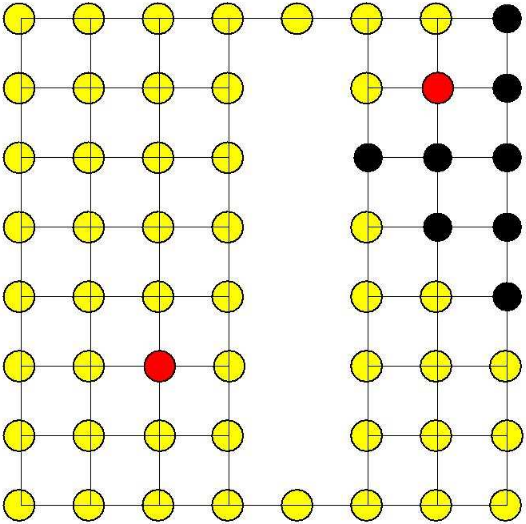
Observation: A shortest s-t-path w.r.t. c is a shortest s-t-path w.r.t. c_π, and vice versa.

Suppose $\mathcal{L}(x)$ is a lower bound on the distance from x to t, and
$\mathcal{L}(v) \leq c(e) + \mathcal{L}(w)$ for each $e = (v, w) \in E(G)$.
Then $\pi(x) := -\mathcal{L}(x)$ is a feasible potential.
$\mathcal{L}(x)$ is also called the future cost at x.

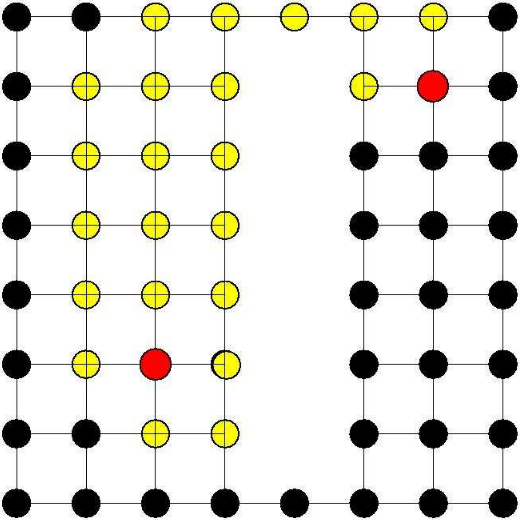
Future cost: example



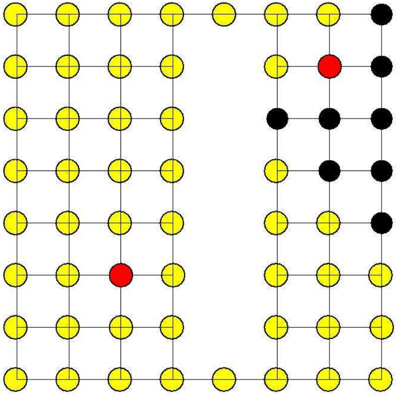
Dijkstra without future cost



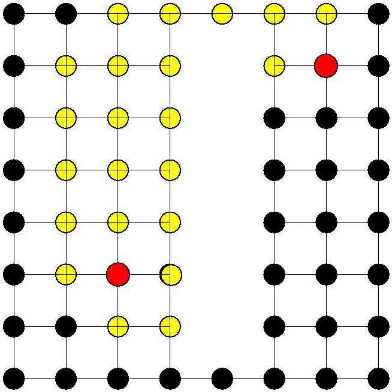
Dijkstra with future cost



Comparison with and without future cost

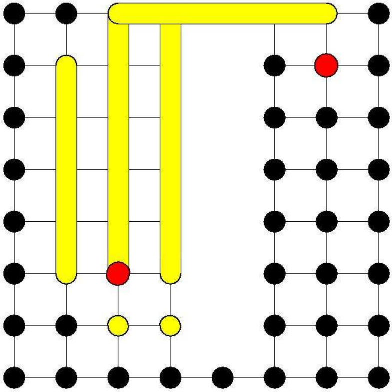


50 points labelled

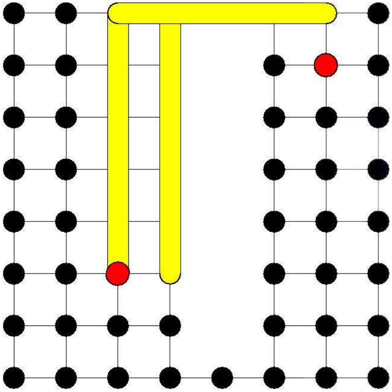


24 points labelled

Comparison with and without future cost



7 intervals labelled



4 intervals labelled

Detailed routing

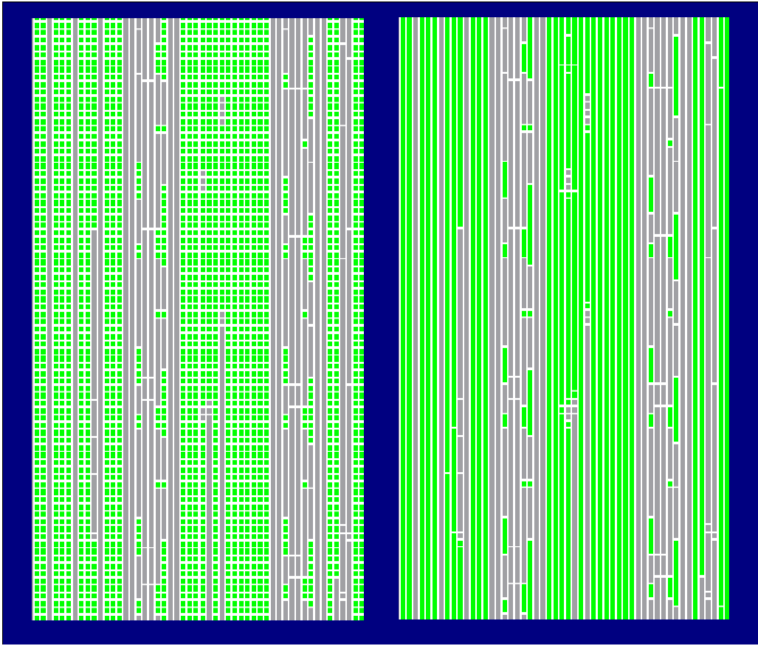
- route nets sequentially, subnets by a variant of Dijkstra's algorithm
- goal-oriented Dijkstra: future cost
- label intervals rather than single points
- restrict path search to small areas (computed by global routing)

Theorem

Running time iof modified Dijkstra is $O((d+1)l \log l)$, *where* d *is the detour (actual length minus lower bound), and* l *is the number of intervals in the search space.*

(Hetzel [1998])

Detailed routing: intervals



Global routing: simplified problem formulation

Instance:

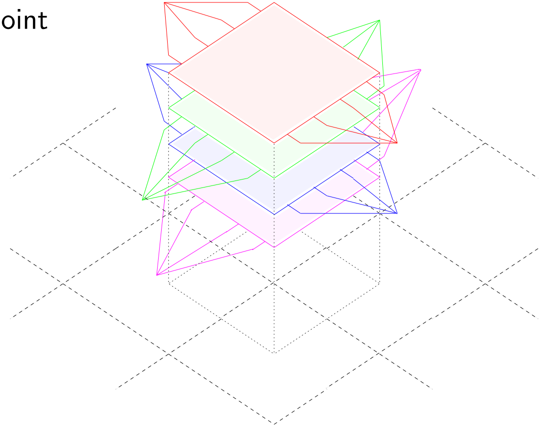
- a global routing (grid) graph with edge capacities
- a set of nets, each consisting of a set of vertices (terminals)

Task: find a Steiner tree for each net such that

- the edge capacities are respected,
- some objective function (e.g., netlength, yield, or power) is optimized,
- and the timing constraints are met.

Capacity estimation

- First route very short nets (within one region or two adjacent regions).
- Then consider each pair of adjacent regions. Assume that planes are mainly used in preferred wiring direction, alternatingly horizontal and vertical.
- Consider the following instance of the edge-disjoint paths problem:



Capacity estimation: fast augmenting path heuristic

- Apply a very fast multicommodity flow heuristic, exploiting the structure of the instances. (Müller [2002])
- Each augmenting path requires only $O(k)$ constant-time bit pattern operations, where k is the number of edges orthogonal to the preferred wiring direction.
- Heuristic finds a feasible integral multicommodity flow solution whose value is approx. 90% of the (weak) max-flow upper bound.
- Complete chip with 300 million paths in 15 minutes (Goldberg-Tarjan runs 1 month)

Global routing is hard

Restriction: EDGE-DISJOINT PATHS PROBLEM
Given a pair of graphs (G, H), find a family $(P_h)_{h \in E(H)}$ of edge-disjoint paths in G such that $P_h + h$ is a circuit for each $h \in E(H)$.
NP-complete even if

- G is a rectangle (Raghavan [1986])
- G is a rectangle, and we allow shortest paths only (Vygen [1994])
- G is a rectangle, and $G + H$ is Eulerian (Marx [2002])
- G is series-parallel (Nishizeki, Vygen, Zhou [2001])
- G is directed and planar, H consists of two sets of parallel arcs (Müller [2002])

Fractional relaxation: Multicommodity Flow Problem

Instance:

- an undirected graph G with capacities $u : E(G) \to \mathbb{Z}_+$ and lengths $l : E(G) \to \mathbb{R}$
- a family $\mathcal{N}$ of nets (terminal pairs) with demands $w : \mathcal{N} \to \mathbb{Z}_+$ and weights $c : \mathcal{N} \to \mathbb{Z}_+$

Task: Find a flow f_N for each N of value $w(N)$ such that

$$\sum_{N \in \mathcal{N}} w_N f_N(e) \leq u(e) \qquad \text{for } e \in E(G),$$

and

$$\sum_{N \in \mathcal{N}} c_N \sum_{e \in E(G)} l(e) f_N(e) \qquad \text{is minimum.}$$

In many applications: congestion costs — heavily used edges are more expensive

Examples: traffic flows, VLSI routing

Global routing: positive results

- There is a combinatorial fully polynomial approximation scheme for the MULTICOMMODITY FLOW PROBLEM (Sharokhi, Matula [1990], Leighton, Makedon, Plotkin, Stein, Tardos, Tragoudas [1991], Plotkin, Shmoys, Tardos [1991], Radzik [1995], Young [1995], Grigoriadis, Khachiyan [1996], Garg, Könemann [1998], Fleischer [2000], Karakostas [2002])
- If edges have sufficient capacity, randomized rounding can be applied to get an integral solution violating capacity constraints only slightly (Raghavan, Thompson [1987,1991], Raghavan [1988])
- This can be applied to Steiner trees instead of paths and works efficiently for large global routing instances (Albrecht [2001])

But this does not take timing constraints and global objectives (power consumption, yield) into account.

Timing constraints in routing

The delay on each path must not exceed its bound. A path can be viewed as a sequence of nets. The delay of a net depends on its electrical capacitance.

- first assume delay-optimal Steiner trees for all nets
- distribute slack optimally (Albrecht, Korte, Schietke, Vygen [2000], Held [2001]) to all nets for which sufficient slack is available. For these nets the slack defines a maximum tolerable capacitance
- call the remaining nets (with no or insufficient slack assigned) critical
- compute weights and a bound on the weighted sum of capacitances for each path containing a critical net

Main design objectives in routing

minimize power consumption

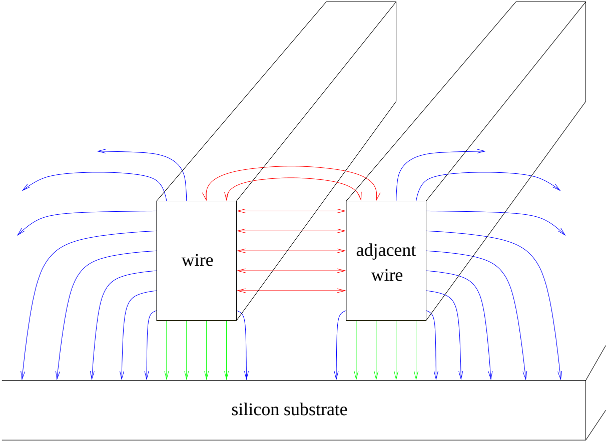
- active power consumption roughly proportional to the electrical capacitance, weighted by switching activity
- leakage power and capacitance of cells not influenced by routing.
- capacitance of nets depends on length, width, plane, and existence of neighbour wires

minimize cost

- minimize number of masks (number of routing planes), maximize yield (spreading), minimize design effort

Capacitance estimation

- area capacitance (parallel plate capacitor) – proportional to length times width
- fringing capacitance – proportional to length
- coupling capacitance – proportional to length if adjacent wire exists



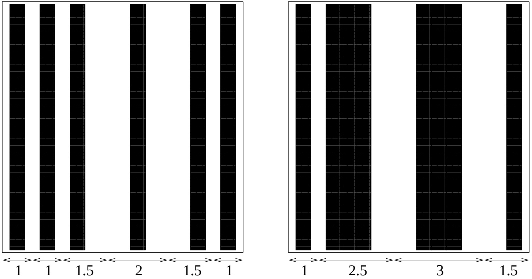
Modeling coupling capacitance

Assume linear dependence on distance to adjacent wire between the following bounds:

- minimum distance → coupling capacitance $\frac{1}{2}v(e)$
- minimum distance plus 1 → coupling capacitance 0

Example:

global routing edge e of capacity $u(e) = 8$, with two global routing solutions:



- Left: six unit width wires use 6–12 channels. Coupling capacitance $v(e)$ times $1, 1, \frac{1}{2}, 0, \frac{1}{2}, 1$
- Right: two unit width wires and two double width wires use 6–10 channels. Coupling capacitance $v(e)$ times $1, \frac{1}{2}, 0, \frac{1}{2}$

Global Routing Problem

Instance:

- An undirected graph G with edge capacities $u : E(G) \to \mathbb{R}_+$,
- a set $\mathcal{N}$ of nets and a set $\mathcal{Y}_N$ of feasible Steiner trees for each net N,
- wire widths $w : E(G) \times \mathcal{N} \to \mathbb{R}_+$,
  extra space $s : E(G) \times \mathcal{N} \to \mathbb{R}_+$,
- maximum capacitances $l : E(G) \times \mathcal{N} \to \mathbb{R}_+$ and
  coupling contributions $v : E(G) \times \mathcal{N} \to \mathbb{R}_+$.
- A family $\mathcal{M}$ of subsets of $\mathcal{N}$ with $\mathcal{N} \in \mathcal{M}$ with capacitance bounds $U : \mathcal{M} \to \mathbb{R}_+$ and weights $c(M, N) \in \mathbb{R}_+$ for $N \in M \in \mathcal{M}$.

Global Routing Problem

Task:

Find a Steiner tree $Y_N \in \mathcal{Y}_N$ and numbers $0 \le y_{e,N} \le 1$ for each $N \in \mathcal{N}$ and $e \in E(Y_N)$, such that

$$\sum_{N \in \mathcal{N} : e \in E(Y_N)} (w(e,N) + s(e,N) y_{e,N}) \le u(e)$$

for each edge $e \in E(G)$,

$$\sum_{N \in M} c(M,N) \sum_{e \in E(Y_N)} (l(e,N) - v(e,N) y_{e,N}) \le U(M)$$

for $M \in \mathcal{M}$, and such that

$$\sum_{N \in \mathcal{N}} c(\mathcal{N},N) \sum_{e \in E(Y_N)} (l(e,N) - v(e,N) y_{e,N})$$

is minimum.

LP relaxation of the Global Routing Problem

$\min \lambda$ subject to

$$\sum_{Y \in \mathcal{Y}_N} x_{N,Y} = 1 \qquad (N \in \mathcal{N})$$

$$\sum_{N \in M} c(M,N) \left(\sum_{Y \in \mathcal{Y}_N} \sum_{e \in E(Y)} l(e,N) x_{N,Y} - \sum_{e \in E(G)} v(e,N) y_{e,N} \right) \leq \lambda U(M) \qquad (M \in \mathcal{M})$$

$$\sum_{N \in \mathcal{N}} \left(\sum_{Y \in \mathcal{Y}_N : e \in E(Y)} w(e,N) x_{N,Y} + s(e,N) y_{e,N} \right) \leq \lambda u(e) \qquad (e \in E(G))$$

$$y_{e,N} \leq \sum_{Y \in \mathcal{Y}_N : e \in E(Y)} x_{N,Y} \qquad (e \in E(G), N \in \mathcal{N})$$

$$y_{e,N} \geq 0 \qquad (e \in E(G), N \in \mathcal{N})$$

$$x_{N,Y} \geq 0 \qquad (N \in \mathcal{N}, Y \in \mathcal{Y}_N)$$

The dual LP

$$\max \sum_{N \in \mathcal{N}} z_N \text{ subject to}$$

$$\sum_{e \in E(G)} u(e)\omega_e + \sum_{M \in \mathcal{M}} U(M)\mu_M = 1$$

$$z_N \leq \sum_{e \in E(Y)} \left(l(e,N) \sum_{M \in \mathcal{M} : N \in M} c(M,N)\mu_M + w(e,N)\omega_e - \chi_{e,N} \right) \quad (N \in \mathcal{N}, Y \in \mathcal{Y}_N)$$

$$\chi_{e,N} \geq v(e,N) \sum_{M \in \mathcal{M} : N \in M} c(M,N)\mu_M - s(e,N)\omega_e \quad (e \in E(G), N \in \mathcal{N})$$

$$\chi_{e,N} \geq 0 \quad (e \in E(G), N \in \mathcal{N})$$

$$\omega_e \geq 0 \quad (e \in E(G))$$

$$\mu_M \geq 0 \quad (M \in \mathcal{M})$$

Edge costs

Let $\omega_e \in \mathbb{R}_+ (e \in E(G))$ and $\mu_M \in \mathbb{R}_+ (M \in \mathcal{M})$, and let us define edge costs

$$\psi_{N,e} \;:=\; \min_{\delta \in \{0,1\}} \left((l(e,N) \,-\, \delta v(e,N)) \sum_{M \in \mathcal{M}: N \in M} c(M,N)\mu_M \right.$$
$$\left. + (w(e,N) + \delta s(e,N))\omega_e \right).$$

Then

$$\frac{\displaystyle\sum_{N \in \mathcal{N}} \min_{Y \in \mathcal{Y}_N} \sum_{e \in E(Y)} \psi_{N,e}}{\displaystyle\sum_{e \in E(G)} u(e)\omega_e + \sum_{M \in \mathcal{M}} U(M)\mu_M}$$

is a lower bound on the optimum LP value.

The fractional global routing algorithm

Input: An instance of the Global Routing Problem with $\mathcal{N} = \{1, \dots, k\}$, $t \in \mathbb{N}$, $\epsilon \in \mathbb{R}_+$.
Output: Feasible solutions to the primal and dual LP.

Initialize:
Set $\omega_e := \frac{1}{u(e)}$ for $e \in E(G)$ and $\mu_M := \frac{1}{U(M)}$ for $M \in \mathcal{M}$.
Set $x_{i,Y} := 0$ for $i := 1, \dots, k$, $Y \in \mathcal{Y}_i$.
Set $y_{e,i} := 0$ for $e \in E(G)$ and $i := 1, \dots, k$.
Set $Y_i := \emptyset$ for $i := 1, \dots, k$.

(Main Loop)

TakeAverage:
Set $x_{i,Y} := \frac{1}{t} x_{i,Y}$ for $i = 1, \dots, k$ and $Y \in \mathcal{Y}_i$.
Set $y_{e,i} := \frac{1}{t} y_{e,i}$ for $e \in E(G)$ and $i = 1, \dots, k$.

The fractional global routing algorithm: main loop

```
For $p := 1$ to $t$ do:
    For $i := 1$ to $k$ do:
        Let $\psi_{i,e}$ be defined as above.
        Let $Y_i \in \mathcal{Y}_i$ with $\sum_{e \in E(Y_i)} \psi_{i,e}$ minimum.
        UpdateVariables:
        Set $x_{i,Y_i} := x_{i,Y_i} + 1$.
        For $e \in E(Y_i)$ do:
            If $v(e,i) \sum_{M \in \mathcal{M}: i \in M} c(M,i)\mu_M < s(e,N)\omega_e$
                then $\delta_e := 0$ else $\delta_e := 1$.
            $y_{e,i} := y_{e,i} + \delta_e$.
            $\omega_e := \omega_e e^{\epsilon \frac{w(i,e)+\delta_e s(e,i)}{u(e)}}$.
            For $M \in \mathcal{M}$ with $i \in M$ do:
                $\mu_M := \mu_M e^{\epsilon c(M,i) \frac{l(e,i)-\delta_e v(e,i)}{U(M)}}$.
```

Global routing algorithm: main theorem

This is a fully polynomial approximation scheme for the primal-dual pair of LPs

Enhanced global routing algorithm

- Compute new Steiner tree for net N only if previous one is longer than $(1+\epsilon_1)z_N$, where z_N is a continuously updated lower bound.
- If a new Steiner tree has to be computed, a $(1+\epsilon_2)$-optimal one suffices.

Theorem

Let λ^* be the optimum LP value and $t\epsilon\lambda^* > \log(m+|\mathcal{M}|)$. Then the algorithm computes feasible primal and dual solutions, whose values differ by at most a factor

$$\frac{\epsilon(1+\epsilon)(1+\epsilon_1)(1+\epsilon_2)}{\epsilon(1-\epsilon(1+\epsilon)(1+\epsilon_1)(1+\epsilon_2)\lambda^*)\left(1-\frac{\log(m+|\mathcal{M}|)}{t\epsilon\lambda^*}\right)}$$

By choosing $\epsilon, \epsilon_1, \epsilon_2, t$ appropriately, we get a $(1+\epsilon_0)$-optimal solution in $\frac{2\ln(m+|\mathcal{M}|)}{\epsilon_0^2}$ iterations, for any $\epsilon_0 > 0$.

(Vygen [2004])

The fractional global routing algorithm (enhanced)

```
For $p := 1$ to $t$ do:
    For $i := 1$ to $k$ do:
        Let $\psi_{i,e}$ be defined as above.
        If $Y_i = \emptyset$ or $\sum_{e \in E(Y_i)} \psi_{i,e} > (1 + \epsilon_1) z_i$ then:
            Let $Y_i \in \mathcal{Y}_i$ with
                $\sum_{e \in E(Y_i)} \psi_{i,e} \le (1 + \epsilon_2) \min_{Y \in \mathcal{Y}_i} \sum_{e \in E(Y)} \psi_{i,e}$.
            Set $z_i := \sum_{e \in E(Y_i)} \psi_{i,e}$.
        UpdateVariables
        For $M \in \mathcal{M}$ and $j \in M$ do:
            $z_j := z_j + (1 + \epsilon_2) \mathcal{L}_j c(M, j) \left( \mu_M^{new} - \mu_M^{old} \right)$
```

Randomized rounding

Let (x, y, λ) be a fractional solution to the primal LP. Compute a rounded solution $(\hat{x}, \hat{y}, \hat{\lambda})$ as follows:

- choose $Y \in \mathcal{Y}_N$ as Y_N with probability $x_{N,Y}$ (independently for all $N \in \mathcal{N}$); then set $\hat{x}_{N,Y_N} := 1$ and $\hat{x}_{N,Y} := 0$ for $Y \in \mathcal{Y}_N \setminus \{Y_N\}$.
- Set $\hat{y}_{N,e} := \frac{y_{N,e}}{\sum_{Y \in \mathcal{Y}_N : e \in E(Y)} x_{N,Y}}$ if $e \in E(Y_N)$ and $\hat{y}_{N,e} := 0$ otherwise.
- Choose $\hat{\lambda}$ minimum possible such that $(\hat{x}, \hat{y}, \hat{\lambda})$ is a feasible solution to the primal LP.

Let $\Lambda \leq \frac{U(M)}{c(M,N) \sum_{e \in E(Y)} (l(e,N) + v(e,N))}$ for $N \in M \in \mathcal{M}$ and $Y \in \mathcal{Y}_N$, and $\Lambda \leq \frac{u(e)}{w(e,N) + s(e,N)}$ for $N \in \mathcal{N}$. Moreover, suppose that $|\mathcal{M}| + |E(G)| < e^{\lambda \Lambda}$.

Then $\hat{\lambda} \leq \lambda \left(1 + (e-1)\sqrt{\frac{\ln(|\mathcal{M}| + |E(G)|}{\lambda \Lambda}}\right)$.

(Vygen [2004])

The global routing algorithm in practice

- In practice, results are much better than theoretical performance guarantees. Usually 10–20 iterations suffice.
- Only few upper bounds are violated; these are corrected easily by *ripup-and-reroute*.
- Detailed routing can realize the solution well, due to excellent capacity estimations.
- Small integrality gap and approximate dual solution implies that an infeasibility proof can be found for most infeasible instances.
- First global routing algorithm to take into account coupling, timing, and power consumption directly. Provably near-optimal.

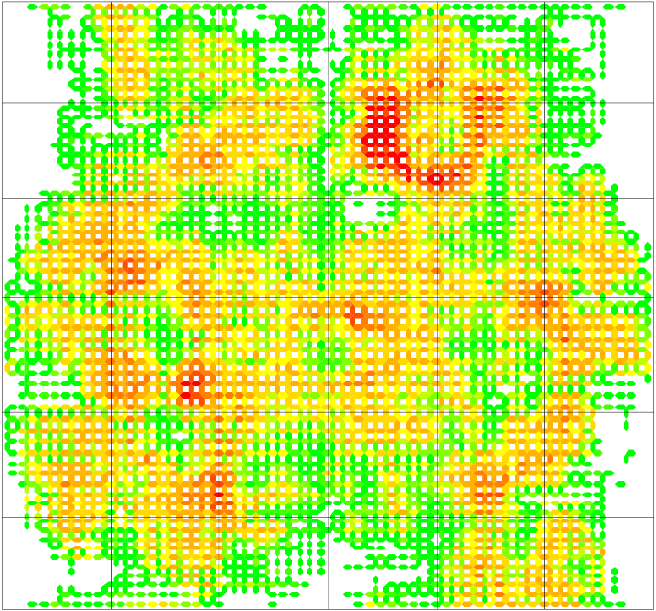
Connection to traffic flows

The global routing problem is equivalent to routing traffic flow

- with hard capacity bounds on edges (streets)
- without capacity bounds on vertices
- in a static setting (flow continuously repeated over time)
- with bounds on weighted sums of travel times
- and with the following transit time model: the transit time along an edge (latency) is constant up to $x\%$ congestion and grows linearly between $x\%$ and 100% congestion

Algorithm is equivalent to selfish routing but with taxes (exponential dependance on congestion)

Example: global routing congestion map



Future cost in global routing

The edge costs

$$\psi_{N,e} \quad := \quad \min_{\delta \in \{0,1\}} \Bigg((l(e,N) - \delta v(e,N)) \sum_{M \in \mathcal{M} : N \in M} c(M,N)\mu_M + (w(e,N) + \delta s(e,N))\omega_e \Bigg)$$

consist of a geometrical length part and a congestion part.

The future cost considers geometrical length only (ℓ_1-distance).

A suitable weighting of the geometrical part can speed up the algorithm considerably.

Future cost: observations in practice

Electrical characteristics or defect sensitivities are encoded in the geometrical part of the edge costs.

Thus future cost quality can degrade with increasing differences of these values

- over different planes
- between a spreaded and an unspreaded wire on the same plane

and also with increasing congestion.

Example

Edge lengths for yield optimization in a recent technology:

- M5 – M7: 1.0 (1 channel extra space), 1.37 (no extra space)
- M1 – M4: 1.76 (1 channel extra space), 2.73 (no extra space)

Future cost and RC-delay

Let N be a two-terminal net, and e an edge on some path connecting these terminals.

The contribution of e to the RC-delay on N is

$$r_e \left(\frac{c_e}{2} + C_e \right),$$

where

- r_e is the resistance of the edge e,
- c_e is its capacitance, and
- C_e is the downstream capacitance “hanging behind” e on the path.

For approximating C_e, the future cost can be used.

Yield analysis: critical area

Consider faults caused by particles with size distribution

$$f(r) := \begin{cases} 0, r < r_0 \\ \frac{c}{r^3}, r \geq r_0 \end{cases}$$

for some $r_0 \in \mathbb{R}_+$ smaller than the smallest possible particle that can cause a fault, and c such that $\int_0^\infty f(r)\mathrm{d}r = 1$.

Then the critical area w.r.t. extra material faults on plane z is

$$\mathrm{CA}_{em}^{z} := \int_x \int_y \int_{t_{em}(x,y,z)}^{\infty} f(r)\mathrm{d}r\mathrm{d}y\mathrm{d}x,$$

where $t_{em}(x, y, z)$ is the smallest size of a particle that causes an extra material fault at location (x, y, z).

Yield analysis: expected number of faults

Weighted sum of critical areas is used to estimate the number of extra material faults per chip:

$$\mathsf{F}_{em} := \sum_{z} w_{em}^{z} \mathsf{CA}_{em}^{z}$$

Analogously define the number of miss material faults on wire planes, F_{wm}, and on via planes, F_{vm}.

Define the estimated total number of faults per chip as $\mathsf{F} := \mathsf{F}_{em} + \mathsf{F}_{wm} + \mathsf{F}_{vm}$.

The percentage of chips without a fault from one of the above classes is estimated by

$$e^{-\mathsf{F}}.$$

The complement $1 - e^{-\mathsf{F}}$ is called the wiring yield loss.

Experimental results: the testbed

Chip	Technology	Image Size (in 1000 channels)	# Nets (in 1000)
Edgar	Cu08	40 x 40	772
Hannelore	Cu08	36 x 33	140
Paul	Cu08	24 x 24	68
Monika	Cu11	35 x 35	1502
Ralf	Cu11	26 x 26	1349
Garry	Cu11	26 x 26	827
Heidi	Cu11	23 x 23	777
Elena	Cu11	19 x 19	421
Lotti	Cu11	14 x 14	132
Dieter	Cu11	19 x 19	58
Ingo	Cu11	19 x 19	58
Bill	Cu11	26 x 26	11
Roland	Cu11	16 x 16	11
Joachim	SA27E	14 x 14	288

Experimental results: total running time (in seconds)

Chip	2D-GR	3D-GR, Netl. Opt.		3D-GR, Yield Opt.	
Edgar	63421	57096	(−10.0%)	91215	(+43.8%)
Hannelore	10847	12766	(+17.7%)	14552	(+34.2%)
Paul	4076	6019	(+47.7%)	5413	(+32.8%)
Monika	65064	62560	(−3.8%)	92995	(+42.9%)
Ralf	61473	55506	(−9.7%)	116221	(+89.1%)
Garry	48382	40399	(−16.5%)	70615	(+46.0%)
Heidi	31431	25936	(−17.5%)	45150	(+43.6%)
Elena	21197	20924	(−1.3%)	38327	(+80.8%)
Lotti	3978	5425	(+36.4%)	5895	(+48.2%)
Dieter	12705	11063	(−12.9%)	11152	(−12.2%)
Ingo	20733	11125	(−46.3%)	15661	(−24.5%)
Bill	4994	3924	(−21.4%)	5448	(+9.1%)
Roland	2528	3025	(+19.7%)	4200	(+66.1%)
Joachim	7432	9024	(+21.4%)	9526	(+28.2%)
Total	358591	325343	(−9.3%)	526819	(+46.9%)

Experimental results: wirelength

Chip	2D-GR	3D-GR, Netl. Opt.		3D-GR, Yield Opt.	
Edgar	211.656 m	212.022 m	(+0.2%)	214.162 m	(+1.2%)
Hannelore	30.110 m	30.239 m	(+0.4%)	31.006 m	(+3.0%)
Paul	9.888 m	9.903 m	(+0.2%)	9.999 m	(+1.1%)
Monika	263.936 m	264.123 m	(+0.1%)	273.793 m	(+3.7%)
Ralf	234.747 m	234.169 m	(–0.2%)	242.094 m	(+3.1%)
Garry	221.950 m	221.989 m	(+0.0%)	227.186 m	(+2.4%)
Heidi	150.775 m	150.863 m	(+0.1%)	153.837 m	(+2.0%)
Elena	92.234 m	92.226 m	(–0.0%)	94.511 m	(+2.5%)
Lotti	18.208 m	18.230 m	(+0.1%)	18.679 m	(+2.6%)
Dieter	13.226 m	13.329 m	(+0.8%)	13.574 m	(+2.6%)
Ingo	13.199 m	13.285 m	(+0.7%)	13.482 m	(+2.1%)
Bill	23.312 m	23.356 m	(+0.2%)	23.542 m	(+1.0%)
Roland	17.351 m	17.397 m	(+0.3%)	17.595 m	(+1.4%)
Joachim	62.250 m	62.432 m	(+0.3%)	63.721 m	(+2.4%)
Total	1363.675 m	1364.404 m	(+0.1%)	1398.024 m	(+2.5%)

Experimental results: number of vias

Chip	2D-GR	3D-GR, Netl. Opt.		3D-GR, Yield Opt.	
Edgar	6151607	6114859	(–0.6%)	8302895	(+35.0%)
Hannelore	795855	804856	(+1.1%)	1096198	(+37.7%)
Paul	474376	449112	(–5.3%)	606733	(+27.9%)
Monika	9335637	8916882	(–4.5%)	12409600	(+32.9%)
Ralf	10314838	9250179	(–10.3%)	12945468	(+25.5%)
Garry	6018048	5740090	(–4.6%)	8555230	(+42.2%)
Heidi	5030429	4790479	(–4.8%)	6821014	(+35.6%)
Elena	2738929	2689970	(–1.8%)	3486325	(+27.3%)
Lotti	669582	649336	(–3.0%)	797861	(+19.2%)
Dieter	426860	421537	(–1.2%)	537206	(+25.9%)
Ingo	441647	429608	(–2.7%)	586823	(+32.9%)
Bill	103812	101471	(–2.3%)	185742	(+78.9%)
Roland	95847	102976	(+7.4%)	191646	(+99.9%)
Joachim	1924130	1937133	(+0.7%)	2026975	(+5.3%)
Total	44594645	42470739	(–4.8%)	58623918	(+31.5%)

Experimental results: expected number of faults per chip

Chip	2D-GR	3D-GR, Netl. Opt.		3D-GR, Yield Opt.	
Edgar	0.09780	0.10493	(+7.3%)	0.08586	(–12.2%)
Hannelore	0.01396	0.01543	(+10.6%)	0.01027	(–26.4%)
Paul	0.00502	0.00568	(+13.2%)	0.00402	(–19.9%)
Monika	0.08744	0.09505	(+8.7%)	0.08055	(–7.9%)
Ralf	0.07832	0.08920	(+13.9%)	0.07361	(–6.0%)
Garry	0.07224	0.08017	(+11.0%)	0.06714	(–7.1%)
Heidi	0.05351	0.05804	(+8.5%)	0.04965	(–7.2%)
Elena	0.03167	0.03314	(+4.6%)	0.02966	(–6.3%)
Lotti	0.00658	0.00688	(+4.5%)	0.00575	(–12.6%)
Dieter	0.00482	0.00516	(+7.2%)	0.00416	(–13.6%)
Ingo	0.00457	0.00505	(+10.4%)	0.00392	(–14.2%)
Bill	0.00707	0.00833	(+17.8%)	0.00376	(–46.8%)
Roland	0.00563	0.00605	(+7.5%)	0.00396	(–29.7%)
Joachim	0.00432	0.00440	(+1.9%)	0.00431	(–0.1%)
Total	0.47336	0.51791	(+9.4%)	0.42703	(–9.8%)

The wiring yield loss is reduced by more than 10 % for most chips.

Introduction

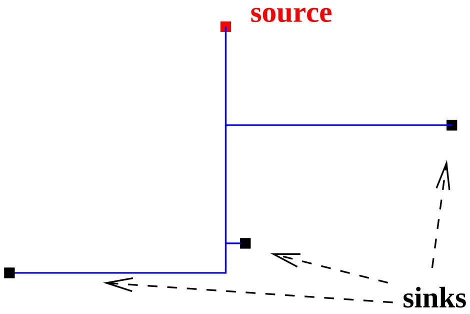
Placement

Routing

Timing Optimization
- Inverter Trees
- Logic Optimization
- Gate Sizing
- Overall Timing Optimization

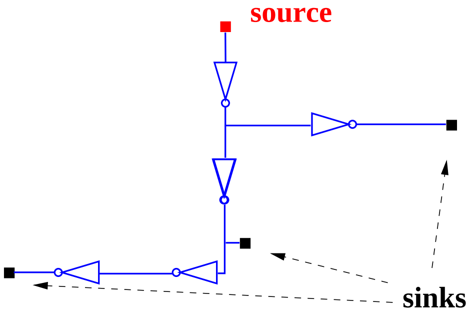
Clocktree Design

The repeater tree problem



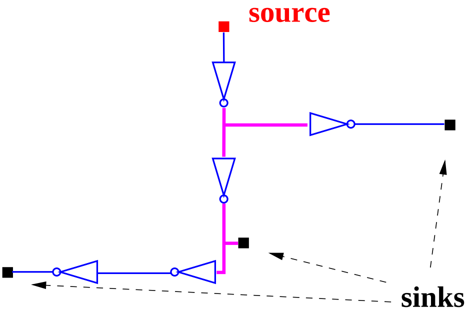
- A signal has to be distributed from a source to a set of sinks.
- The delay on a source-sink path increases
    - quadratically in the path length within the tree.

The repeater tree problem



- A signal has to be distributed from a source to a set of sinks.
- The delay on a source-sink path increases
  - linearly in path length (assuming ideal repeater insertion).

The repeater tree problem



- A signal has to be distributed from a source to a set of sinks.
- The delay on a source-sink path increases
  - linearly in path length (assuming ideal repeater insertion),
  - with every bifurcation on the path.

Importance of repeater trees

- As feature sizes decrease, the wire resistances increase.
- More and more repeaters are needed:
  - $10 - 20\%$ repeaters in 130nm technology
  - $20 - 30\%$ repeaters in 90nm technology
  - $30 - 40\%$ repeaters in 65nm technology
- The speed, robustness and power consumption depend heavily on repeater insertion algorithms.
- Up to 30 Mio. instances are solved during timing closure.
  $\Rightarrow$ algorithms must be fast.

The repeater tree problem

Instance:

- a root $r \in \mathbb{R}^2$,
- a finite set $S \subset \mathbb{R}^2$ of sinks,
- for each sink $s \in S$ a time interval $[t_{\min}(s), t_{\max}(s)]$ in which the signal must arrive at s, and a parity ($+$ or $-$),
- a library L of repeaters (inverters and buffers).

Task: Compute a repeater tree, i.e.

- an arborescence A rooted at r whose set of sinks is S,
- $\psi : V(A) \setminus (\{r\} \cup S) \to L \times \mathbb{R}^2$, and
- an arrival time t(r) at the root

such that $t(r) + delay_{(A,\psi)}(r, s) \in [t_{\min}(s), t_{\max}(s)]$ for all $s \in S$, the number of inversions is correct for each sink, and

- $t(r)$ is maximum or
- the power consumption is minimum.

Fast repeater trees

- In many cases, $t_{\min}(s) = -\infty$ for all $s \in S$.
- In this case, we cannot be too fast.
- Let $RAT(s) := t_{\max}(s)$ for $s \in S$.
- The slack at s is $RAT_s - t(r) - delay(r, s)\}$.
- Maximizing $t(r)$ is equivalent to maximizing the worst slack w.r.t. $t(r) = 0$, i.e.

$$\sigma_r := \min_{s \in S}\{RAT_s - delay(r, s)\}$$

Previous work

- Repeater insertion into given topology and a finite number of admissible locations $\mathcal{L}$.
  - Dynamic Programming with $O(|\mathcal{L}|^2)$ running time (van Ginneken [1990])
  - Running time was improved to $O(|\mathcal{L}| \log |\mathcal{L}|)$ (Shi, Li [2003,2005])
- No satisfying solution exists for topology generation:
  - optimum Steiner trees.
    Minimum power but poor delays due to long paths.
  - Bounded radius Steiner trees.
  - Heuristical splitting into critical and non-critical sub-trees.

Fast and short repeater trees: summary

In many cases, $t_{\min}(s) = -\infty$ for all $s \in S$.
A rough model for delay is

$$delay_{(A,\psi)}(r,s) = \sum_{(v,w)\in A[r,s]} \left(dist(v,w) + (|\delta^+(v)| - 1) \right),$$

where *dist* denotes ℓ_1-distance, $A[r,s]$ is the r-s-path in A.
A rough model for power consumption is the total wirelength:
$\sum_{(v,w)\in E(A)} dist(v,w)$

Fact 1: Huffman code yields minimum latency (maximum $t(r)$), but with length $\sum_{s\in S} dist(r,s)$.
Fact 2: Starting with an isolated root and successively inserting a closest sink is a $\frac{3}{2}$-approximation for the Steiner tree problem.

Fast and short repeater trees: summary

Proposed Algorithm:

- Sort the sinks by $t_{\max}(s) - \mathit{dist}(r, s)$, in nondecreasing order.
- Start by connecting the first sink to the root.
- Then successively insert the sinks in the above order. Insert s into edge $e \in E(A)$ such that $\min\{t_{\max}(s') - \mathit{delay}_A(r, s') : s' \in V(A)\}$ is maximum, or the total length is minimum, or a linear combination.

Theorem

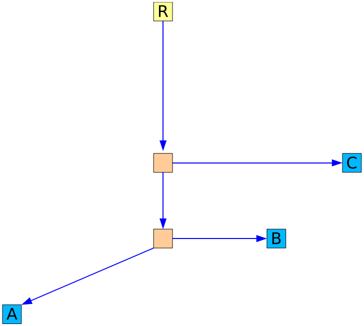
If all distances are zero, this also results in the optimum latency, namely with $t(r) = -\left\lceil \log_2 \left(\sum_{s \in S} 2^{-t_{\max}(s)} \right) \right\rceil$.

Experimental results show that in average, these trees are 0.66% longer or 0.22ps worse than the optimum.
(Bartoschek, Held, Rautenbach, Vygen [2006])

Definition (topology)

A topology T is an arborescence rooted at r, whose set of leaves is a subset of S, whose internal nodes are points in $\mathbb{R}^2$, and with $\delta^+(r) = 1$ and $\delta^+(u) = 2$ for all internal nodes u.

Example:



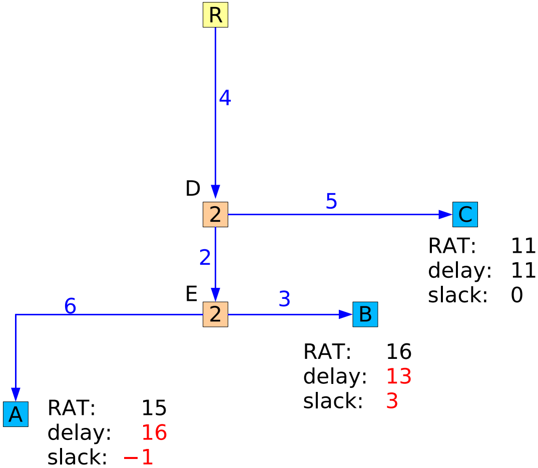
Delay model

The delay from r to a sink s is modeled as:

$$c_{node} \cdot (|E(T_{[r,s]})| - 1) \ + \sum_{(u,v) \in E(T_{[r,s]})} c_{wire} \cdot dist(u, v)$$

- c_{node}: delay penalty for bifurcation
- c_{wire}: delay per unit length
- Typical values are $c_{node} = 10 - 20$ ps and $c_{wire} = 100 - 200$ ps/mm.

Delay model: example



RAT: 11
delay: 11
slack: 0

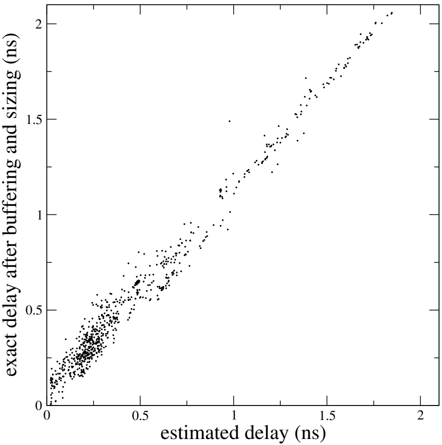
RAT: 16
delay: 13
slack: 3

RAT: 15
delay: 16
slack: −1

$c_{wire} = 1, c_{node} = 2.$

Justification of delay model

Relation between critical path delays in our model (estimated delay) and after repeater insertion and exact timing analysis:



Bounds on slack and wirelength

Lower bound on wirelength

A lower bound on the wire length is given by an optimum Steiner tree.

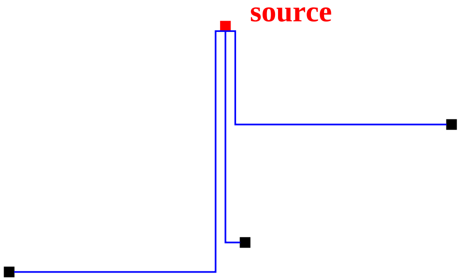
Upper bound on slack: Theorem

The maximum possible slack σ_{max} with respect to our delay model is at most:

$$-c_{node} \cdot \log_2 \left(\sum_{s \in S} 2^{-\left(\frac{RAT_s - c_{wire} dist(r,s)}{c_{node}} \right)} \right).$$

Proof.

The maximum possible slack can be obtained by a topology T where all internal nodes are at the root location:



All distance delays are minimum: $c_{wire} \cdot dist(r, s)$ for all $s \in S$.
In other words, we may assume $c_{wire} = 0$.

Proof (continued).

- Kraft's inequality: There exists a rooted binary tree with n leaves at depths $l_1, l_2, \ldots, l_n \Leftrightarrow$

$$\sum_{i=1}^{n} 2^{-l_i} \leq 1.$$

- Slack at root σ_r is minimum over all sink slacks $\Rightarrow$

$$delay(r,s) = c_{node} \cdot (|E(T_{[r,s]})| - 1) \leq RAT_s - \sigma_r \qquad \forall s \in S.$$

$\Longrightarrow$ The maximum slack achievable by any topology is bounded by

$$\max \left\{ \sigma \left| \sum_{s \in S} 2^{\frac{-RAT_s + \sigma}{c_{node}}} \leq 1 \right. \right\} = -c_{node} \cdot \log_2 \left(\sum_{s \in S} 2^{-\frac{RAT_s}{c_{node}}} \right)$$

$\square$

Improving the bound

Drawbacks of closed formula

- Closed formula ignores discrete structure of the problem
- Computation creates numerical problems

Huffman coding

- No closed formula
- Slightly better bounds
- Numerical stable and linear-time computation

Topology generation algorithm

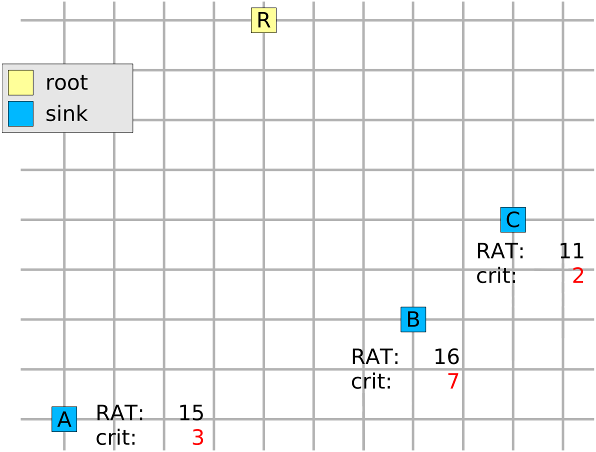
- Define criticality of $s \in S$ by $RAT_s - c_{wire} \cdot dist(r, s)$
- Start with partial topology $T' = (r, \emptyset)$;
- Connect most critical sink $s \in S$ to r.
- While unconnected sinks exist do:

```
    Choose most critical unconnected sink $s \in S \setminus V(T')$.
    Connect $s$ to an arc $e = (u, v) \in E(T')$ such that
        $\xi \cdot \sigma_e + (\xi - 1) \cdot c_{wire} \cdot dist(s, Area(e))$
    is maximized.
```

where

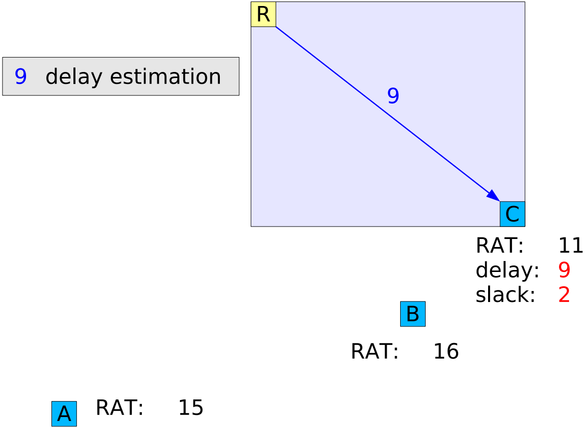
- ξ is a parameter weighing the objectives timing and power,
- σ_e is the slack at the root after connecting s to e.
- $Area(e)$ is the area covered by the union of all shortest u-v-paths.

Topology generation: example



$c_{wire} = 1, c_{node} = 2$

Topology generation: example



9 delay estimation

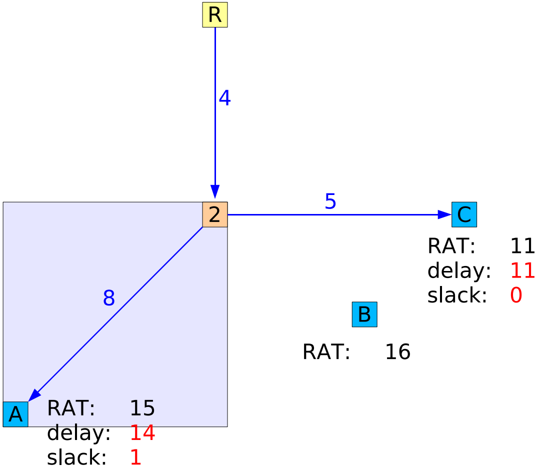
RAT: 11
delay: 9
slack: 2

RAT: 16

RAT: 15

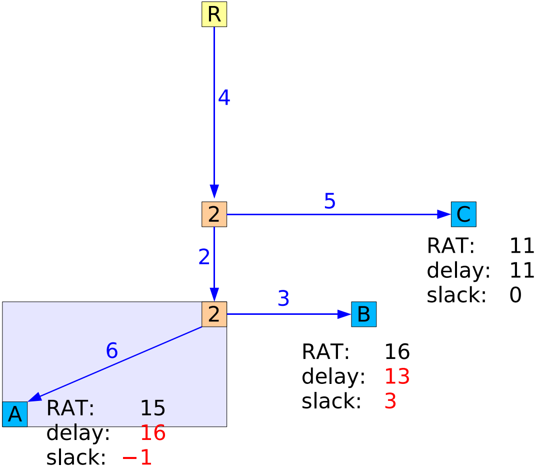
$c_{wire} = 1, c_{node} = 2$

Topology generation: example



$c_{wire} = 1, c_{node} = 2$

Topology generation: example



$c_{wire} = 1, c_{node} = 2$

Theorem

For $c_{wire} = 0$, $c_{node} = 1$, $\xi > 0$ and integer values for RAT_s, $s \in S$, the algorithm generates a topology that realizes the maximum possible slack

$$-\left\lceil \log_2 \left(\sum_{s \in S} 2^{-RAT_s} \right) \right\rceil .$$

Proof.

Induction on the number of sinks.
Assume the sinks in $S' \subset S$ are already connected optimally in T'.
Let $s' \in S \setminus S'$.

- If all $s \in S'$ have the same slack $\sigma_{S'}$ in T':
  - They are connected at maximum possible slack.
  - The best possible slack for the set $S' \cup \{s'\}$ equals $\sigma_{S'} + 1$.
  - s' can be connected to any existing edge in T' such that its slack is $\leq \sigma_{S'} + 1$.
- Otherwise s' can be connected to any non-critical edge. $\square$

Prim heuristic for Steiner trees

Wire length minimization ($\xi = 0$):

- Instead of choosing next critical sink:
- Choose sink, which is closest to the preliminary topology T'.
- Well known heuristic existing in many variants.

Hwang's theorem $\Longrightarrow$ $\frac{3}{2}$-approximation algorithm for the Steiner tree problem.

Running time

The running time is $O(|S|^2 \cdot \Psi)$, where Ψ is the running time for computing the union of all shortest paths between a sink and a union of paths. ($\Psi = 1$ for ℓ_1-distances)

Handling large instances

- Pre-clustering if $|S| > 10\,000$
- Facility location approximation (Massberg, Vygen [2005])
- Runtime: $O(|S| \log |S|)$

Experimental results

- 2.3 million instances with up to 10 000 sinks were taken from current 90nm designs.
- The extreme cases $\xi \in \{0, 1\}$ are compared against
    1. Length bound (optimum Steiner tree for $|S| \leq 30$, heuristic for $|S| > 30$).
    2. Slack bound (Huffman coding).
- 4.6 million topologies were computed in ≤ 100 seconds on a 2.6 GHz Opteron.

Results of topology generation

		Wirelength optimization: $\xi = 0$				Slack optimization: $\xi = 1$			
		Wirelength Deviation (%)		Slack Deviation (ps)		Wirelength Deviation (%)		Slack Deviation (ps)	
# Sinks	# Instances	avg.	worst	avg.	worst	avg.	worst	avg.	worst
1	1547517	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	319759	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	165448	0.00	0.00	13.89	82.72	12.19	99.60	0.12	20.00
4	86377	0.16	19.65	23.72	312.98	10.93	190.27	0.27	40.00
5	44301	0.16	21.51	33.40	174.51	14.01	188.15	0.34	52.45
6	27854	0.28	23.84	41.92	118.27	14.38	268.06	1.04	52.93
7	20523	0.45	22.24	52.19	285.43	22.26	248.77	0.42	52.51
8	19300	0.44	30.73	64.01	332.29	19.39	268.49	2.08	69.13
9	11085	0.81	26.26	71.11	465.77	29.58	250.04	3.36	60.00
10	11942	0.74	28.68	76.46	367.39	23.61	296.47	1.45	54.87
11-20	38184	1.60	28.00	101.16	427.25	32.57	426.68	1.73	76.80
21-30	11104	3.20	30.80	144.27	520.00	35.86	805.45	2.51	84.18
31-50	8647	2.99	33.16	226.05	793.70	70.29	1091.17	6.55	161.81
51-100	6621	4.06	26.34	344.88	1486.06	105.90	1782.56	12.23	203.48
101-200	1863	5.82	16.91	606.26	2019.90	135.84	1498.34	19.78	351.25
201-500	824	6.22	24.00	920.37	3711.47	209.77	2127.34	26.91	304.92
501-1000	205	7.62	19.40	1686.15	3563.61	569.58	2242.49	48.57	257.65
> 1000	31	6.99	14.74	2929.08	7872.96	211.40	1124.99	17.78	89.88
Total	2321585	**0.66**	33.16	**9.92**	7872.96	**19.35**	2242.49	**0.21**	351.25
> 2 sinks	774068	**1.31**	33.16	**50.69**	7872.96	**38.34**	2242.49	**1.08**	351.25

Results from 90 nm technology; $c_{node} = 20$

Buffering a topology: inserting repeaters

- Repeaters are inserted bottom-up whenever the optimum load capacitance is reached
- Special care has to be taken for merging branches that require different parity
- 4.4 Mio. trees are constructed in less than 10 minutes
- Repeater sizes are refined by global gate sizing
- Future work: unbalanced distribution of c_{node}
- Future work: better use of different wiring planes and wire widths
- Future work: better consideration of blockages and wiring congestion (using shortest path computations in grid graphs)

Further topics in timing optimization

- Logic optimization (fanin trees)
- Gate sizing, V_t-assignment
- Overall timing optimization: On the largest designs, the core timing optimization (inverter trees and gate sizing) runs only 6 hours instead of 3 days.
- The total runtime for timing closure (several iterations of placement and timing optimization) was reduced from more than a week to 26 hours.

Introduction

Placement

Routing

Timing Optimization

Clocktree Design
    Clock Skew Scheduling
    Clock Tree Synthesis
    Facility Location and Network Design with Service Capacities

Slack distribution

Instance:

- directed graph G, $c : E(G) \to \mathbb{R}$
- a set $F_0 \subseteq E(G)$ and a partition $\mathcal{F}$ of $E(G) \setminus F_0$
- weights $w : E(G) \setminus F_0 \to \mathbb{R}_{>0}$

Task: Find a potential $\pi : V(G) \to \mathbb{R}$ with
$c_\pi(e) := c(e) + \pi(x) - \pi(y) \geq 0$ for $e = (x, y) \in F_0$
such that the vector

$$\left(\min \left\{ \frac{c(e) + \pi(x) - \pi(y)}{w(e)} : e = (x, y) \in F \right\} \right)_{F \in \mathcal{F}}$$

(after sorting entries in non-decreasing order) is lexicographically maximal.

Slack distribution: optimum balance

Theorem

Let $\pi : V(G) \to \mathbb{R}$ *with* $c_\pi(e) \geq 0$ *for* $e \in F_0$. *Let*

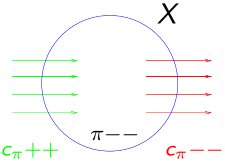
$$E_\pi := \left\{ f \in F \in \mathcal{F} : \frac{c_\pi(f)}{w(f)} \text{ minimal in } F \right\}.$$

Then π *is an optimum solution with* E_π *minimal iff for each* $X \subset V(G)$:

$$\min_{e \in E_\pi \cap \delta^-(X)} \frac{c_\pi(e)}{w(e)} \geq \min_{e \in E_\pi \cap \delta^+(X)} \frac{c_\pi(e)}{w(e)}$$

or

$$\min_{e \in \delta^+(X) \cap F_0} c_\pi(e) = 0.$$



(Vygen [2003])

Algorithms for slack distribution

- unit weights: $O(mn + n^2 \log n)$ (Albrecht, Korte, Schietke, Vygen [2002], generalizing Schneider, Schneider [1991], Young, Tarjan, Orlin [1991])
- general weights: $O(\min\{\ n^4 \log^2 n + n^2 m \log m,\ n^4 \log n + n^2 m \log^2 n \log\log n,\ w_{\max}(mn + n^2 \log n)\})$ (Held [2001])
- running times much better in practice, in particular when ignoring large slacks

Application to clock skew scheduling:

first maximize late-slacks, then early-slacks, then length of time intervals for clock signals (Albrecht, Korte, Schietke, Vygen [2002], Held, Maßberg, Korte, Ringe, Vygen [2003])

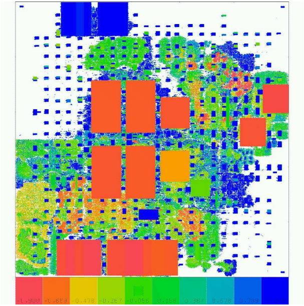
Clock skew scheduling: the critical cycle

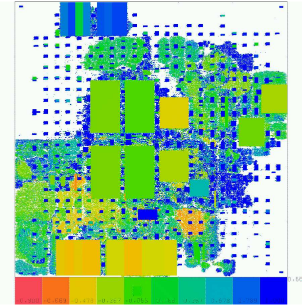
Looking for the first (most critical) entry of the slack vector only reduces the minimum mean mean cycle problem (Karp [1978]) or minimum ratio problem (Megiddo [1983], Radzik [1993])

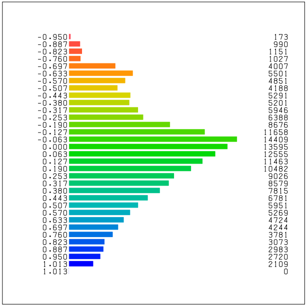
Theorem

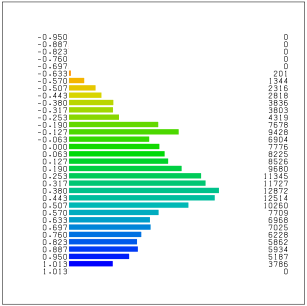
Suppose the whole chip runs at the same frequency. Then the best cycle time that can be achieved by clock skew scheduling equals the maximum mean weight (delay) of a cycle in the latch graph.

Clock skew scheduling: example

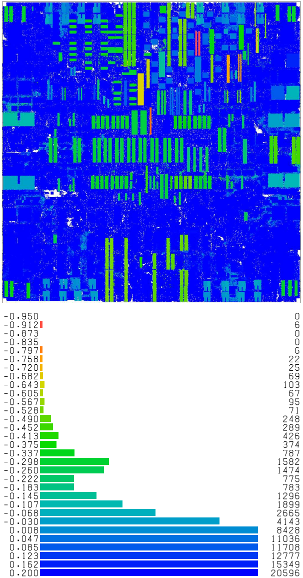




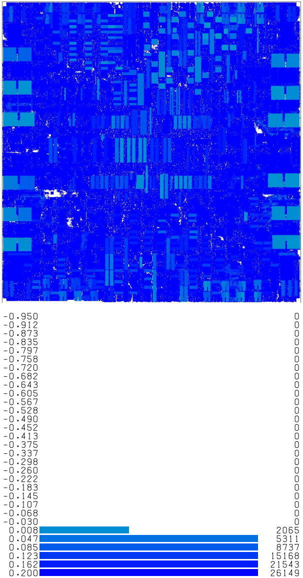




Clock skew scheduling: a high-end ASIC

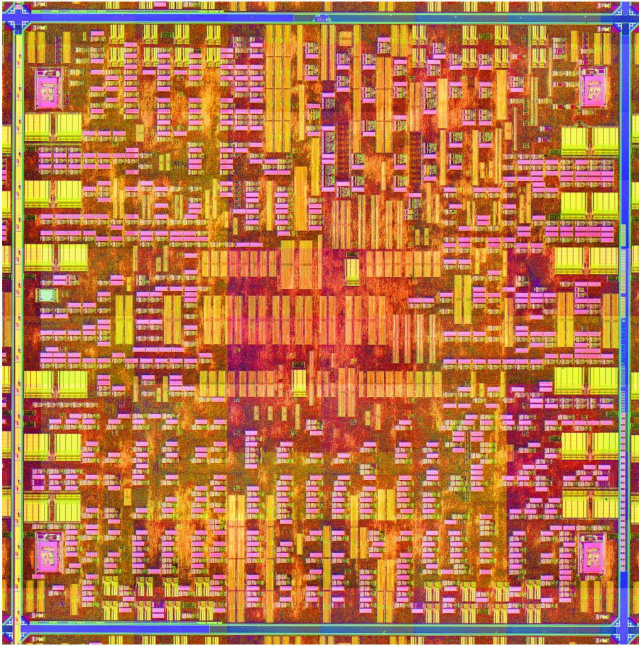


701 MHz with zero skew

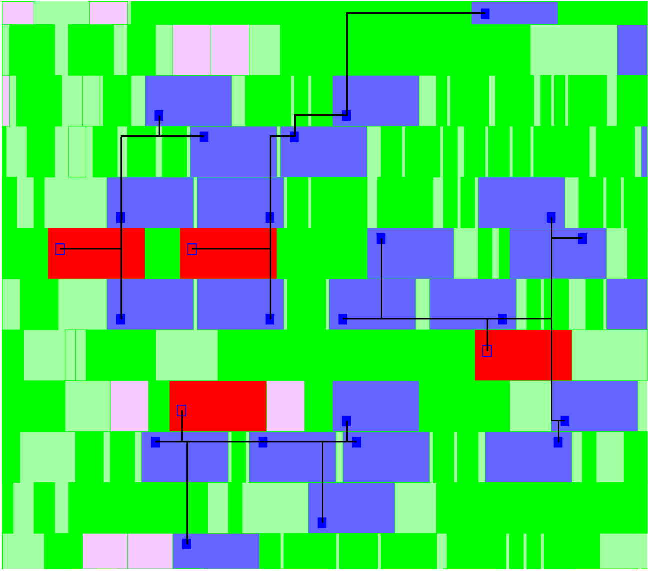


900 MHz with BonnClock
1033 MHz in hardware

The same ASIC in hardware



Distributing a signal to several terminals

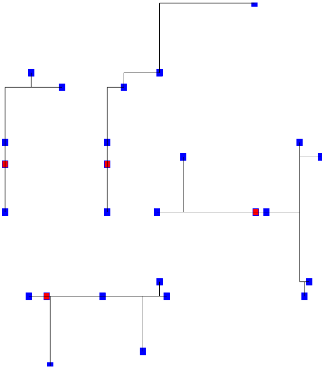


blue: terminals red: facilities

Facility location and network design with service capacities

Instance:

- metric space (V, c),
- finite set $\mathcal{D} \subseteq V$ (terminals/clients),
- demands $d : \mathcal{D} \to \mathbb{R}_+$,
- facility opening cost $f \in \mathbb{R}_+$,
- capacity $u \in \mathbb{R}_+$.



Find a partition $\mathcal{D} = D_1 \dot{\cup} \cdots \dot{\cup} D_k$ and Steiner trees T_i for D_i $(i = 1, \ldots, k)$ with

$$c(E(T_i)) + d(D_i) \leq u$$

for $i = 1, \ldots, k$ such that

$$\sum_{i=1}^{k} c(E(T_i)) + kf$$

is minimum.

Complexity results

(All the following results are by Maßberg and Vygen [2005])

Proposition

- *There is no* $(1.5 - \epsilon)$*-approximation algorithm (for any* $\epsilon > 0$*) unless* $P = NP$.
- *There is no* $(2 - \epsilon)$*-approximation algorithm (for any* $\epsilon > 0$*) for any class of metrics where the Steiner tree problem cannot be solved exactly in polynomial time.*
- *There is a 2-approximation algorithm for geometric instances (similar to Arora's approximation scheme for the TSP). However, this is not practically efficient.*

Lower bound: spanning forests

Let F_1 be a minimum spanning tree for $(\mathcal{D}, c)$.
Let $e_1, \ldots, e_{n-1}$ be the edges of F_1 so that $c(e_1) \geq \ldots \geq c(e_{n-1})$.
Set $F_k := F_{k-1} \setminus \{e_{k-1}\}$ for $k = 2, \ldots, n$.

Lemma

F_k is a minimum weight spanning forest in $(\mathcal{D}, c)$ with exactly k components.

Proof.

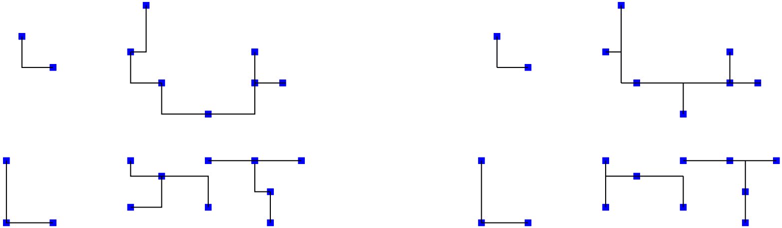
By induction on k. Trivial for $k = 1$. Let $k > 1$.
Let F^* be a minimum weight k-spanning forest.
Let $e \in F_{k-1}$ such that $F^* \cup \{e\}$ is a forest. Then

$$c(F_k) + c(e_{k-1}) = c(F_{k-1}) \leq c(F^*) + c(e) \leq c(F^*) + c(e_{k-1}).$$

$\square$

Lower bound: Steiner forests

A k-Steiner forest is a forest F with $\mathcal{D} \subseteq V(F)$ and exactly k components.



Lemma

$\frac{1}{\alpha}c(F_k)$ is a lower bound for the cost of a minimum weight k-Steiner forest, where α is the Steiner ratio.

Lower bound: number of facilities

Let t' be the smallest integer such that

$$\frac{1}{\alpha} c(F_{t'}) + d(\mathcal{D}) \leq t' \cdot u$$

Lemma

t' is a lower bound for the number of facilities of any solution.

Let t'' be an integer in $\{t', \ldots, n\}$ minimizing

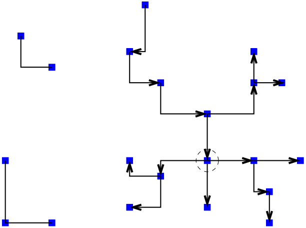
$$\frac{1}{\alpha} c(F_{t''}) + t'' \cdot f.$$

Theorem

$\frac{1}{\alpha} c(F_{t''}) + t'' \cdot f$ is a lower bound for the cost of an optimal solution.

Algorithm A

1. Compute a minimum spanning tree on $(\mathcal{D}, c)$.
2. Compute t'' and spanning forest $F_{t''}$ as above.
3. Split up overloaded components by a bin packing approach.



It can be guaranteed that for each new component at least $\frac{u}{2}$ of load will be removed from the initial forest.

Analysis of algorithm A

Recall: $\frac{1}{\alpha}c(F_{t''}) + t'' \cdot f$ is a lower bound for the optimum.

We set $L_r := \frac{1}{\alpha}c(F_{t''})$ and $L_f := t'' \cdot f$.

Observe: $L_r + d(\mathcal{D}) \leq \frac{u}{f}L_f$.

The cost of the final solution is at most

$$c(F_{t''}) + t''f + \frac{2}{u}\Big(c(F_{t''}) + d(\mathcal{D})\Big)f$$

$$= \alpha L_r + L_f + \frac{2f}{u}\big(\alpha L_r + d(\mathcal{D})\big)$$
$$\leq \alpha L_r + L_f + 2\alpha L_f$$

Theorem

Algorithm A is a $(2\alpha + 1)$*-approximation algorithm.*

Algorithm B

Define metric c' by $c'(v, w) := \min\{c(v, w), \frac{uf}{u+2f}\}$.

1. Compute a Steiner tree F for $\mathcal{D}$ in (V, c') with some β-approximation algorithm.
2. Remove all edges e of F with $c(e) \geq \frac{uf}{u+2f}$.
3. Split up overloaded components of the remaining forest as in algorithm A.

Theorem

Algorithm B has perfomance ratio 3β.

Using the Robins-Zelikovsky Steiner tree approximation algorithm we get a 4.648-approximation algorithm.

With a more careful analysis of the Robins-Zelikovsky algorithm we can get a 4.099-approximation algorithm in $O(n^{2^{10000}})$ time.

Algorithm C

Define metric c'' by $c''(v,w) := \min\{c(v,w), \frac{uf}{u+f}\}$

1. Compute a tour F for $\mathcal{D}$ in (V, c'') with some γ-approximation algorithm.
2. Remove the longest edge of F.
3. Remove all edges e of F with $c(e) \geq \frac{uf}{u+f}$.
4. Split up overloaded components of the remaining forest as in algorithm A.

Theorem

Algorithm C has perfomance ratio 3γ.

Using Christofides' TSP approximation algorithm we get a 4.5-approximation algorithm in $O(n^3)$ time.

Comparison of the three approximation algorithms

- Algorithm A computes a minimum spanning tree.
- Algorithm B calls the Robins-Zelikovsky algorithm.
- Algorithm C calls Christofides' algorithm.
- Then each algorithm deletes expensive edges and splits up overloaded components.

algorithm	metric	perf.guar.	runtime
A	$(\mathbb{R}^2, \ell_1)$	4	$O(n \log n)$
A	general	5	$O(n^2)$
B	general	4.099	$O(n^{2^{10000}})$
C	general	4.5	$O(n^3)$

Experimental results

Algorithm A on six real-world instances:

	inst1	inst2	inst3	inst4	inst5	inst6
# terminals	3675	17140	45606	54831	109224	119461
MST length	13.72	60.35	134.24	183.37	260.36	314.48
t'	117	638	1475	2051	3116	3998
L_r	8.21	31.68	63.73	102.80	135.32	181.45
$L_r + L_f$	23.07	112.70	251.06	363.28	531.05	689.19
# facilities	161	947	2171	2922	4156	5525
service cost	12.08	54.23	101.57	159.93	234.34	279.93
total cost	32.52	174.50	377.29	531.03	762.15	981.61
gap (factor)	1.41	1.55	1.59	1.46	1.44	1.42

Reduction of power consumption

Algorithm A on four chips, compared to the previously used heuristic:

chip	Jens	Katrin	Bert	Alex
technology	180nm	130nm	130nm	130nm
# clocktrees	1	3	69	195
total # sinks	3805	137265	40298	189341
largest instance	375	119461	16260	35305
power (W, old)	0.100	0.329	0.306	2.097
power (W, new)	0.088	0.287	0.283	1.946
difference	−11.1%	−12.8%	−7.5%	−7.2%

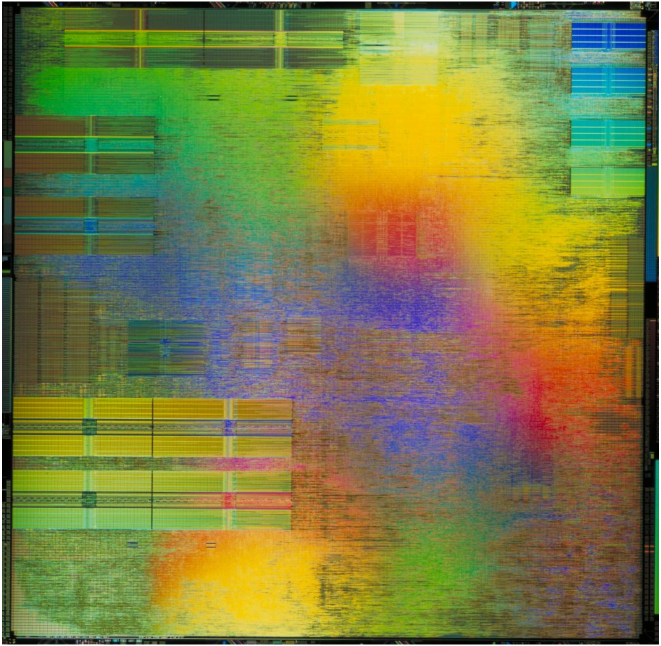
Some open problems

- Improve the approximation ratio for the problem with service capacities (in $(\mathbb{R}^2, \ell_1)$, with a practically efficient algorithm).
- In some real-world instances, there exists an interval graph on the terminals, and we have to partition this graph into cliques. Is there a constant factor approximation algorithm for the resulting problem? (We have a $\log n$-approximation algorithm.)
- What other interesting problems combining facility location with network design, or routing, can be approximated?
- What about multi-stage extensions?

Conclusion

- VLSI design is probably the richest application area of combinatorial optimization
- All classical and many new combinatorial optimization problems are directly applied
- Rapidly developping technology poses constantly new problems
- Instances sizes pose challenges to algorithm design and implementation
- Better chips by better mathematics
- There is still a lot to be done...

Thank you!



Some references for further reading

- Albrecht, C., Korte, B., Schietke, J., and Vygen, J. [2002]: Maximum mean weight cycle in a digraph and minimizing cycle time of a logic chip. Discrete Applied Mathematics 123 (2002), 103–127
- Bartoschek, C., Held, S., Rautenbach, D., and Vygen, J. [2006]: Efficient generation of short and fast repeater tree topologies. Proceedings of the International Symposium on Physical Design (2006), to appear
- Brenner, U., and Struzyna, M. [2005]: Faster and Better Global Placement by a New Transportation Problem. Proceedings of the 42nd IEEE/ACM Design Automation Conference (2005), 591–596
- Brenner, U., and Vygen, J. [2001]: Worst-case ratios of networks in the rectilinear plane. Networks 38 (2001), 126–139
- Brenner, U., and Vygen, J. [2004]: Legalizing a Placement with Minimum Total Movement. IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems 23 (2004), 1597–1613

- Held, S., Korte, B., Maßberg, J., Ringe, M., and Vygen, J. [2003]: Clock scheduling and clocktree construction for high performance ASICs. Proceedings of the IEEE International Conference on Computer-Aided Design (2003), 232–239
- Korte, B., and Vygen, J. [2006]: Combinatorial Optimization: Theory and Algorithms. Third edition. Springer, Berlin 2006
- Maßberg, J., and Vygen, J. [2005]: Approximation Algorithms for Network Design and Facility Location with Service Capacities. Proceedings of the 8th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX 2005); Springer, Berlin 2005, pp. 158-169
- Vygen, J. [2004]: Near-optimum global routing with coupling, delay bounds, and power consumption. Proceedings of the 10th International IPCO Conference; Springer, Berlin 2004, pp. 308–324
- Vygen, J. [2005]: Geometric quadrisection in linear time, with application to VLSI placement. Discrete Optimization 2 (2005), 362–390