Automated Layout and Migration in Ultra-Deep Submicron VLSI

June 25, 1999

Cyrus Bamji — Cadence Design Systems, Inc.
Maarten Berkens, Chris Strolenberg — Sagantec, Inc.
Andrew B. Kahng — UCLA CS Dept.

Cyrus Bamji

- Cyrus Bamji (b. Bombay, India) studied in Paris, then received his BS (1982), MS (1985) and Ph.D. (1989) from MIT in EECS. He joined Cadence Design Systems, Inc. after graduation and has been working there ever since. He worked for several years in the field of layout compaction. Currently, he is an Architect at Cadence, working in the field of timing optimization and characterization. Dr. Bamji has received 3 Best Paper awards, holds 7 patents and is also the author of a book on Compaction.

Maarten Berkens

- Maarten Berkens is currently CTO of Sagantec. Berkens re-joined Sagantec in 1987 and is responsible for the technical direction of the company and is currently working on next generation compaction technology. Prior to Sagantec he was at Philips managing the team responsible for developing non-volatile memories. Prior to Philips Berkens was at Sagantec working on design tools and migration technologies for libraries and logic blocks. Berkens graduated in 1984 from Eindhoven University of Technology in the Netherlands. His graduation topic in electronics engineering was related to circuit simulation. Berkens' scientific interests include deep submicron (physical) design problems, combinatorial optimization and linking this to real life applications.

Chris Strolenberg

- Chris Strolenberg received his masters degree in Computing Science from Eindhoven University of Technology in 1988, based on a method for regular datapath floorplanning. Since joining Sagantec in 1985, he has been involved in many aspects of VLSI design automation, including global routing, timing analysis and HDL-based design entry. Since 1993, Mr. Strolenberg has held the position of Product Architect for Sagantec's DREAM tool. His interests include migration and optimization techniques for deep submicron designs, and methods for productivity enhancement in layout design.

Andrew B. Kahng

- Andrew B. Kahng (b. Oct. 1963, San Diego, CA) received the A.B. degree in applied mathematics (physics) from Harvard College, and the M.S. and Ph.D. degrees in computer science from the University of California at San Diego. He joined the computer science faculty at UCLA in July 1989, and is currently professor and vice-chair for graduate studies. From April 1996 through September 1997, he was on sabbatical leave and leave of absence from UCLA, as a Visiting Scientist at Cadence Design Systems, Inc. Dr. Kahng has received NSF Research Initiation and Young Investigator awards, and a DAC Best Paper award. He was the founding General Chair of the 1997 ACM/IEEE International Symposium on Physical Design, and defined the physical design roadmap as a member of the Design Tools and Test working group for the 1997 renewal of the SIA NTRS. He is currently a member of the EDA Council's EDA 200X task force, and the Design Tools and Test working group for the 1999 SIA ITRS renewal. He co-organized and presented a DAC-94 tutorial.

Tutorial Overview

- UDSM technology trends and implications
  - new issues and problems in USDM design
  - current context: cell-based place-and-route
- New solutions: Custom layout design
- New solutions: Layout-level modifications for performance and yield
- Applications: Hard-IP reuse and optimization

Silicon Complexity and Design Complexity

- Silicon complexity: physical effects cannot be ignored
  - fast but weak gates; resistive and cross-coupled interconnects
  - subwavelength lithography from 350nm generation onward
  - delay, power, signal integrity, manufacturability, reliability all become first-class objectives along with area
- Design complexity: more functionality and customization, in less time
  - reuse-based design methodologies for SOC
- Interactions increase complexity
  - need robust, top-down, convergent design methodology

Guiding Philosophy in the Back-End

- Many opportunities to leave $$$ on table
  - physical effects of process, migratability
  - design rules more conservative, design waivers up
  - device-level layout optimizations in cell-based methodologies
- Verification cost increases
- Prevention becomes necessary complement to checking
- Successive approximation = design convergence
  - upstream activities pass intentions, assumptions downstream
  - downstream activities must be predictable
  - models of analysis/verification = objectives for synthesis
- More “custom” bias in automated methodologies

Overall Roadmap Technology Characteristics

YEAR OF FIRST PRODUCT SHIPMENT	1997	1999	2002	2005	2008	2011	2014
TECHNOLOGY NODE DENSE LINES (DRAM HALF-PITCH) (nm)	250	180	130	100	70	50	35
ISOLATED LINES (MPU GATES) (nm)	200	140	100	70	50	35	25
Logic (Low-Volume—ASIC)‡							
Usable transistors/cm2 (auto layout)	8M	14M	24M	40M	64M	100M	160M
Nonrecurring engineering cost /usable transistor (microcents)	50	25	15	10	5	2.5	1.3
Number of Chip I/Os – Maximum							
Chip-to-package (pads) (high-performance)	1515	1867	2553	3492	4776	6532	8935
Chip-to-package (pads) (cost-performance)	758	934	1277	1747	2386	3268	4470
Number of Package Pins/Balls – Maximum							
Microprocessor/controller (cost-performance)	568	700	957	1309	1791	2449	3350
ASIC (high-performance)	1136	1400	1915	2619	3581	4898	6700
Package cost (cents/pin) (cost-performance)	0.78-2.71	0.70-2.52	0.60-2.16	0.51-1.85	0.44-1.59	0.38-1.36	0.33-1.17
Power Supply Voltage (V)							
Minimum logic Vdd (V)	1.8–2.5	1.5–1.8	1.2–1.5	0.9–1.2	0.6–0.9	0.5–0.6	0.37-0.42
Maximum Power							
High-performance with heat sink (W)	70	90	130	160	170	175	183
Battery (W)—(Hand-held)	1.2	1.4	2	2.4	2.8	3.2	3.7

Overall Roadmap Technology Characteristics (Cont'd)

YEAR OF FIRST PRODUCT SHIPMENT	1997	1999	2002	2005	2008	2011	2014
TECHNOLOGY NODE DENSE LINES (DRAM HALF-PITCH) (nm)	250	180	130	100	70	50	35
Chip Frequency (MHz)							
On-chip local clock (high-performance)	750	1250	2100	3500	6000	10000	16903
On-chip, across-chip clock (high-performance)	375	1200	1600	2000	2500	3000	3674
On-chip, across-chip clock (high-performance ASIC)	300	500	700	900	1200	1500	1936
On-chip, across-chip clock (cost-performance)	400	600	800	1100	1400	1800	2303
Chip-to-board (off-chip) speed (high-performance, reduced-width, multiplexed bus)	375	1200	1600	2000	2500	3000	3674
Chip-to-board (off-chip) speed (high-performance, peripheral buses)	250	480	885	1035	1285	1540	1878
Chip Size (mm2) (@sample/introduction)							
DRAM	280	400	560	790	1120	1580	2240
Microprocessor	300	340	430	520	620	750	901
ASIC [max litho field area]	480	800	900	1000	1100	1300	1482
Lithographic Field Size (mm2)	22 x 22 484	25 x 32 800	25 x 36 900	25 x 40 1000	25 x 44 1100	25 x 52 1300	25 x 59 1482
Maximum Number Wiring Levels	6	6–7	7	7–8	8–9	9	10

Technology Scaling Trends

- Interconnect
  - Impact of scaling on parasitic capacitance
  - Impact of scaling on inductance coupling
  - Impact of new materials on parasitic capacitance & resistance
  - Trends in number of layers, routing pitch
- Device
  - V_{dd}, V_t, sizing
  - Circuit trends (multithreshold CMOS, multi-supply, dynamic CMOS)
  - Impact of scaling on power and reliability

Technology Scaling Trends

- Scaling of x0.7 every three years
  - .25u .18u .13u .10u .07u .05u
  - 1997 1999 2002 2005 2008 2011
  - 5LM 6LM 7LM 7LM 8LM 9LM
- Interconnect delay dominates system performance
  - consumes 70% of clock cycle
- cross coupling capacitance is dominating
  - cross capacitance -> 100%, ground capacitance ->0%
  - 90% in .18u
  - harms signal integrity

New Materials Implications

- Lower dielectric
  - Reduces total capacitance
  - Doesn't change cross/ground proportions
- Copper metallization
  - Reduces RC delay
  - Avoids electronmigration
  - Thinner deposition reduces cross cap
- Multi layers of routing
  - Relative routing pitch may increase
  - Room for shielding

Tutorial Overview

- UDSM technology trends and implications
  - new issues and problems in USDM design
  - current context: cell-based place-and-route
- New solutions: Custom layout design
- New solutions: Layout-level modifications for performance and yield
- Applications: Hard-IP reuse and optimization

UDSM Technology Trends and Implications

June 25, 1999

Session Overview

- New issues and problems arising in UDSM technology
  - catastrophic yield: critical area, antennas
  - parametric yield: density control (filling) for CMP
  - parametric yield: subwavelength lithography implications
    - optical proximity correction (OPC)
    - phase-shifting mask design (PSM)
  - signal integrity
    - crosstalk and delay uncertainty
    - DC electromigration
    - AC self-heat
    - hot electrons
- Current context: cell-based place-and-route methodology
  - placement and routing formulations, basic technologies
  - methodology contexts

Technical Issues in VDSM Design

- Manufacturability (chip can't be built)
  - antenna rules
  - minimum area rules for stacked vias
  - CMP (chemical mechanical polishing) area fill rules
  - layout corrections for optical proximity effects in subwavelength lithography; associated verification issues
- Signal integrity (failure to meet timing targets)
  - crosstalk induced errors
  - timing dependence on crosstalk
  - IR drop on power supplies
- Reliability (design failures in the field)
  - electromigration on power supplies
  - hot electron effects on devices
  - wire self heat effects on clocks and signals

Why Now?

- These effects have always existed, but become worse at UDSM sizes because of:
  - finer geometries
    - greater wire and via resistance
    - higher electric fields if supply voltage not scaled
  - more metal layers
    - higher ratio of cross coupling to grounded capacitance
  - lower supply voltages
    - more current for a given power
  - lower device thresholds
    - smaller noise margins

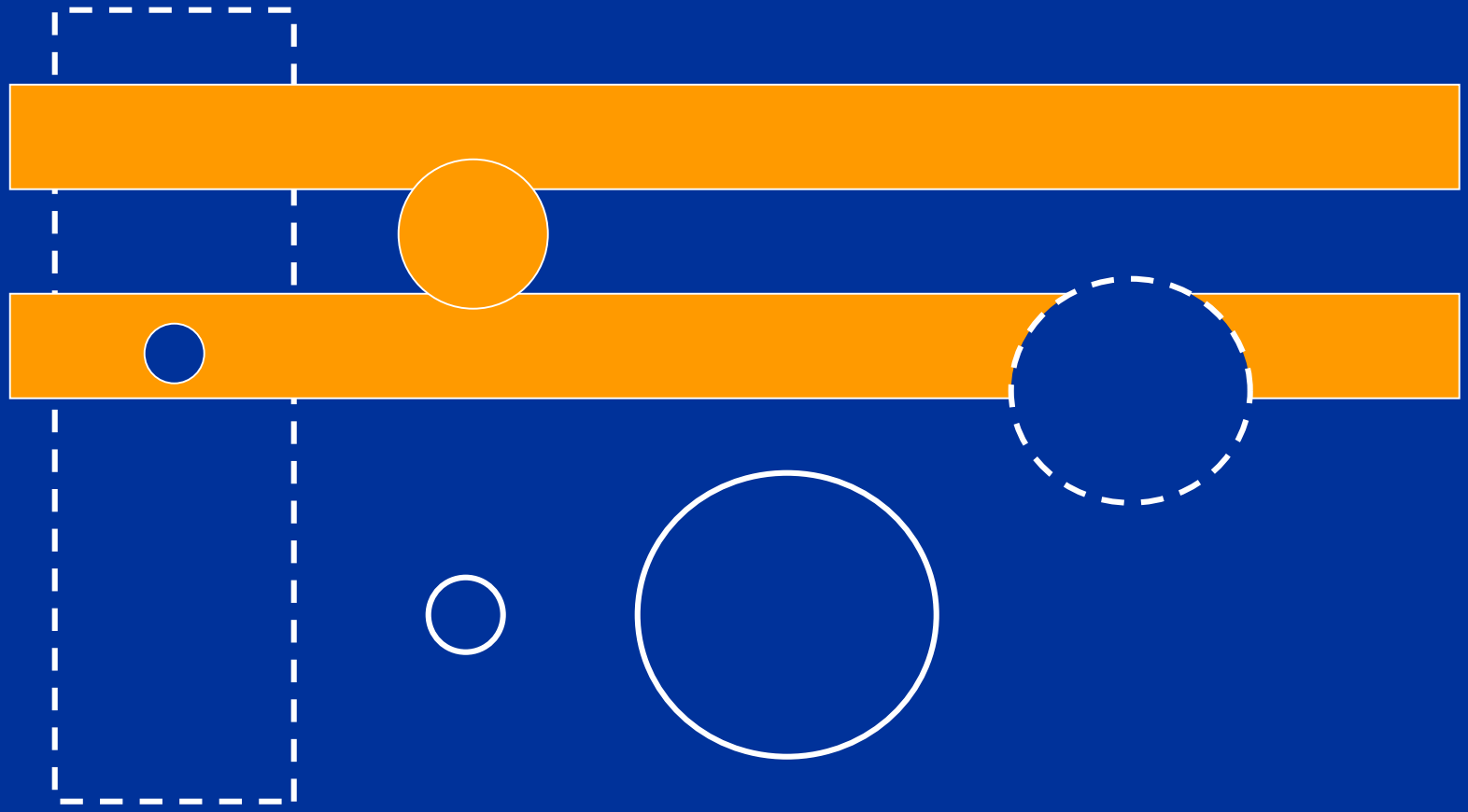
Why Now?

- Focus on interconnect
  - susceptible to patterning difficulties
    - CMP, optical exposure, resist development/etch, CVD, ...
  - susceptible to defects
    - critical area, critical volume

Defect-related Yield Loss

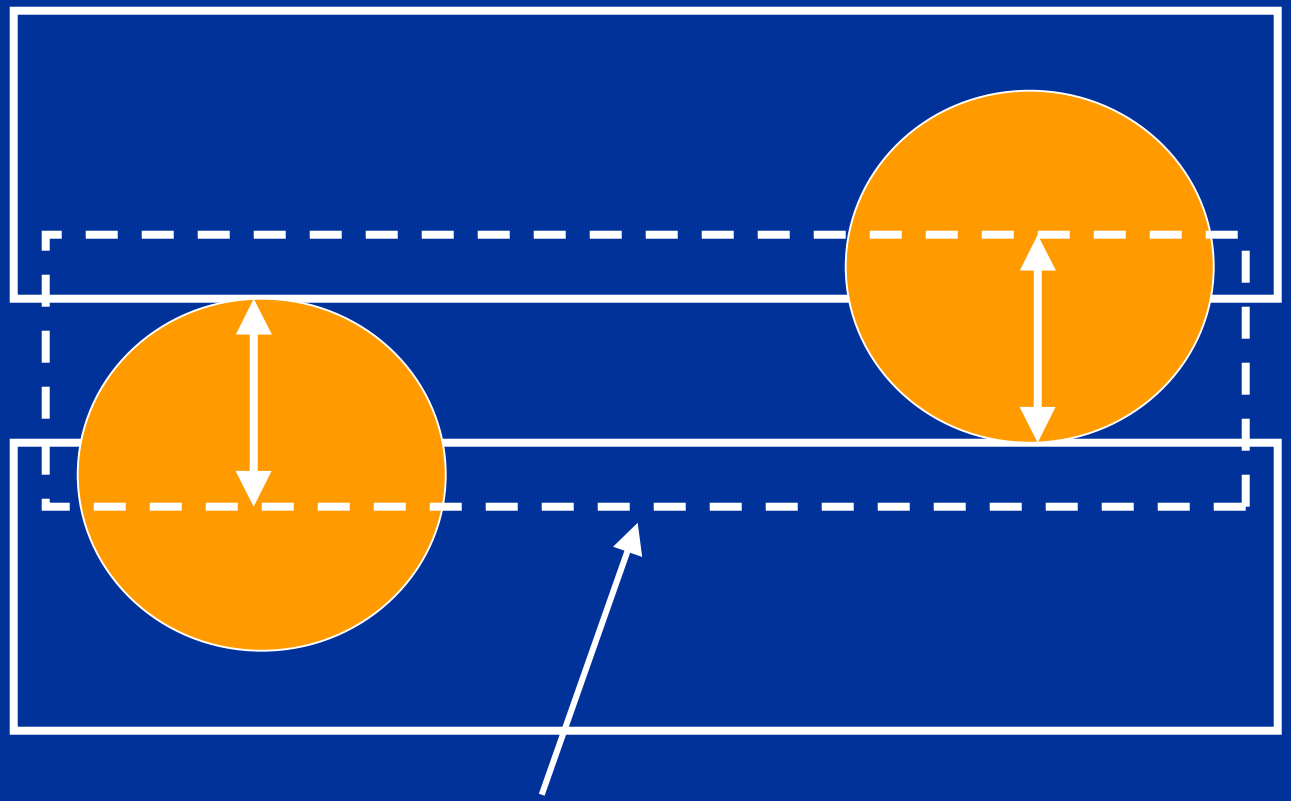
- High susceptibility to spot defect-related yield loss, particularly in metallization stages of process
- Most common failure mechanisms: shorts or opens due to extra or missing material between metal tracks
- Design tools fail to realize that values in design manuals are minimum values, not target values
- Spot defect yield loss modeling
  - extremely well-studied field
  - first-order yield prediction: Poisson yield model
  - critical-area model much more successful
  - fatal defect types (two types of short circuits, one type of open)

Defect-related Yield Loss

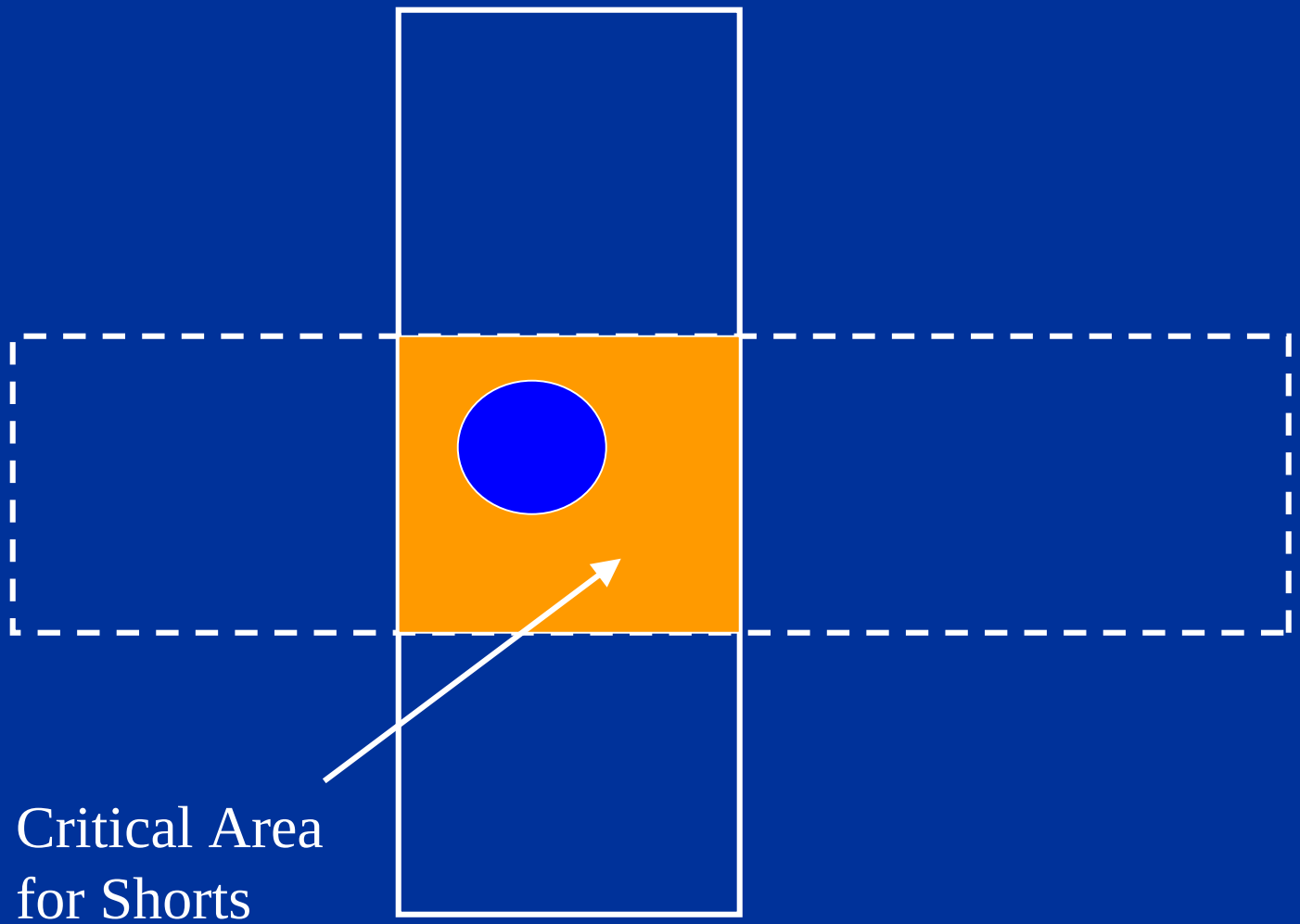


fatal defect types (two types of short circuits, one type of open)

Critical Area for Short Circuits



Critical Area for Short Circuits



Approaches to Spot Defect Yield Loss

- Modify wire placements to minimize critical area
- Router issue
  - router understands critical-area analyses, optimizations
  - spread, push/shove (gridless, compaction technology)
  - layer reassignment, via shifting (standard capabilities)
  - related: via doubling when available, etc.
- Post-processing approaches in PV are awkward
  - breaks performance verification in layout (if layout has been changed by physical verification)
  - no easy loop back to physical design: convergence problems

Antennas

- Charging in semiconductor processing
  - many process steps use plasmas, charged particles
  - charge collects on conducting poly, metal surfaces
  - capacitive coupling: large electrical fields over gate oxides
  - stresses cause damage, or complete breakdown
  - induced V_t shifts affect device matching (e.g., in analog)

Antennas

- Charging in semiconductor processing
- Standard solution: limit antenna ratio
  - antenna ratio = (A_{poly} + A_{M1} + ...) / $A_{gate\text{-}ox}$
  - e.g., antenna ratio < 300
  - A_{Mx} ≡ *metal (x)* area electrically connected to node without using *metal (x+1)*, and not connected to an active area

Antennas

- Charging in semiconductor processing
- Standard solution: limit antenna ratio
- General solution == bridging (break antenna by moving route to higher layer)
- Antennas also solved by protection diodes
  - not free (leakage power, area penalties)

Session Overview

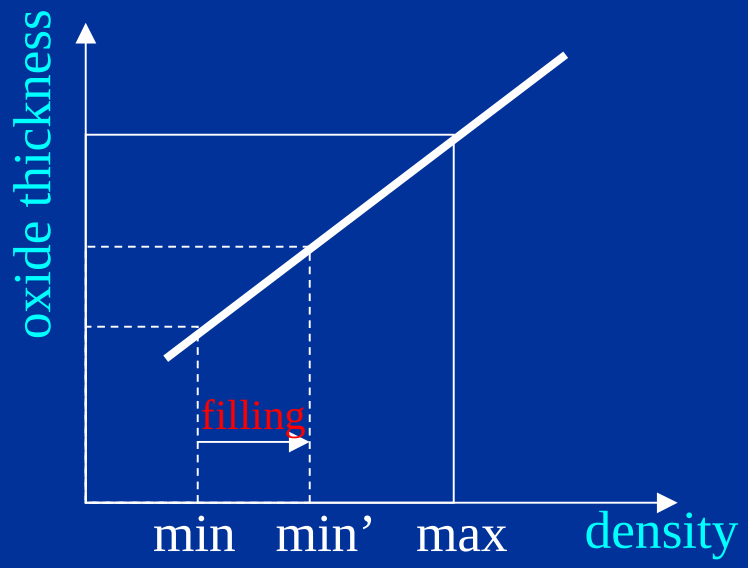
- New issues and problems arising in UDSM technology
  - catastrophic yield: critical area, antennas
  - parametric yield: density control (filling) for CMP
  - parametric yield: subwavelength lithography implications
    - optical proximity correction (OPC)
    - phase-shifting mask design (PSM)
  - signal integrity
    - crosstalk and delay uncertainty
    - DC electromigration
    - AC self-heat
    - hot electrons
- Current context: cell-based place-and-route methodology
  - placement and routing formulations, basic technologies
  - methodology contexts

Density Control for CMP

- Chemical-mechanical polishing (CMP)
  - applied to interlayer dielectrics (ILD) and inlaid metals
  - polishing pad wear, slurry composition, pad elasticity make this a very difficult process step
- Cause of CMP variability
  - pad deforms over metal feature
  - greater ILD thickness over dense regions of layout
  - "dishing" in sparse regions of layout
  - huge part of chip variability budget used up (e.g., 4000Å ILD variation across-die)

Min-Variation Objective

- Relationship between oxide thickness and local feature density



- Minimizing variation in window density over layout preferable to satisfying lower and upper density bounds

Density Control for CMP

- Layout density control
  - density rules minimize yield impact
  - uniform density achieved by post-processing, insertion of dummy features
- Performance verification (PV) flow implications
  - accurate estimation of filling is needed in PD, PV tools (else broken performance analysis flow)
  - filling geometries affect capacitance extraction by > 50%
  - is a multilayer problem (coupling to critical nets, contacting restrictions, active layers, other interlayer dependencies)

Density Rules

- Modern foundry rules specify layout density bounds to minimize impact of CMP on yield
- Density rules control local feature density for $w \times w$ windows
  - e.g., for metal layer every 2000um $\times$ 2000um window must be between 35% and 70% filled
- Filling = insertion of "dummy" features to improve layout density
  - typically via layout post-processing in PV / TCAD tools
  - affects vital design characteristics (e.g., RC extraction)
  - accurate knowledge of filling is required during physical design and verification

Need for Density Awareness in Layout

- Performance verification flow:

- Filling/slotting geometries affect RC extraction

VICTIM LAYER TOTAL CAPACITANCE (10^{-15} F)			
Same layer-i neighbors?	Fill layers i-1, i+1?	$\varepsilon = 3.9$	$\varepsilon = 2.7$
N	N	2.43 (1.0)	1.68 (1.0)
N	Y	3.73 (1.54)	2.58 (1.54)
Y	N	4.47 (1.84)	3.09 (1.84)
Y	Y	5.29 (2.18)	3.66 (2.18)

- Up to 1% error in extracted capacitance
- Reliability also affected (e.g. slotting of power stripes)

Need for Density Awareness in Layout

- Performance verification flow:

- Can be considered as ``single-layer'' problem

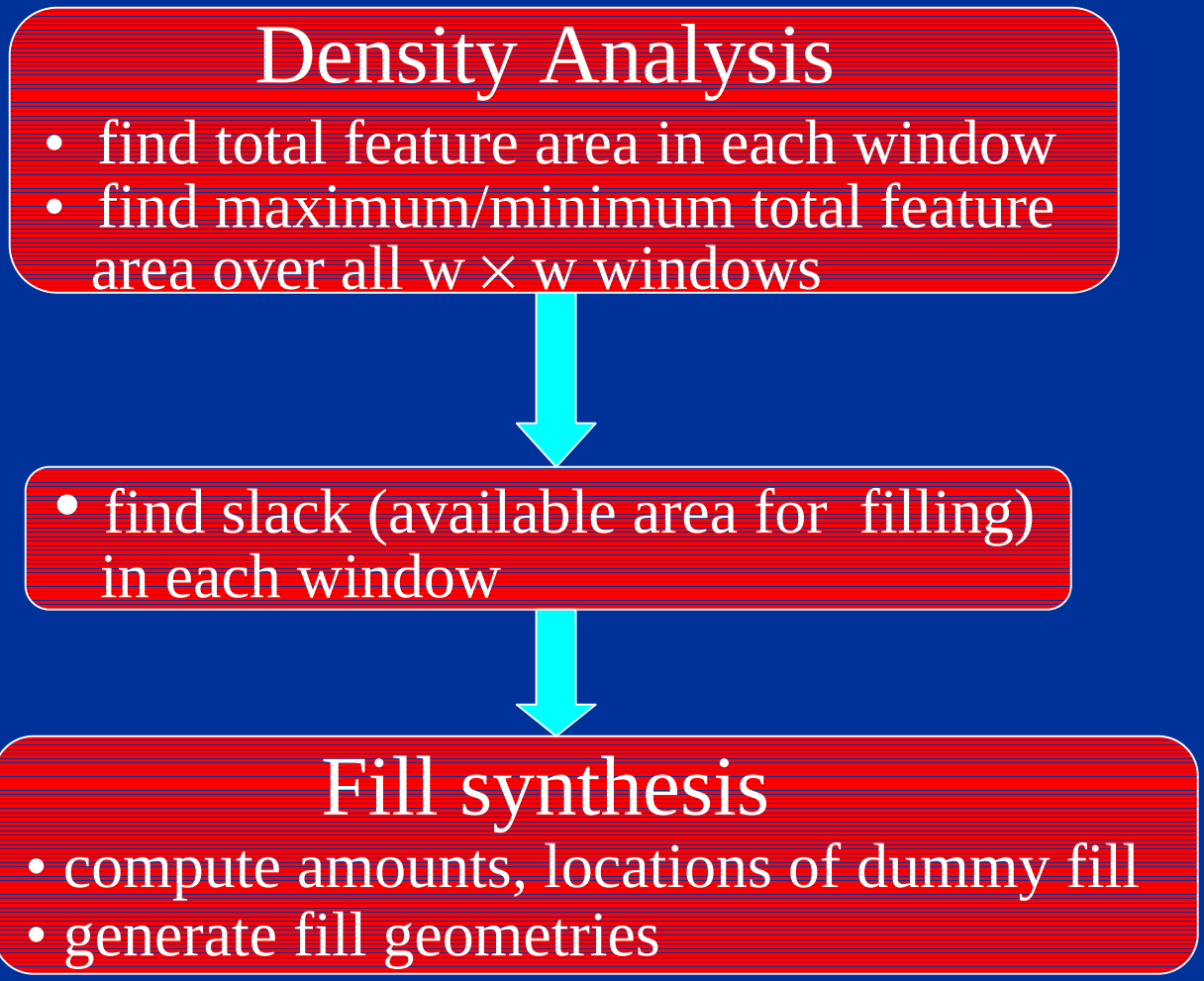
Middle Victim Conductor Total Capacitance (10^{-15} F)			
Fill layer offset	Fill geometry	$\varepsilon = 3.9$	$\varepsilon = 2.7$
N	10×1	3.776 (1.0)	2.614 (1.0)
N	1×1	3.750 (0.99)	2.596 (0.99)
Y	10×1	3.777 (1.00)	2.615 (1.00)
Y	1×1	3.745 (0.99)	2.593 (0.99)

- Caveat: contacting, active layers, other interlayer dependencies

Limitations of Current Techniques

- Current techniques for density control have three key weaknesses:
  (1) only the average ***overall*** feature density is constrained, while local variation in feature density is ignored
  (2) density analysis does not find ***true*** extremal window densities - instead, it finds extremal window densities only over fixed set of window positions
  (3) fill insertion into layout does not minimize the maximum variation in window density

Layout Density Control Flow



Session Overview

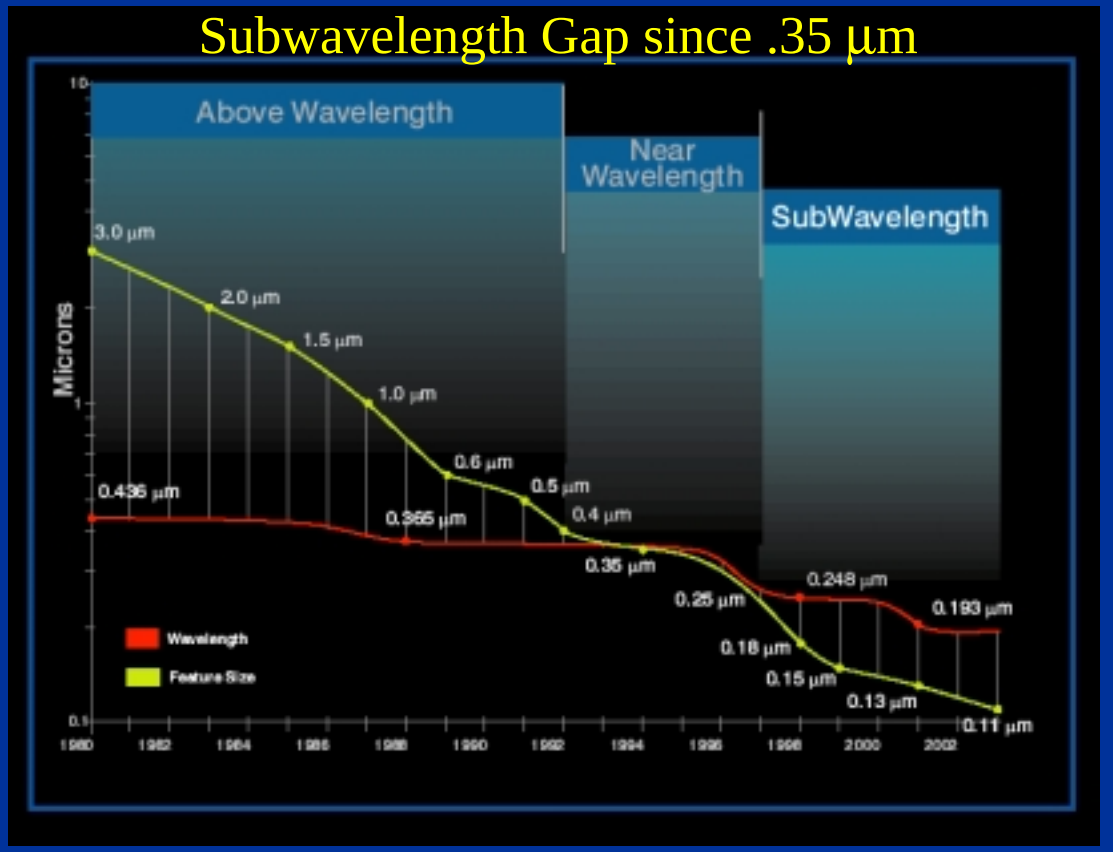
- New issues and problems arising in UDSM technology
  - catastrophic yield: critical area, antennas
  - parametric yield: density control (filling) for CMP
  - parametric yield: subwavelength lithography implications
    - optical proximity correction (OPC)
    - phase-shifting mask design (PSM)
  - signal integrity
    - crosstalk and delay uncertainty
    - DC electromigration
    - AC self-heat
    - hot electrons
- Current context: cell-based place-and-route methodology
  - placement and routing formulations, basic technologies
  - methodology contexts

Subwavelength Optical Lithography — Technology Limits

- Implications of Moore's Law for feature sizes
- Steppers not available; WYSIWYG fails after .35μm generation
- Optical lithography
  - circuit patterns optically projected onto wafer
  - feature size limited by diffraction effects
  - Rayleigh limits
    - resolution R proportional to λ / NA
    - depth of focus DOF proportional to λ / NA^2
- Available knobs
  - amplitude (aperture): OPC
  - phase: PSM

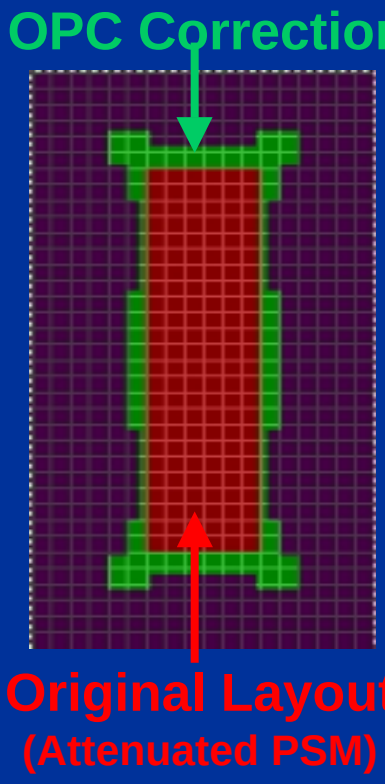
Next-Generation Lithography and the Subwavelength Gap

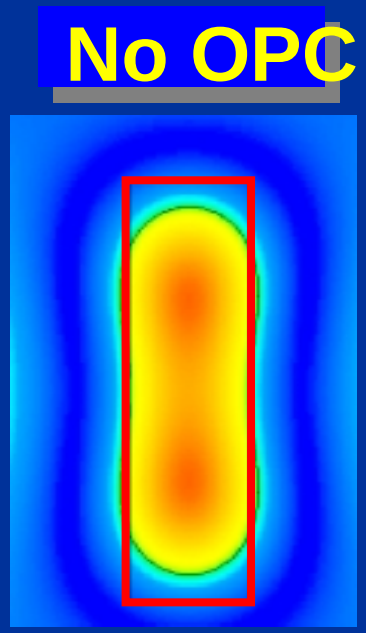
- EUV
- X-rays
- E-beams
- All at least 10 years away; require significant R&D, major infrastructure changes
- > 30 years of infrastructure and experience supporting optical lithography

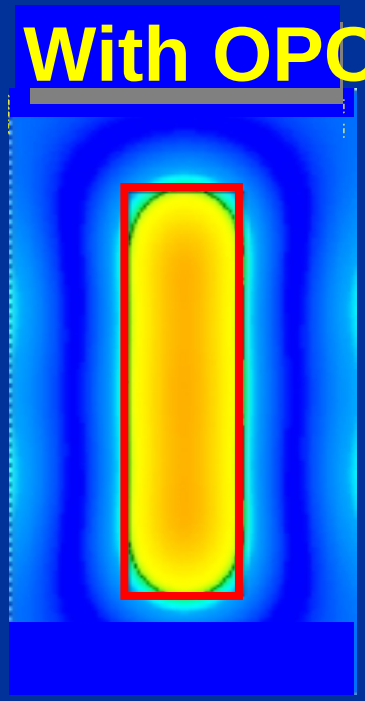


Optical Proximity Correction (OPC)

- Corrective modifications to improve process control
  - improve yield (process latitude)
  - improve device performance







Optical Proximity Correction (OPC)

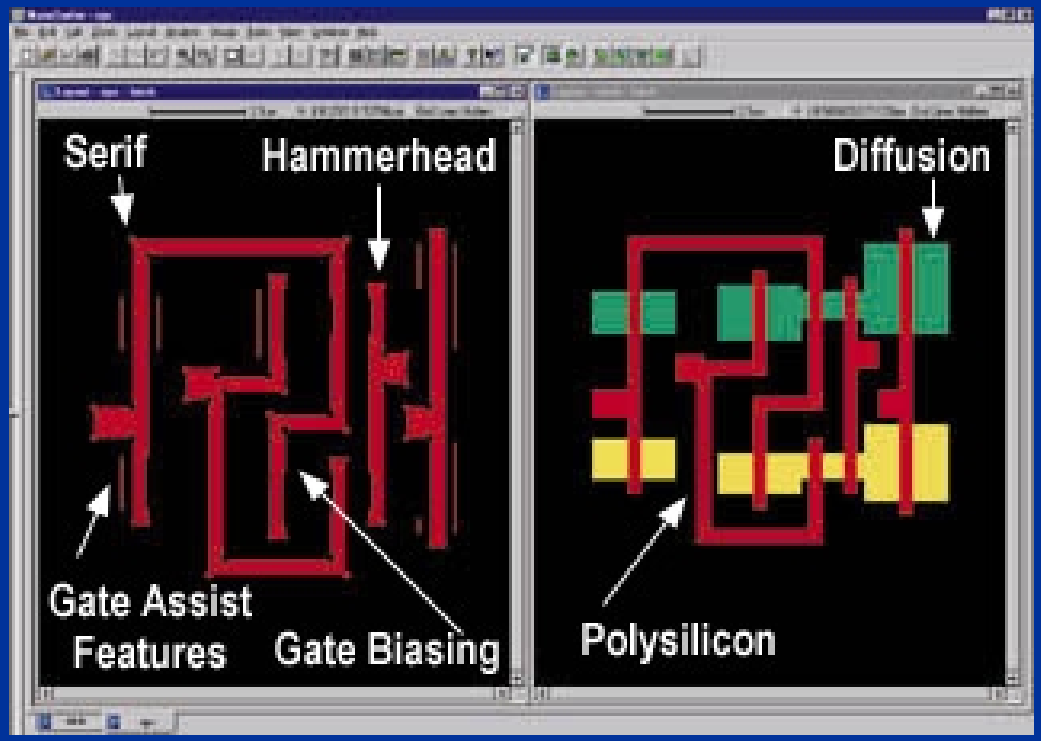
- Mostly cosmetic corrections; complicates mask manufacturing and dramatically increases cost (with little benefit?)
- Post-design verification is essential

Optical Proximity Correction (OPC)

- Rule-based OPC
  - apply corrections based on a set of predetermined rules
  - fast design time, lower mask complexity
  - suitable for less aggressive designs
- Model-based OPC
  - use process simulation to determine corrections on-line
  - longer design time, increased mask complexity
  - suitable for aggressive designs

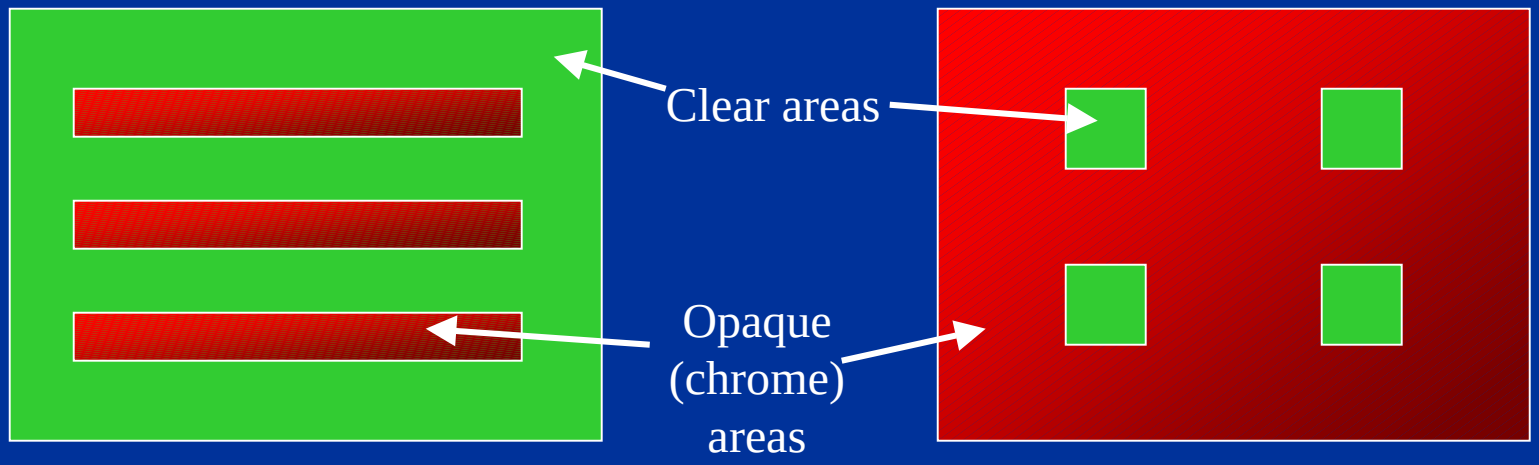
OPC Features

- Serifs - for corner rounding
- Hammerheads - for line-end shortening
- Gate assists (subresolution scattering bars) - for CD control
- Gate biasing - for CD control
- Issues for custom, hierarchical and reuse-based layout methodologies



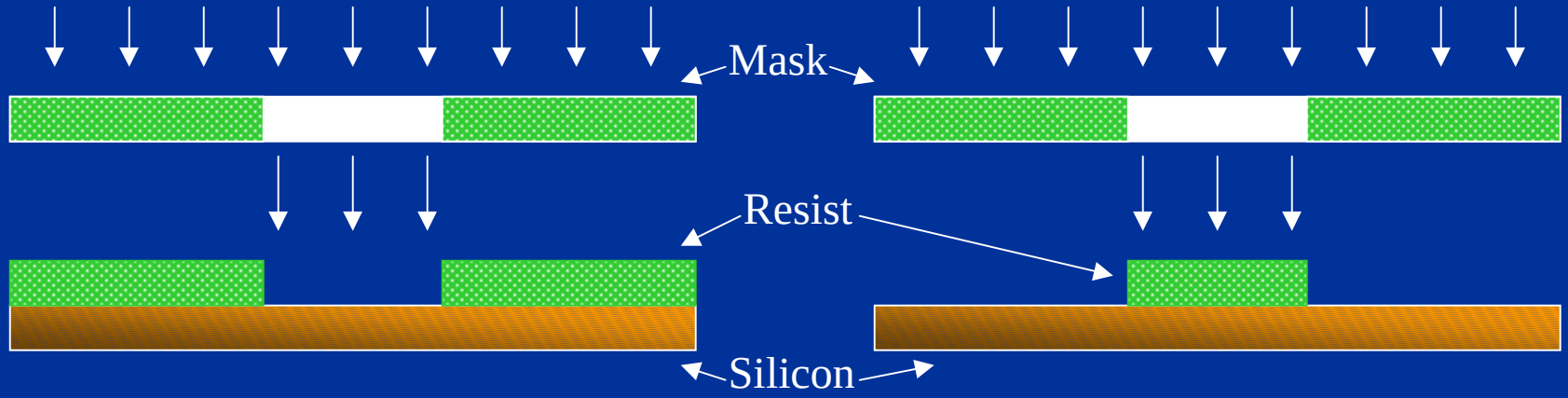
Mask Types

- Bright field masks
  - opaque features defined by chrome
  - background is transparent
  - used, e.g., for poly and metal
- Dark field masks
  - transparent features defined
  - background is opaque (chrome)
  - used, e.g., for contacts
  - used also for damascene metals



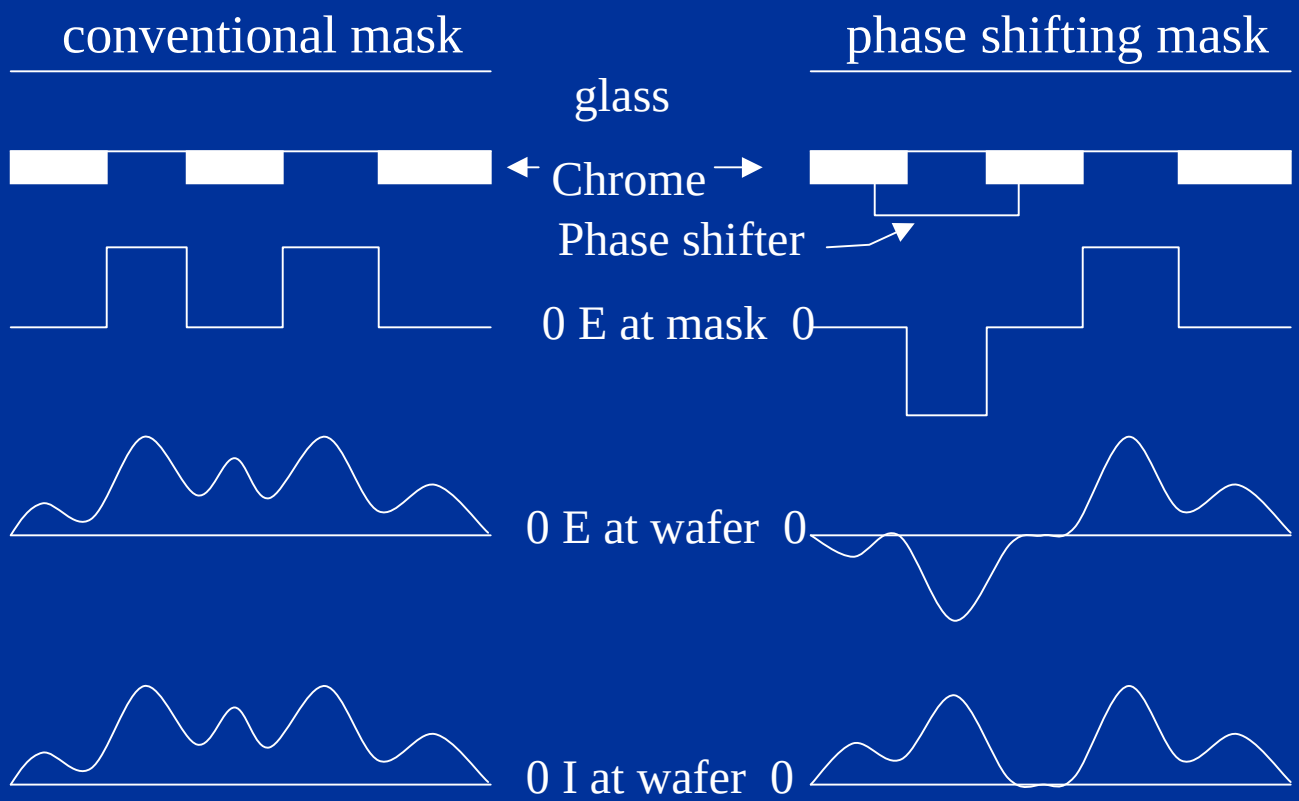
Photoresist Types

- Positive resists
  - material is removed from *exposed* areas during development
  - most widely used
- Negative resists
  - material is removed from *unexposed* areas during development
  - less mature



Post development profile for positive and negative photoresists

Phase Shifting Masks

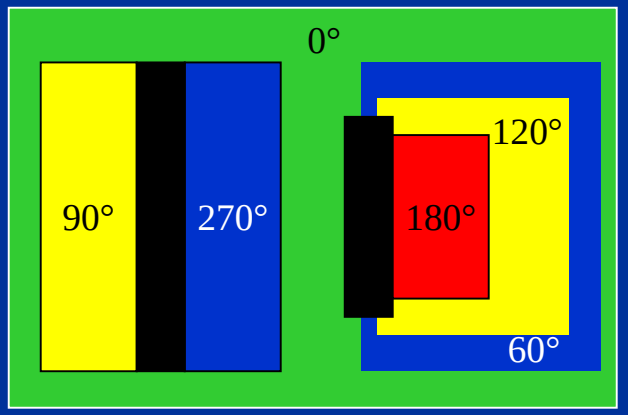


Phase Shifting Masks

- no phase shifting: poor contrast due to diffraction
- phase shifting by 180°: reverse electric field on mask, destructive interference yields zero-intensity on wafer (high contrast)
- Background: invented in 1982 by Levenson at IBM
- Heightened interest in early 1990s, but near wavelength → no pressing need
- Many forms of phase-shifting proposed
- Key issues: manufacturability, design tools
- Today: subwavelength gap forces PSM into every process (example: Motorola 90nm gates announced in early 1999)

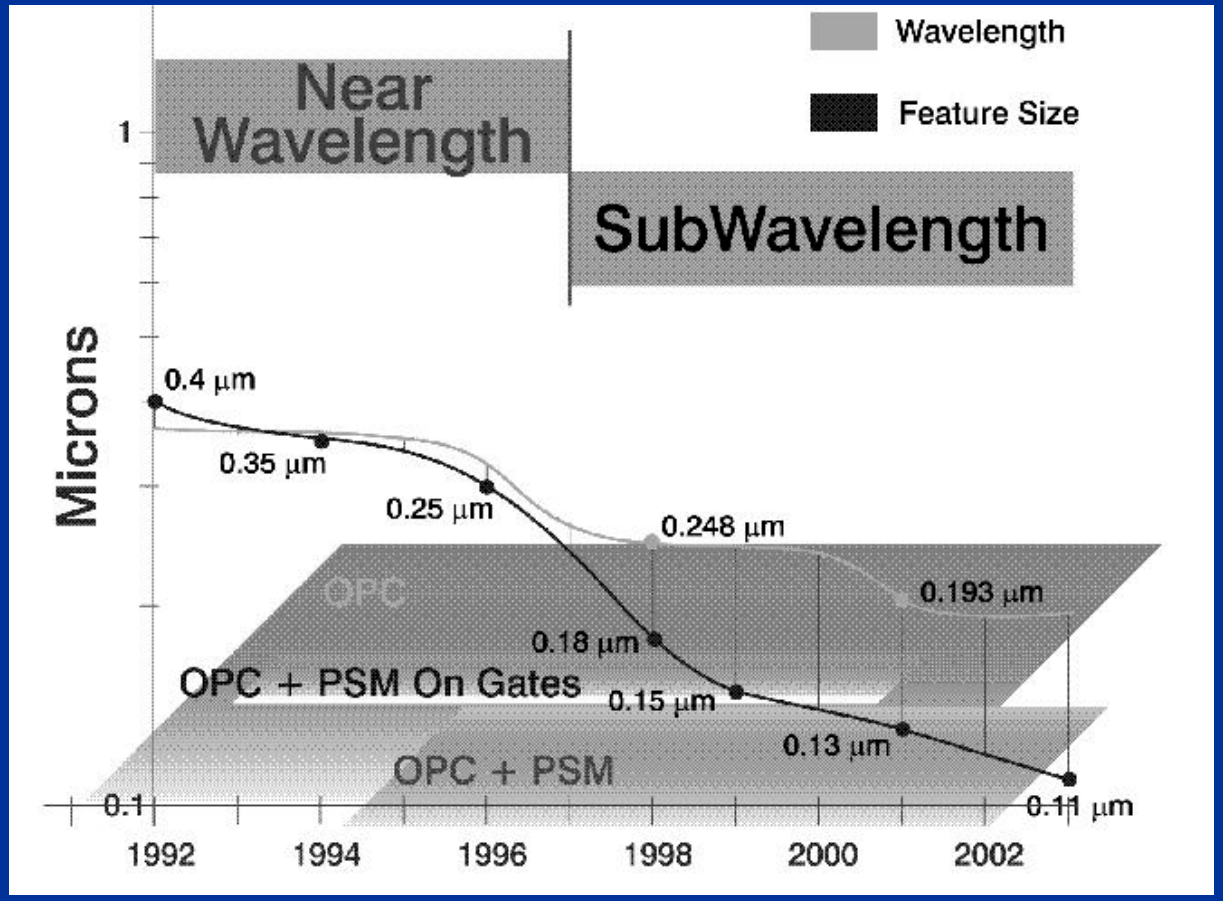
Forms of PSM

- Bright Field Phase-Shifting
  - single exposure
    - phase transitions required, e.g., 0-60-120-180 or 90-0-270 to avoid printing phase edges
    - throughput unaffected
    - limited improvement in process latitude
    - mask manufacturing difficult, mask cost very high
  - double exposure
    - PSM with 0 and 180 degree phase shifters
    - define only critical features ("locally bright-field"), rest of mask is chrome
    - second exposure with clear-field binary mask protects critical features, defines non-critical features as well
    - excellent process latitude
    - decrease in throughput (double exposure)

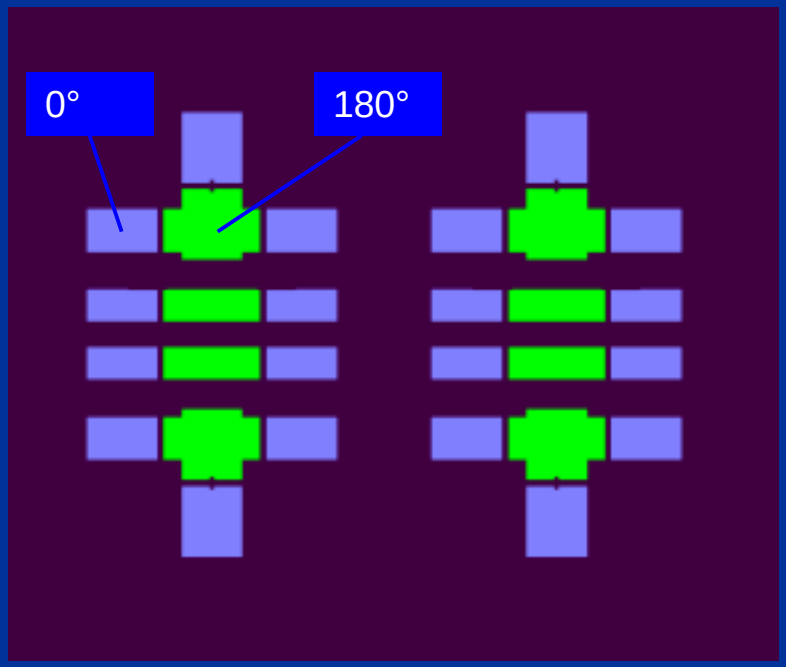


Applicability of OPC and PSM

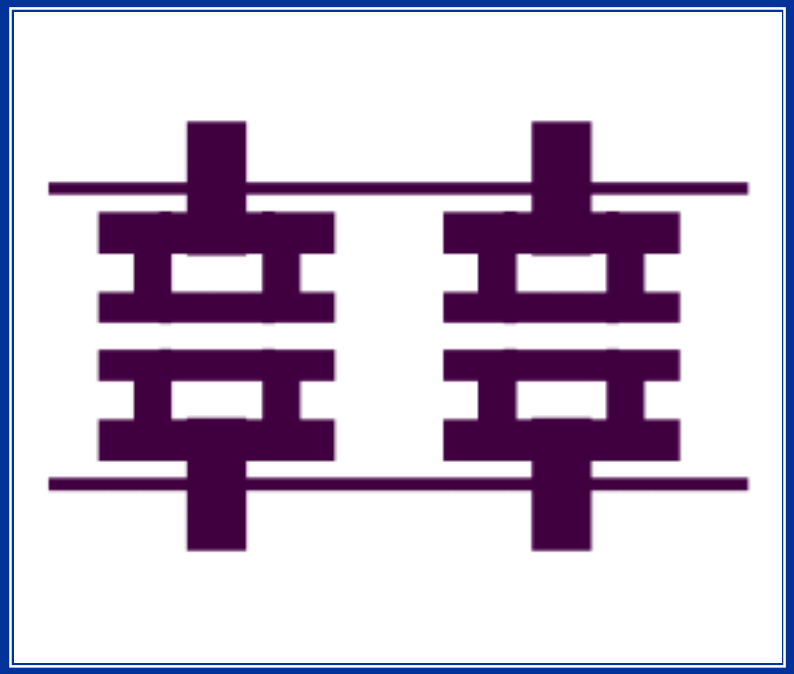
Bridging the SubWavelength Gap: Creating the 0.09 μm stepper



Double-Exposure Alternating PSM

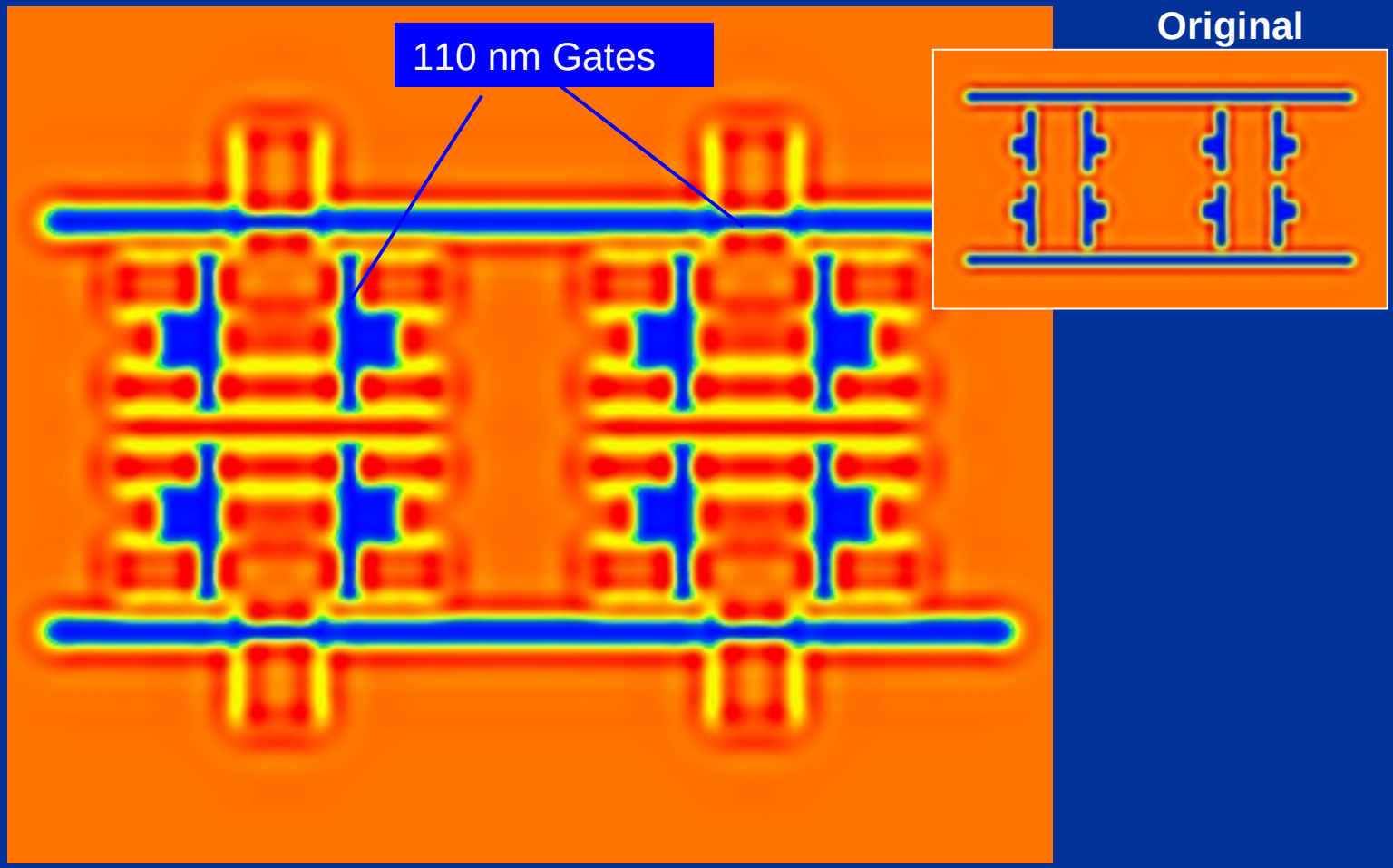


1. Alternate PSM Mask

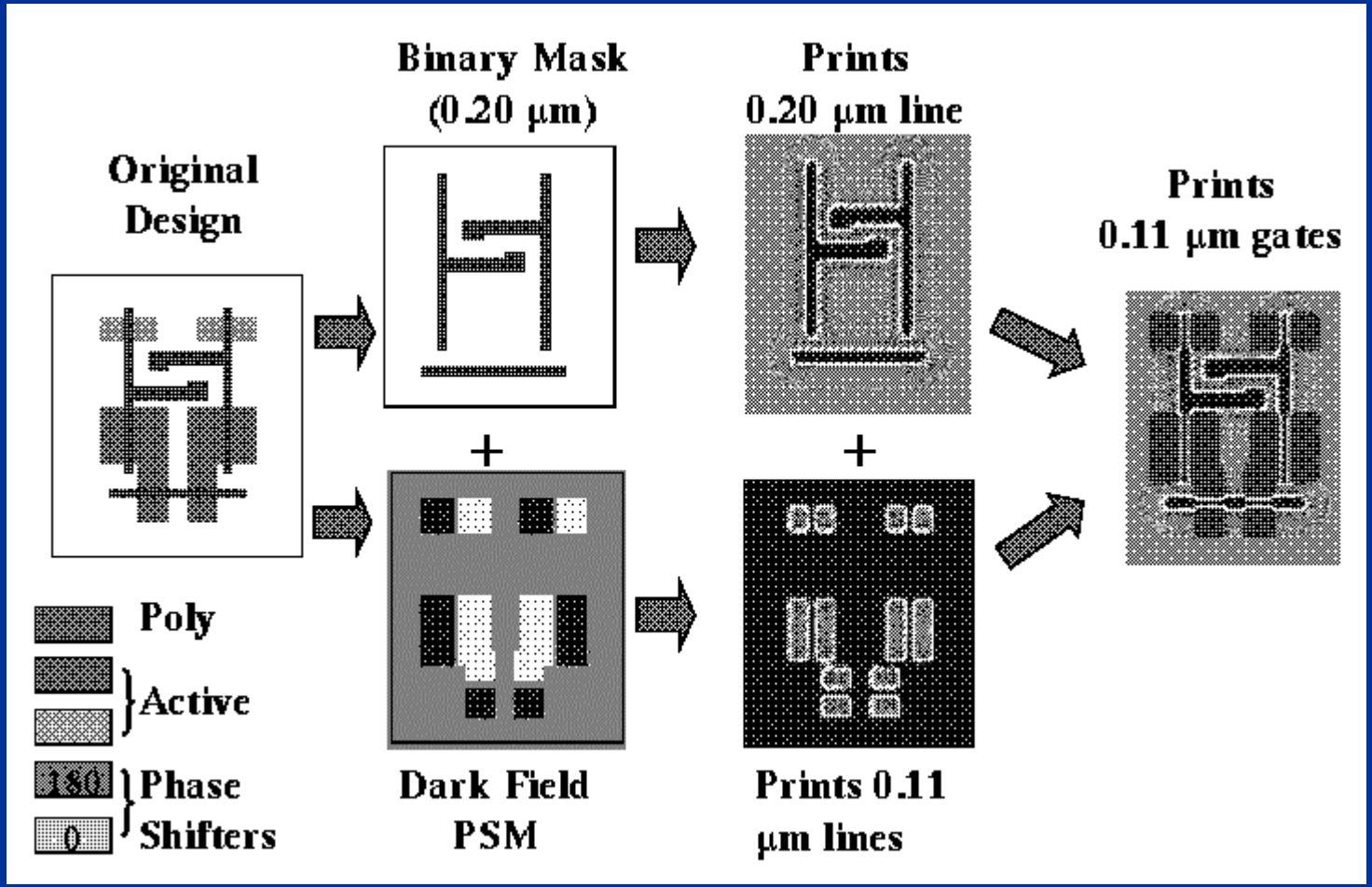


2. Trim Mask (COG)

Benefits of PSM



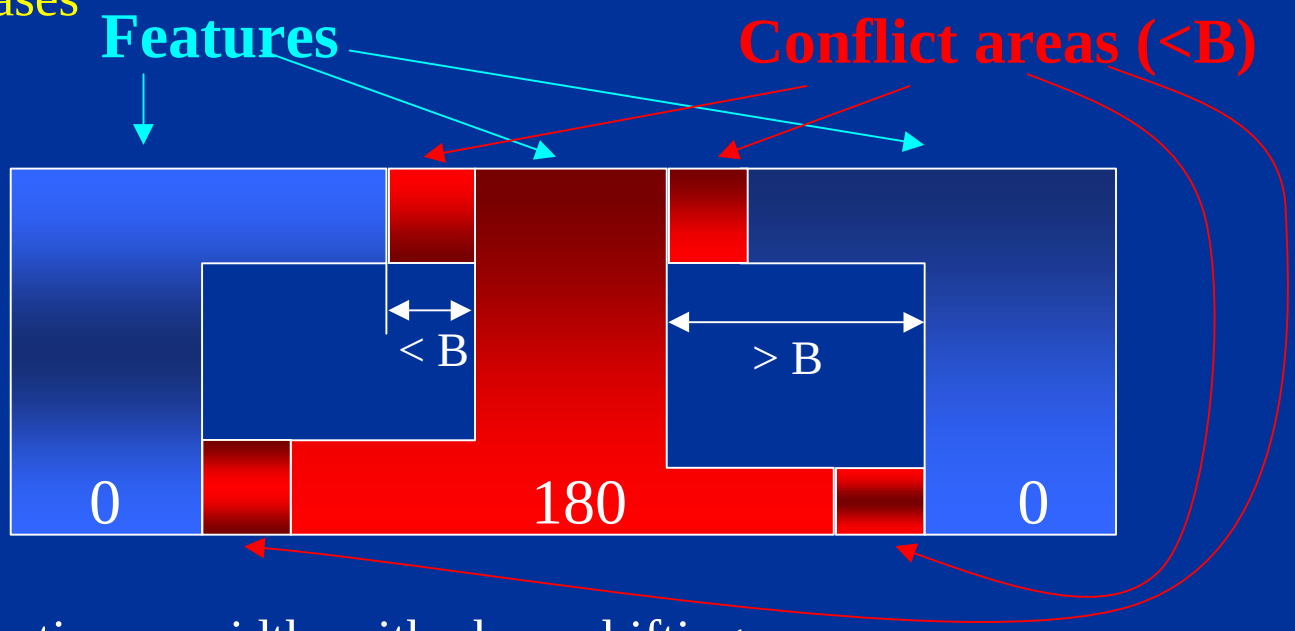
Gate Shrinking and CD Control Using Phase Shifting



The Phase Assignment Problem in PSM

Assign 0, 180 phase regions such that

- (dark field) feature pairs with separation < B have opposite phases
- (bright field) features with width < B are induced by adjacent phase regions with opposite phases

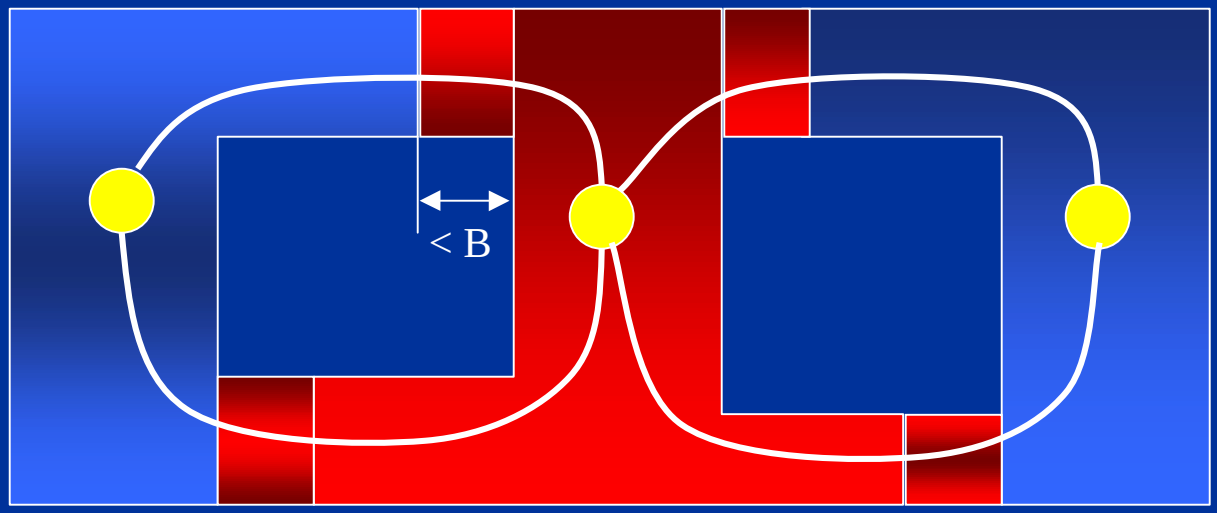


b ≡ minimum separation or width, with phase shifting

B ≡ minimum separation or width, without phase shifting

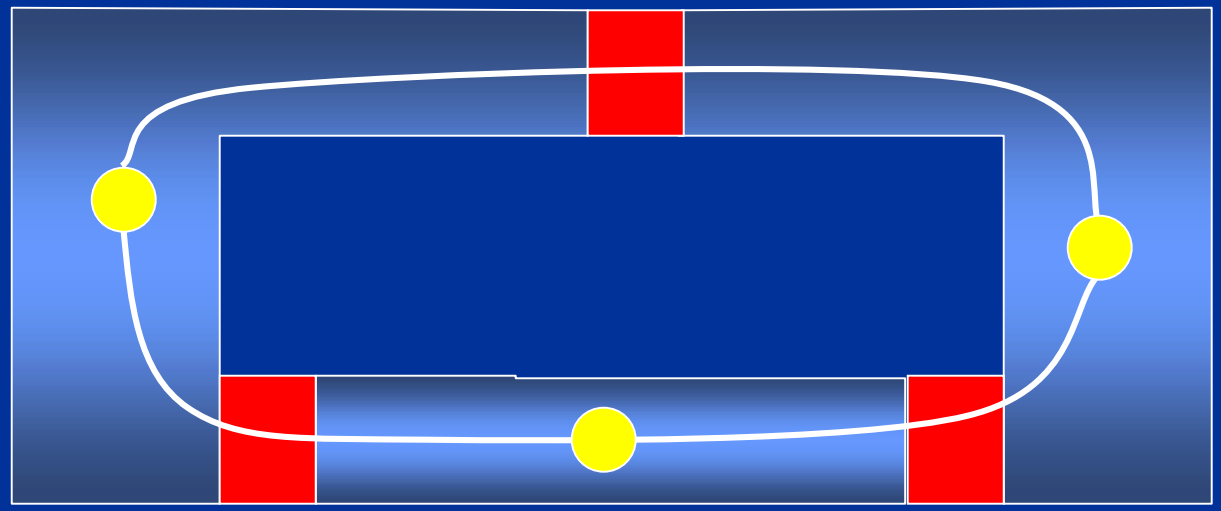
Phase Conflict and the Conflict Graph

- Vertices: features (or phase regions)
- Edges: “conflicts” (necessary phase contrasts)
  (feature pairs with separation < B)



Odd Cycles in Conflict Graph

- Self-consistent phase assignment is not possible if there is an odd cycle in the conflict graph
- Phase-assignable ≡ bipartite ≡ no odd cycles

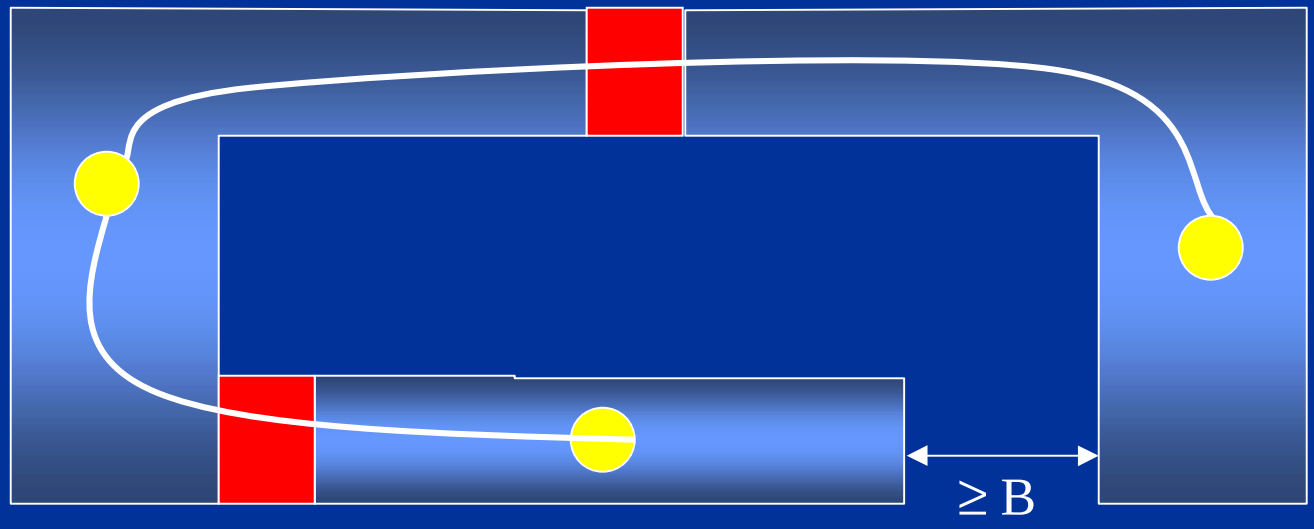


Phase Conflict and the Conflict Graph

- Self-consistent phase assignment is not possible if there is an odd cycle in the conflict graph
- Phase-assignable = bipartite = no odd cycles
- Breaking odd cycles: must change the layout!
  - change feature dimensions, and/or change spacings
  - degrees of freedom include layer reassignment for interconnects

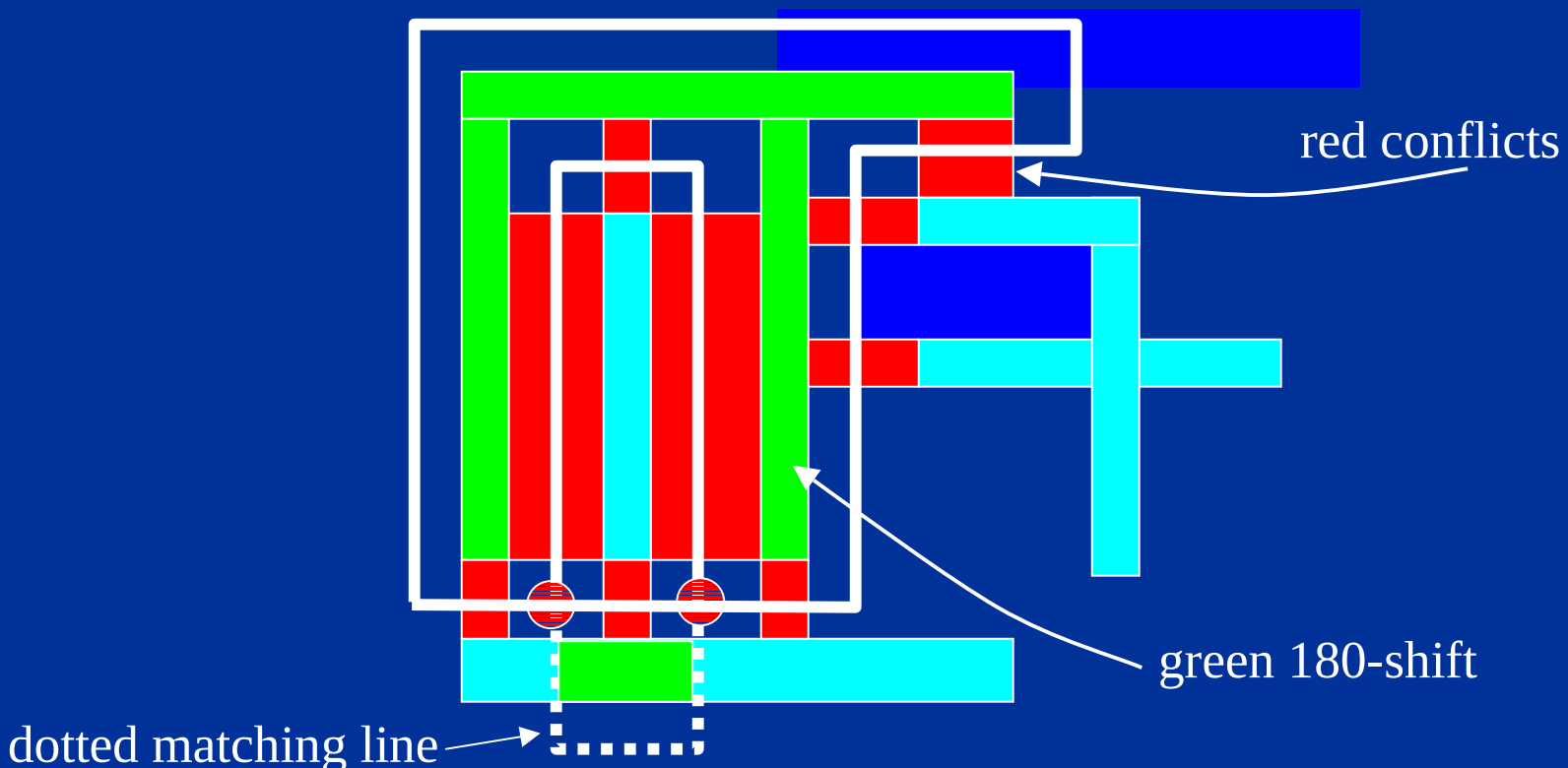
Breaking Odd Cycles

- Must change the layout:
  - change feature dimensions, and/or
  - change spacings
  - PSM phase-assignability is a layout, not verification, issue



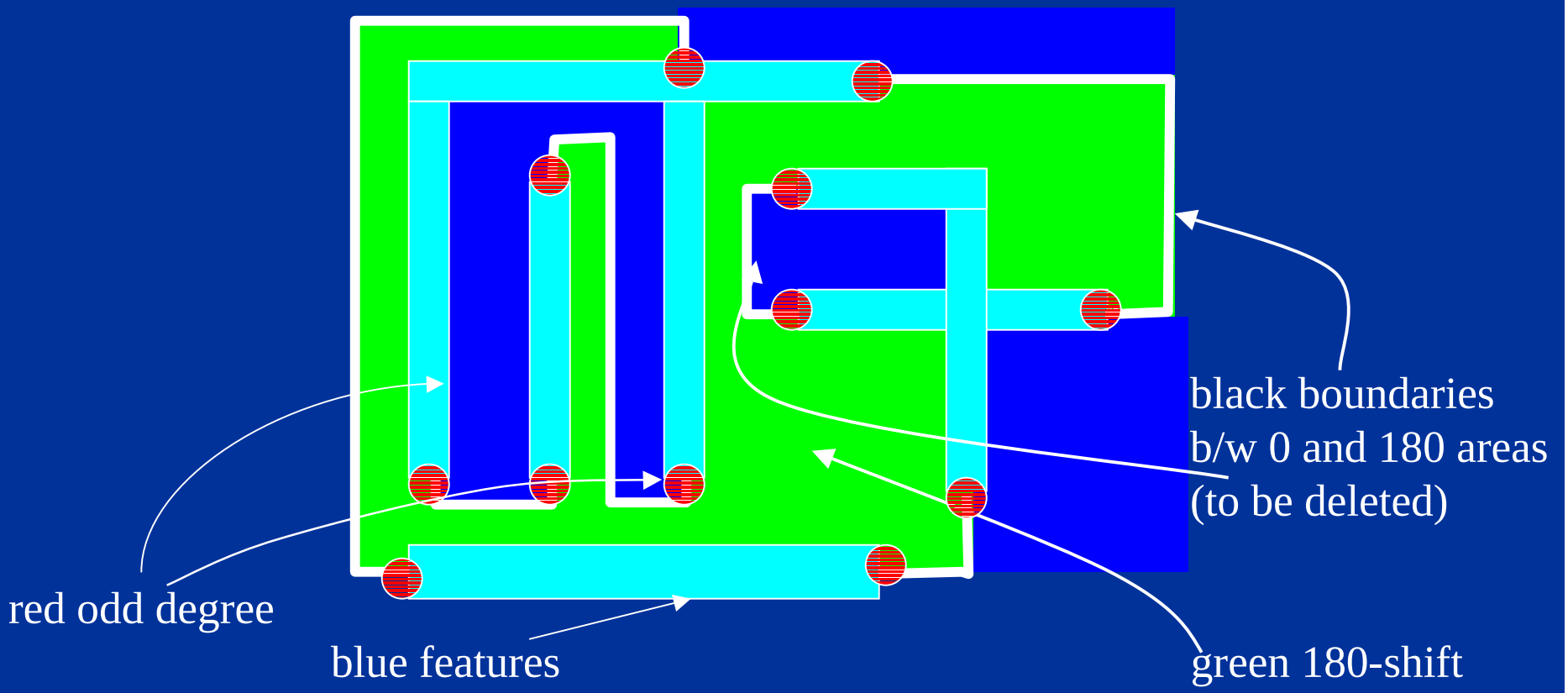
Phase Assignment - Dark Field

- Dark Field (odd-cycle breaking formulation)



Phase Assignment - Bright Field

- Bright Field (dense criticality regime)



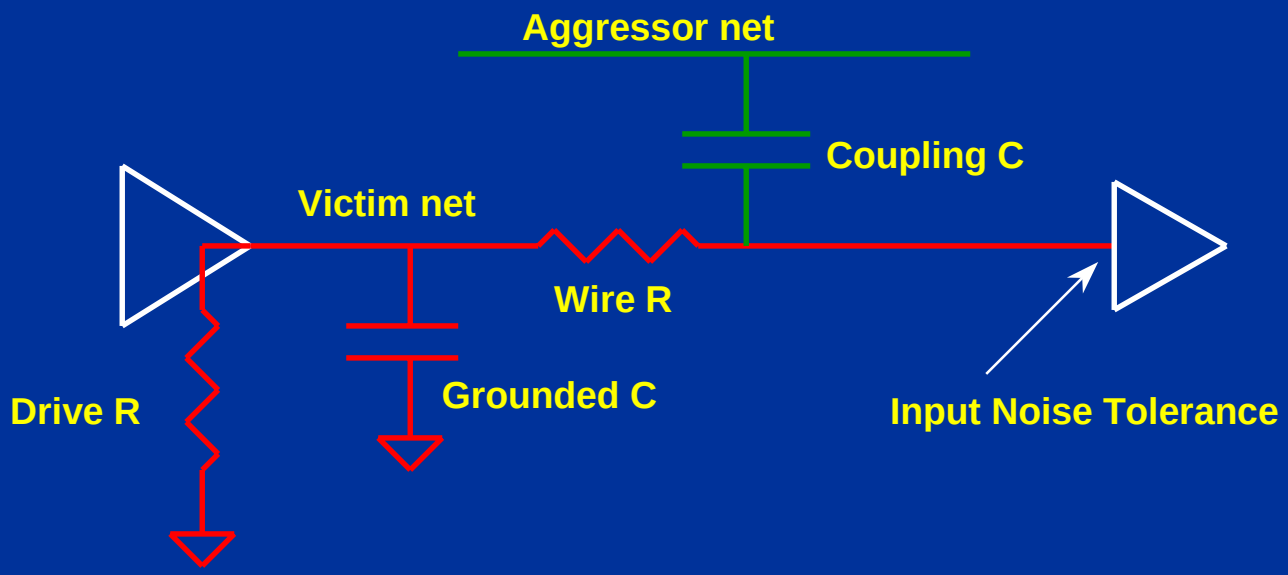
ucla cadence sagantec

Session Overview

- New issues and problems arising in UDSM technology
  - catastrophic yield: critical area, antennas
  - parametric yield: density control (filling) for CMP
  - parametric yield: subwavelength lithography implications
    - optical proximity correction (OPC)
    - phase-shifting mask design (PSM)
  - signal integrity
    - crosstalk and delay uncertainty
    - IR drop
    - DC electromigration
    - AC self-heat
    - hot electrons
- Current context: cell-based place-and-route methodology
  - placement and routing formulations, basic technologies
  - methodology contexts

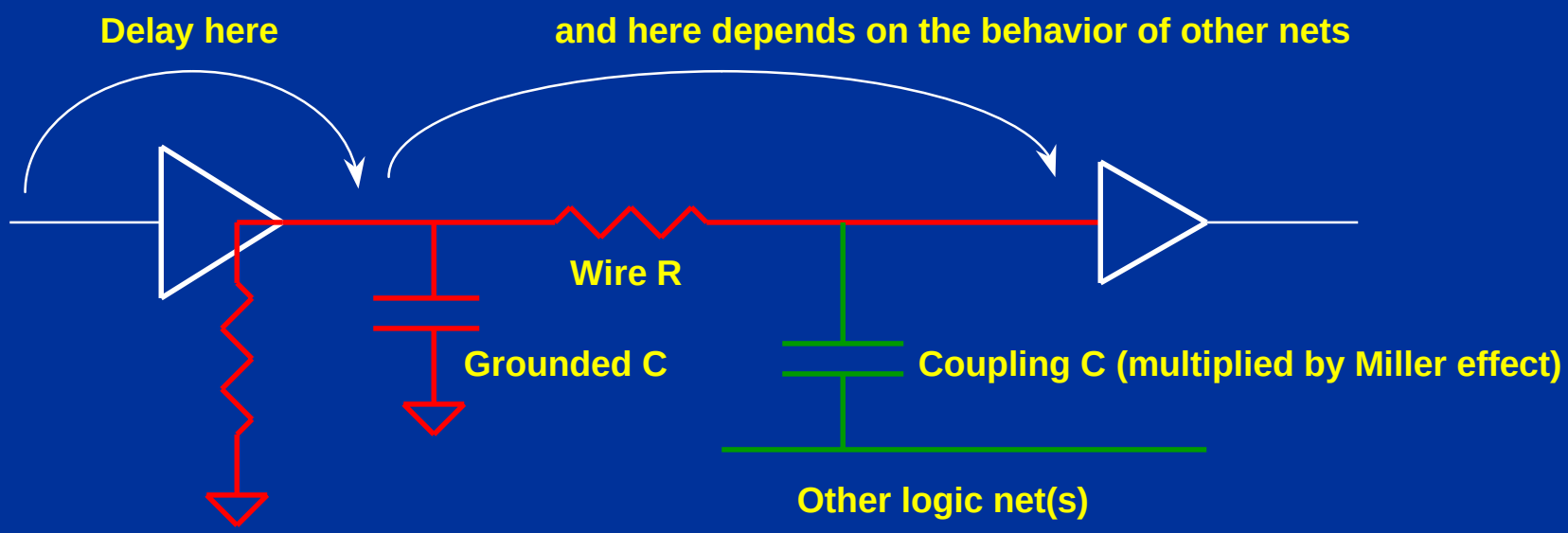
Crosstalk Induced Errors

- Transition on an adjoining signal causes unintended logic transition
- Symptom: chip fails (repeatably) on certain logic operations

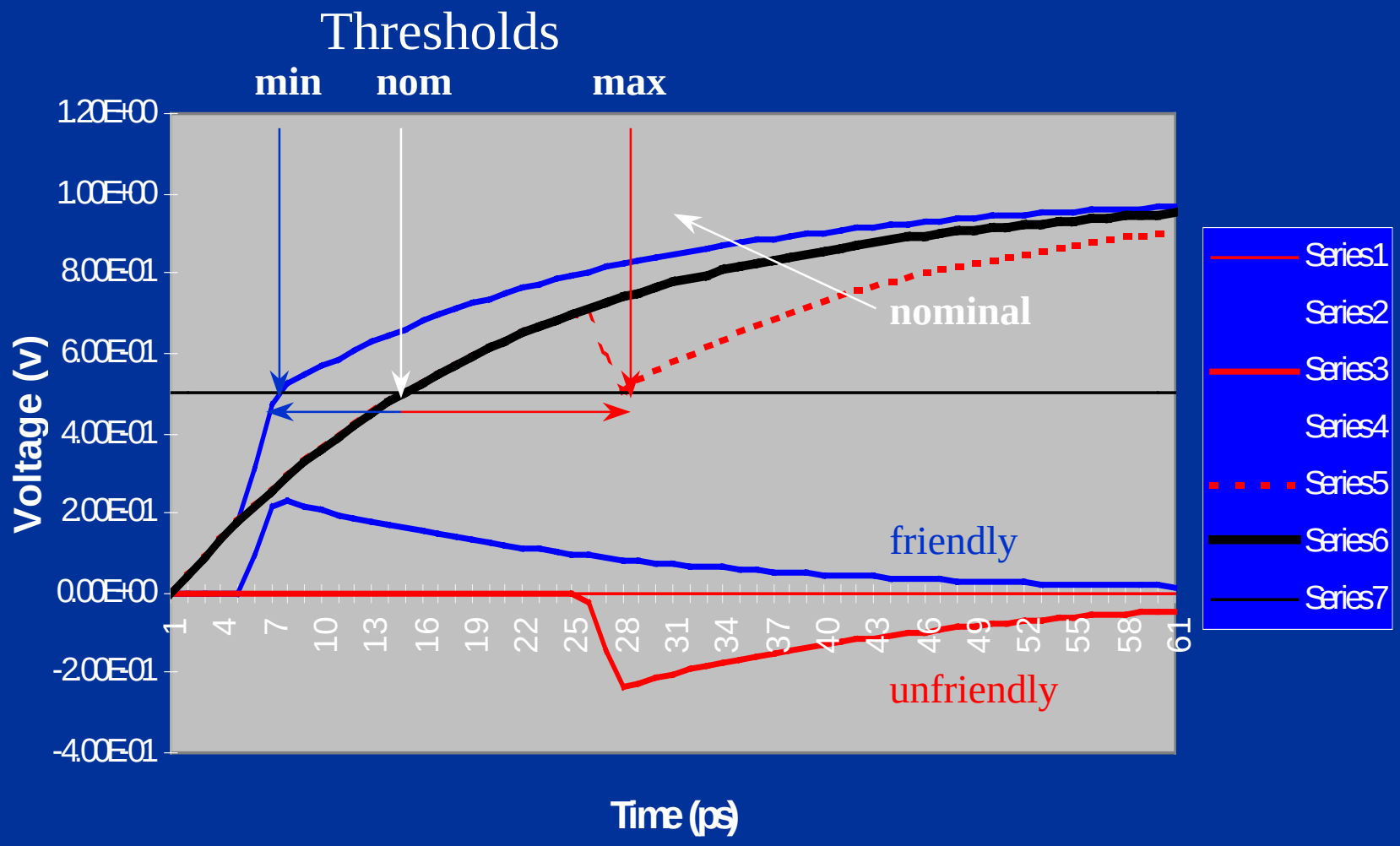


Crosstalk Induced Errors

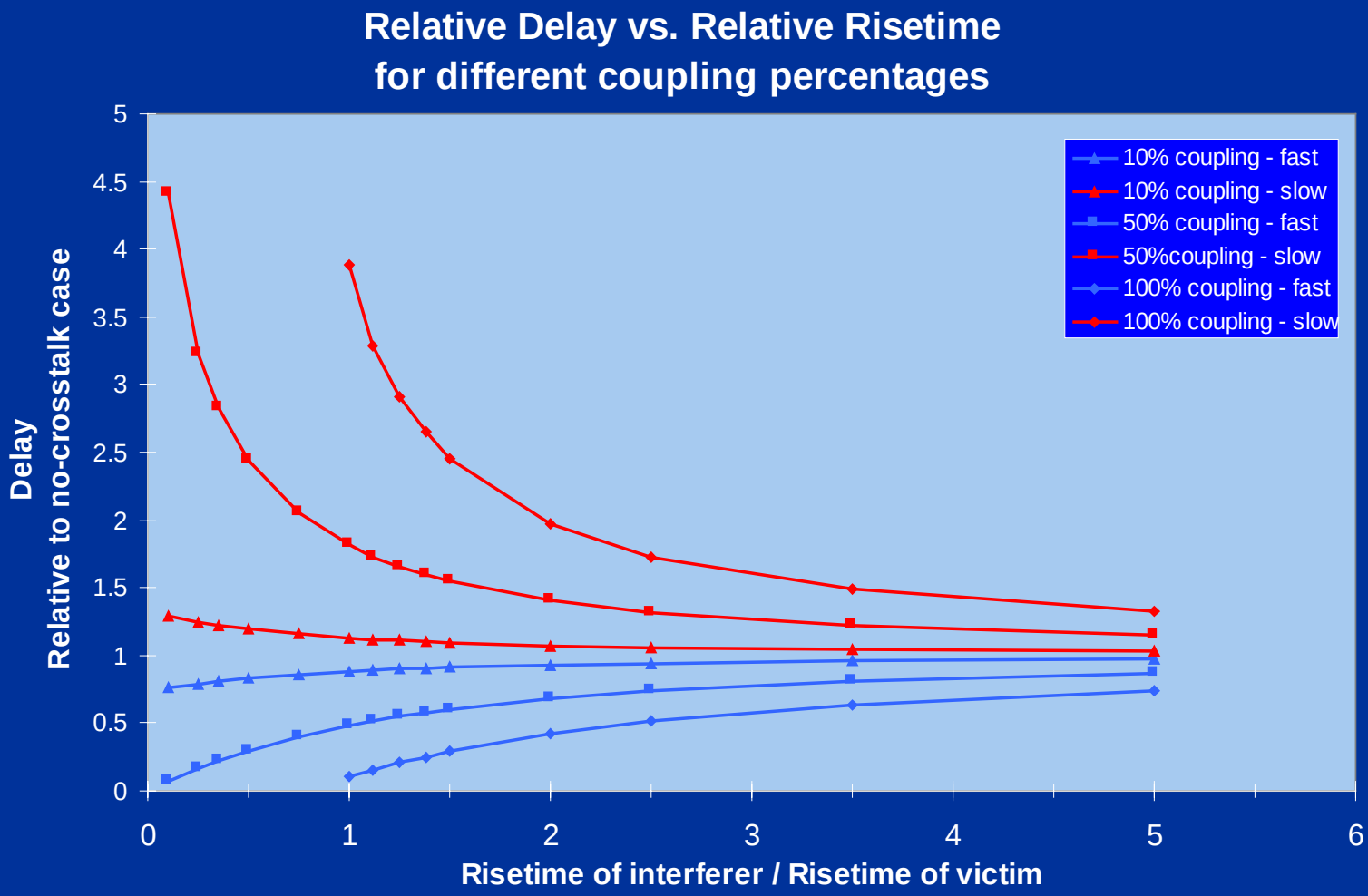
- Timing dependence on crosstalk
  - timing depends on behavior of adjoining signals
  - symptom: timing predictions inaccurate compared to silicon (effect can be large: 3:1 on individual nets)



Effects of Crosstalk: Delay Uncertainty



Effects of Crosstalk: Delay Uncertainty

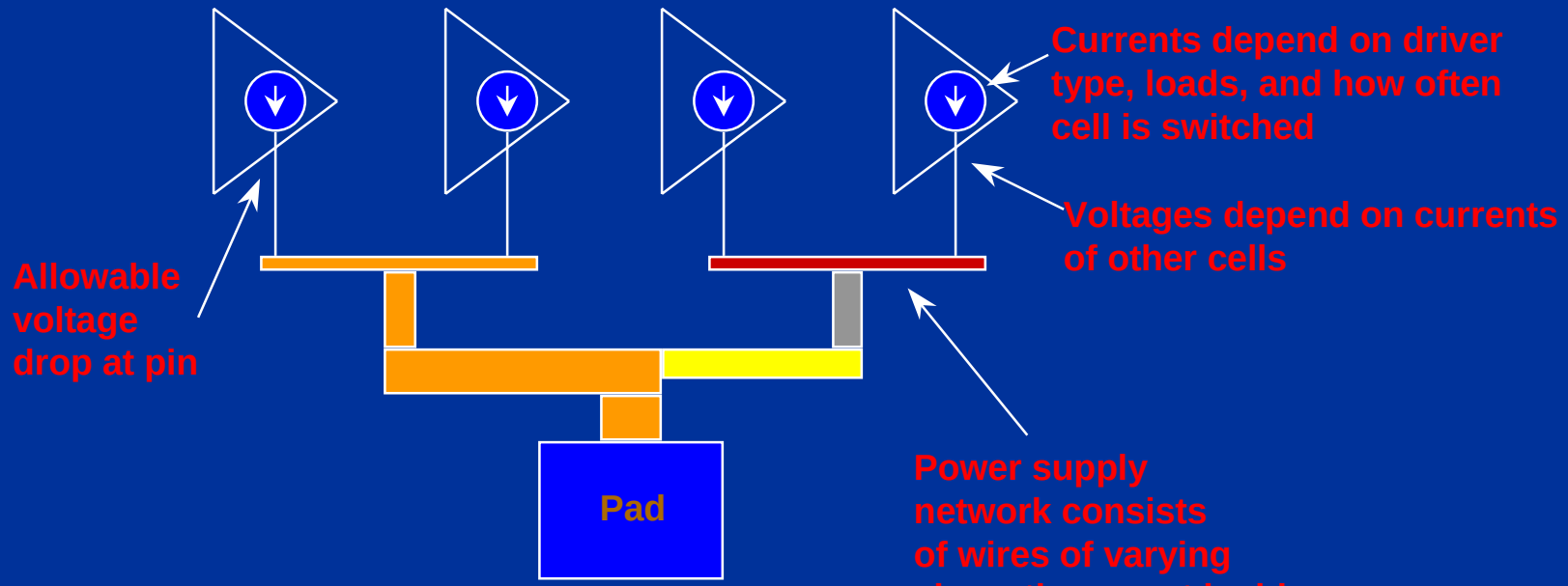


Crosstalk Prevention Strategies

- Placement phase
  - don't know adjacencies, layer assignments, or global routes
  - do know net length, est. wire R/C, driver strength, signal slews
  - establish metrics to tell if net is likely to have problems
  - fixes include driver sizing, buffering
- Global route phase
  - don't know adjacencies, but have idea of congestion
  - do know layer assignments, better R/C estimates
- Can apply timing windows
  - only consider signals that can change at the same time
  - data comes from static timing analysis
- Detailed routing - detailed analysis and routing ECOs

IR Drop

- Voltage drop in supply lines from currents drawn by cells
- Symptom: chip malfunctions on certain vectors
- Biggest problem: what's the worst-case vector?

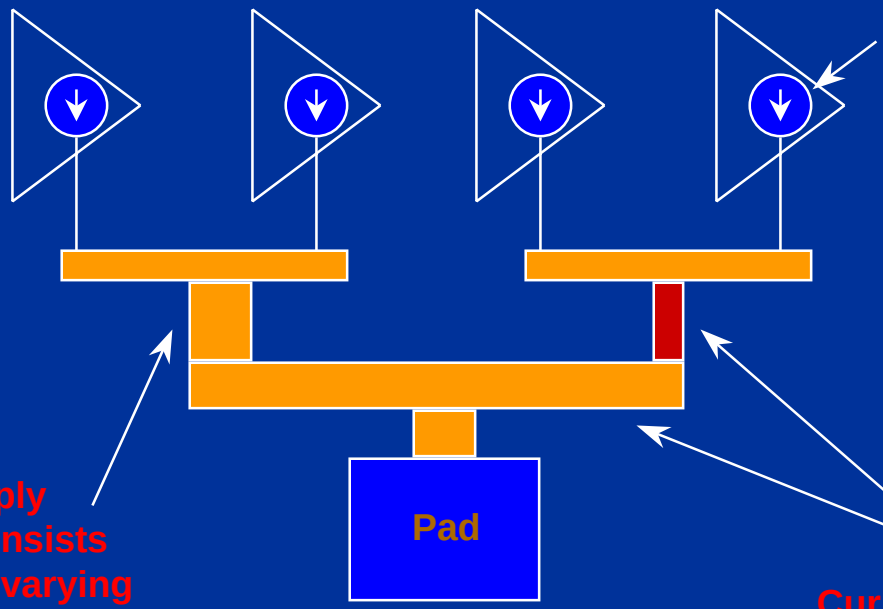


IR Drop

- Analysis
  - model I/O P/G supply; C extraction must distinguish decoupling cap between P/G and coupling cap between signals, P/G
- Prevention (good design)
  - P/G lines on same layer, close to each other; large decoupling on chip; process solutions (e.g., DEC Alpha)

Electromigration

- Power supply lines fail due to excessive current
- Symptom: chip eventually fails in the field when wire breaks



Currents depend on driver type, loads, and how often cell is switched

Currents depend on currents of other cells

Power supply network consists of wires of varying sizes; they must be big enough, but too big wastes area

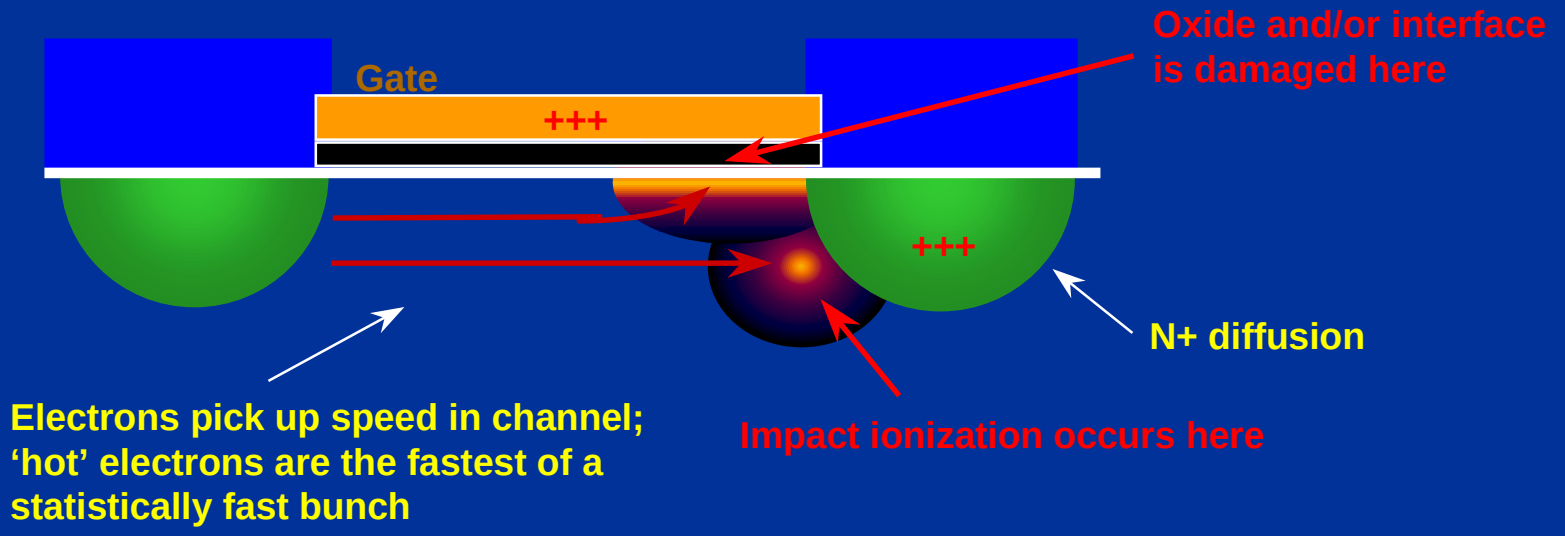
Current limit depends on wire size

Electromigration

- Prevention: wire cross-section to current rules
- Maximum current density for particular material (via, layer)
- Modified Black's equation; waveform models
- Higher limits for short, thin wires due to grain effects
- Copper: 100x resistance to EM $\rightarrow$ not a problem any more?

Hot Electron Effects

- Also called short-channel effect
- Caused by extremely high electric fields in the channel
  - occurs when voltages are not scaled as fast as dimensions
- Effect becomes worse as devices are turned on harder
- Symptom: thresholds shift over time until chip fails



Hot Electron Prevention Strategies

- Allowable region for input slew and output load
- Fluence per transition is function of input slew, output load
- Set maximum allowed degradation over life of device

  (estimate of total number of transitions) ≡ fluence limit
- Size device as needed
- Output load vs. driver sizes

Wire Self-Heat

- May also be called signal wire electromigration
- Wire heats above oxide temperature as pulses go through
- Symptom: chip eventually fails when wire breaks
- Depends on metal composition, signal frequency, wire sizes, slew rates, and amount of capacitance driven
- Requires different data/formulas from power supply EM

Session Overview

- New issues and problems arising in UDSM technology
  - catastrophic yield: critical area, antennas
  - parametric yield: density control (filling) for CMP
  - parametric yield: subwavelength lithography implications
    - optical proximity correction (OPC)
    - phase-shifting mask design (PSM)
  - signal integrity
    - crosstalk and delay uncertainty
    - IR drop
    - DC electromigration
    - AC self-heat
    - hot electrons
- Current context: cell-based place-and-route methodology
  - placement and routing formulations, basic technologies
  - methodology contexts

Cell-Based P&R: Classic Context

- Architecture design
  - golden microarchitecture design, behavioral model, RT-level structural HDL passed to chip planning
  - cycle time and cycle-accurate timing boundaries established
  - hierarchy correspondences (structural-functional, logical (schematic) and physical) well-established
- Chip planning
  - hierarchical floorplan, mixed hard-soft block placement
  - block context-sensitivity: no-fly, layer usage, other routing constraints
  - route planning of all global nets (control/data signals, clock, P/G)
  - induces pin assignments/orderings, hard (partial) pre-routes, etc.
- Individual block design -- various P&R methodologies
- Chip assembly -- possibly implicit in above steps

Global Placement Overview

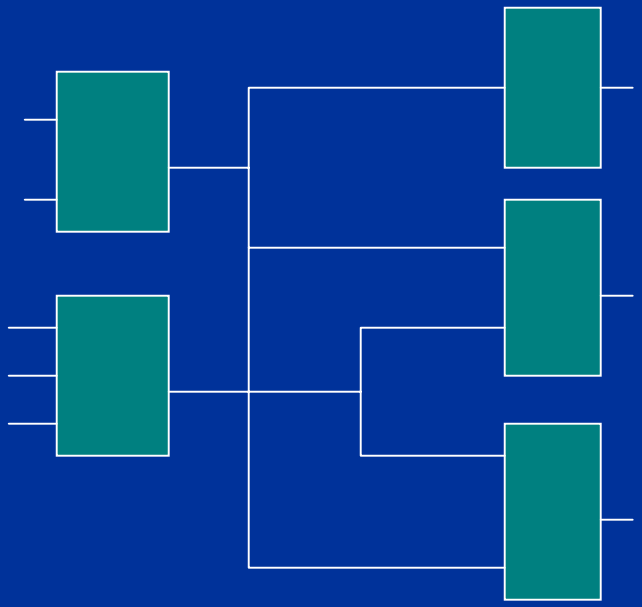
- Context
  - timing- and routability-driven placement of 10^6 cells and up
  - interconnect more important than transistors
- Formalization
  - weighted hypergraph represents netlist
  - cell shapes ignored; cells can overlap
  - constrained vertex locations, e.g., I/O pads
  - minimize objective function of unknown vertex locations

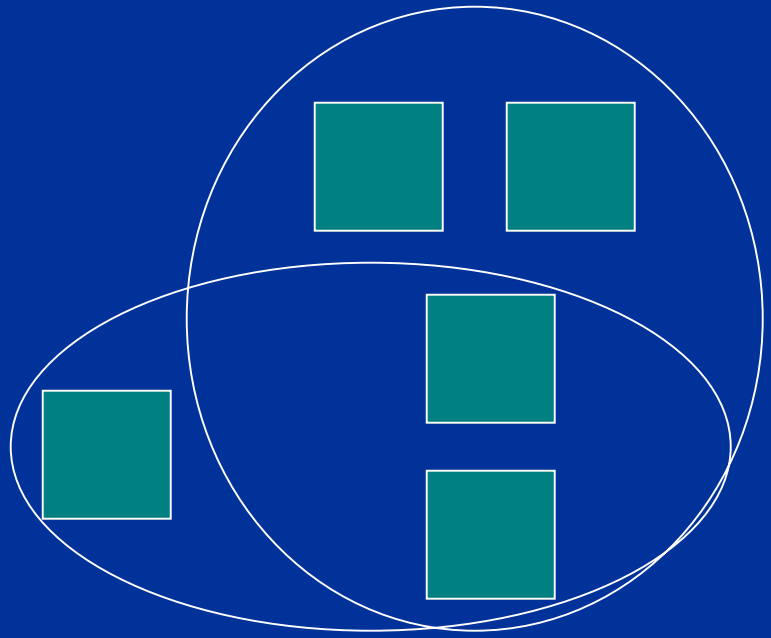
Global Placement Overview

- Cell areas must be "distributed uniformly"
- Top-down hierarchical placement
  - solve a “top-level” problem first
  - apply successive refinements
  - e.g., divide/conquer: split design in two pieces, then split each part, continue recursively until pieces are trivial
- Analytic placement
  - based on mathematical programming, e.g., minimize objective function by funding zeros of derivative

Placement Model

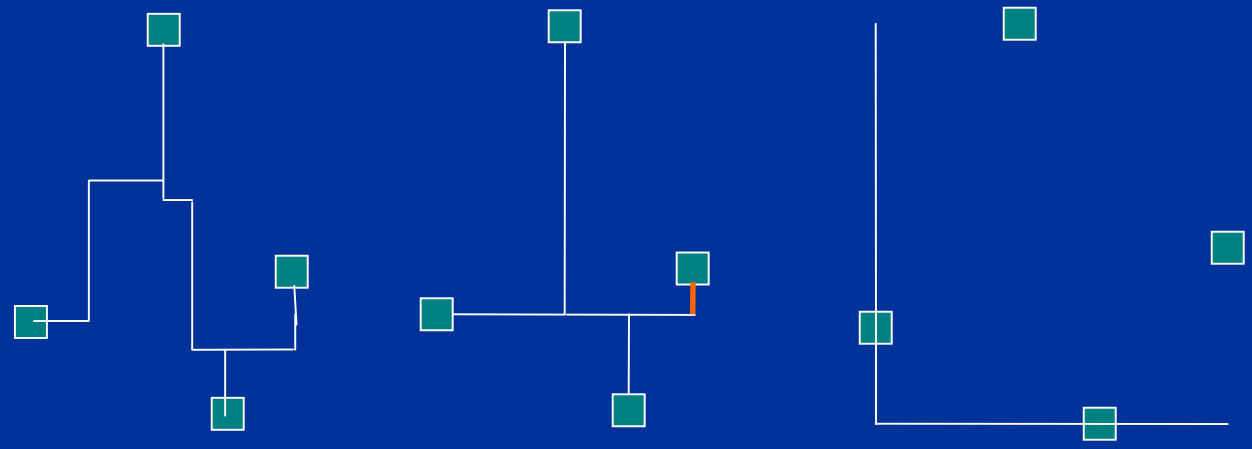
- Hypergraphs
  - netlist represented by hypergraph
  - cells represented by vertices (“with area”)
  - all pins on a cell are placed in the center





Placement Model

- Objectives
  - Rectilinear Steiner Minimal Tree (RSMT)
  - half-perimeter wirelength (BBox)
  - routing congestion



Approaches To Placement

- Top-down partitioning based
  - divide and conquer strategy
  - divide = hypergraph partitioning
- Simulated annealing
  - iterative-improvement move-based
- Analytical
  - LP-style approach
- Hybrids are of course possible

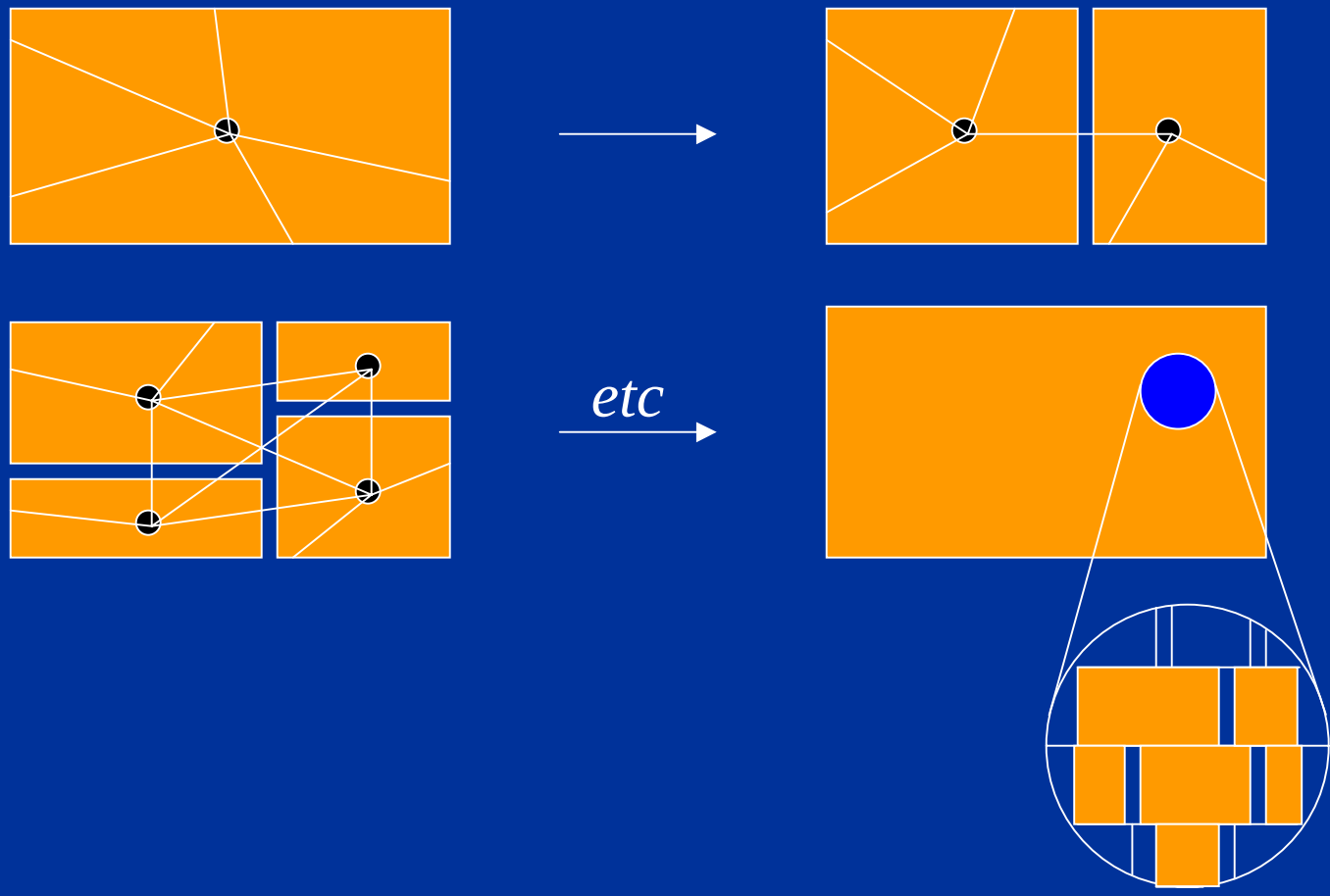
Top-Down Placers

- Partitioning-driven placers: divide/conquer
  - analytic engines can be used as plug-ins
  - annealing can be used as post-optimization
- Core algorithms
  - min-cut partitioning of large hypergraphs
  - end cases, e.g., 15 cells
- Modern implementations scale well, parallelize naturally

Top-Down Placers

- Use model
  - batch mode (no support for interactivity or ECO)
  - some constraints handled well, but not timing-critical paths
  - SA post-processing (detailed placement) to satisfy additional constraints
- Performance
  - reasonably fast; best quality of several starts is stable
  - basis for leading-edge commercial tools

Top-Down Placer Detail: Hypergraph Partitioning



Top-Down Placer Detail: Hypergraph Partitioning

- Balanced hypergraph partitioning is NP-hard
- Randomized heuristics with many starts
- Best ones based on Fiduccia-Mattheyses 82
  - spectral, annealing, etc. methods not competitive
- Significantly improved in last 2 years with multilevel FM
- runtime for circuits of 10^6 nodes: 10 seconds

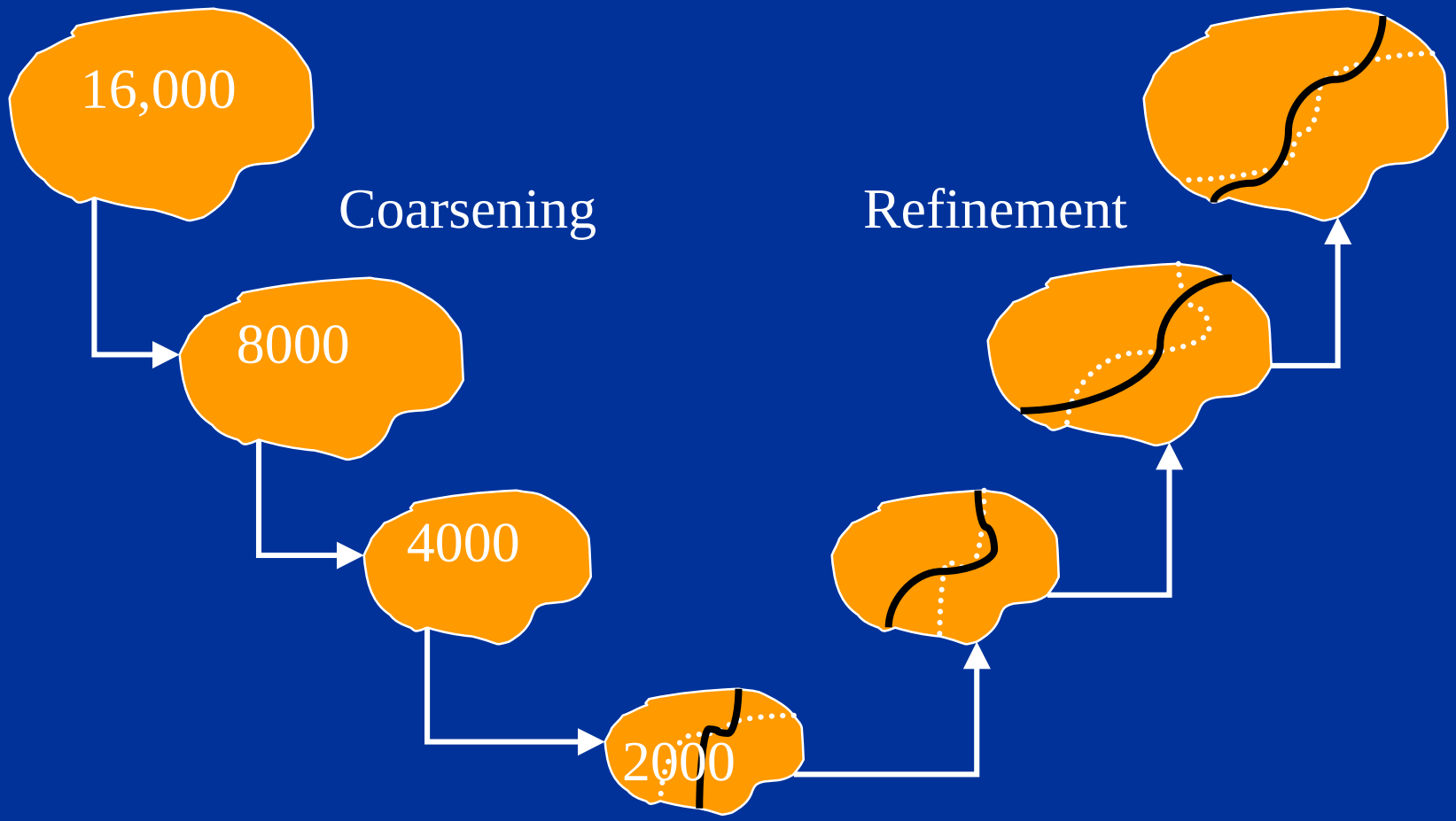
Fiduccia-Mattheyses Approach

- Fiduccia-Mattheyses (1982)
  - start with some initial solution
  - perform passes until a pass fails to improve solution quality
- Pass:
  - start with all vertices free to move to the other partition (unlocked)
  - label each possible move with immediate change in cost that it causes (gain)
  - iteratively select and execute a move with highest gain, lock the moving vertex, and update gains
  - best solution seen during the pass is adopted as starting solution for next pass

Fiduccia-Mattheyses Approach

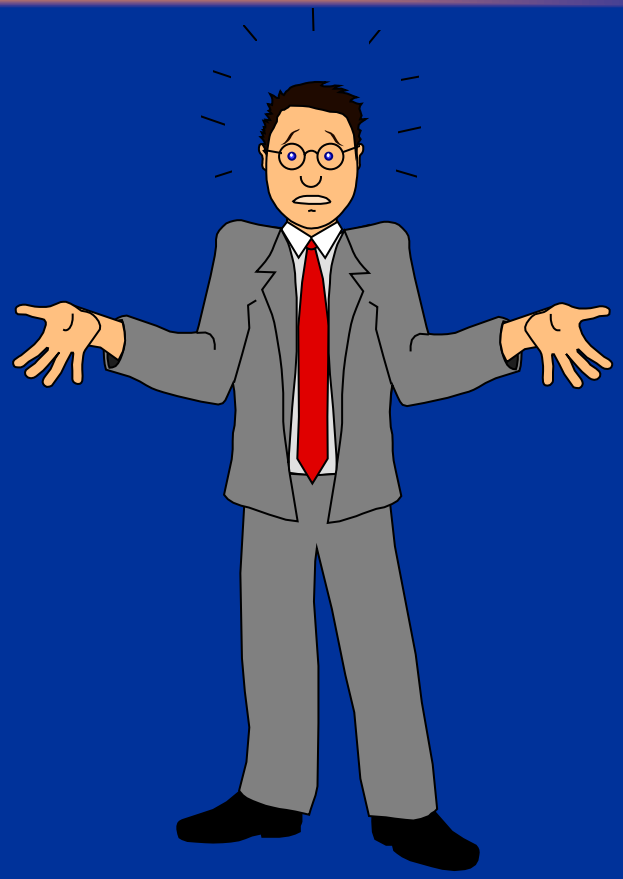
- Key elements:
  - three main operations
    - computation of initial gain values at beginning of pass
    - retrieval of the best-gain (feasible) move
    - update of all affected gain values after a move is made
  - contribution of Fiduccia and Mattheyses
    - circuit hypergraphs are sparse
    - move gain is bounded between +2 * , -2 * max vertex degree
    - hash moves by gains (gain bucket structure)
    - each affected gain can be updated in constant time
    - linear time complexity, O(#pins), per pass

The Multilevel Approach



You Must Have Multilevel

- Extremely efficient
- More chances to climb out of local minima
- Superior solution quality
- Flexible objectives
- Scales with circuit size



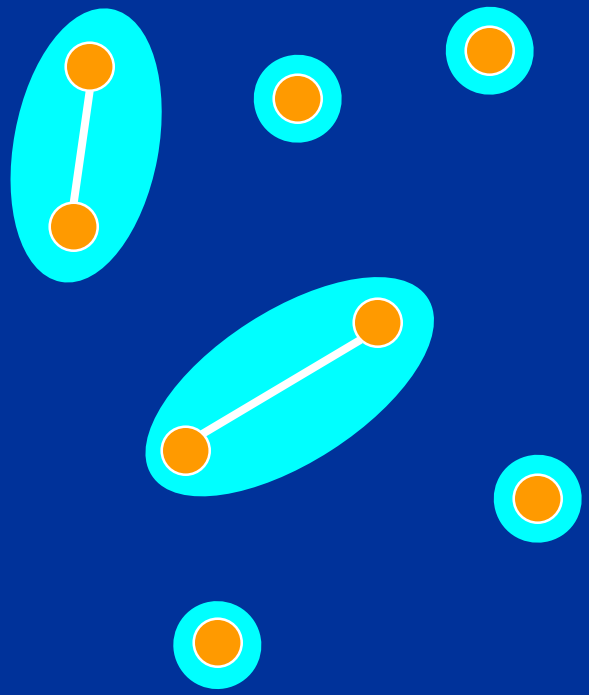
Multilevel approaches are the industry standard

Previous Work

- Scientific Computing
  - Chaco (Hendrickson and Leland, 1993)
  - Metis (Karypis and Kumar, 1995)
- VLSI CAD
  - Cong and Smith, 1993
  - Hauck and Borriello, 1995
  - MLc (Alpert and Kahng, 1996)
  - hMetis (Karypis and Kumar, 1997)

Matching-based Coarsening

- Randomly sort vertices
- Try to match current vertex to best vertex
- Stop when number of matched vertices < M|V|



Multilevel FM and Advanced Techniques

- Key implementation decisions
  - clustering
  - tie-breaking
  - handling balance constraints and cell/cluster areas
  - efficient data structures and pitfalls
- Other objectives
- Other issues: relaxations, 2-way vs. k-way, etc.

Circuit Clustering

- Taxonomy of top-down and bottom-up circuit clustering methods
- Hierarchical edge- and hyperedge clustering methods
  - heavy-edge matching
  - absorption objective
  - continuous vs. batched updating

Analytic Engines

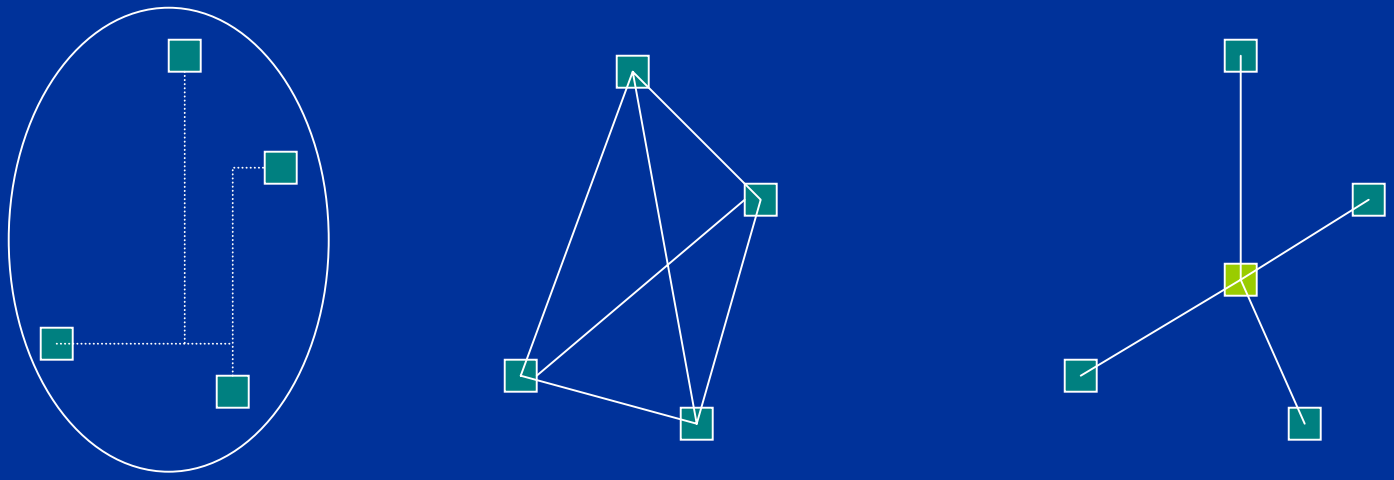
- Core algorithms
  - minimization of convex functions
  - linear algebra: solving sparse linear systems
  - very well studied in applied mathematics, deterministic
- Use models
  - simple objective functions and linear constraints supported
  - discrete constraints are hard to deal with
  - little or no support for interactivity and ECO
  - solutions can only be interpreted as hints to other placers (too many cell overlaps; solutions must be "legalized")

Analytic Engines

- Performance
  - deterministic; predictable runtime and stable quality
- Additional benefits
  - major components (sparse-matrix solver) available off-the-shelf
  - codes are small

Analytic Placement Details

- Reduction of hypergraphs to graphs
  - clique and star models for nets



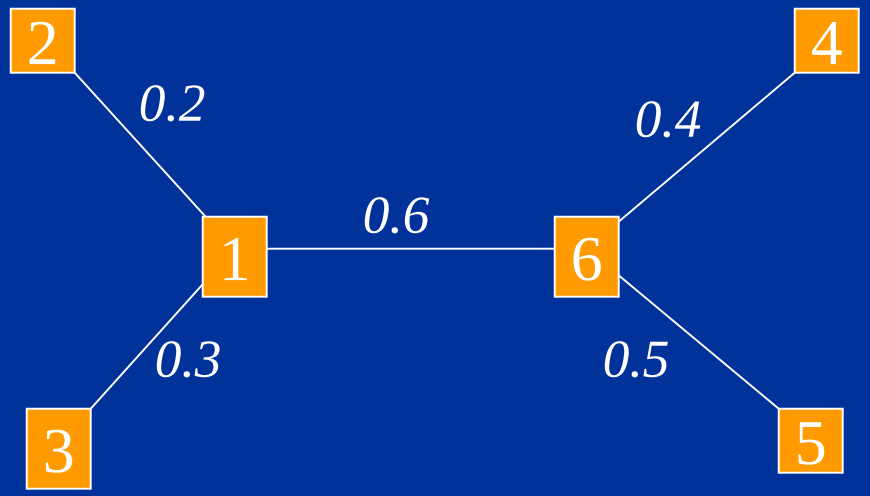
- Objective functions
  - total weighted "wirelength" of all edges
    - linear (Manhattan) WL
    - squared (Euclidean) WL

Analytic Placement Details

- Linear vs. quadratic WL
  - linear (Manhattan) WL
    - closer to reality: better models timing and congestion
    - hard to minimize; multiple optimal solutions exist
    - can be approximated by a sequence of squared WL functions
  - squared aka quadratic (Euclidean) WL
    - always unique optimal solution (strictly convex function)
    - very easy to minimize: set derivative to zero
    - numerics: solve one system of linear equations
    - implementations easily available

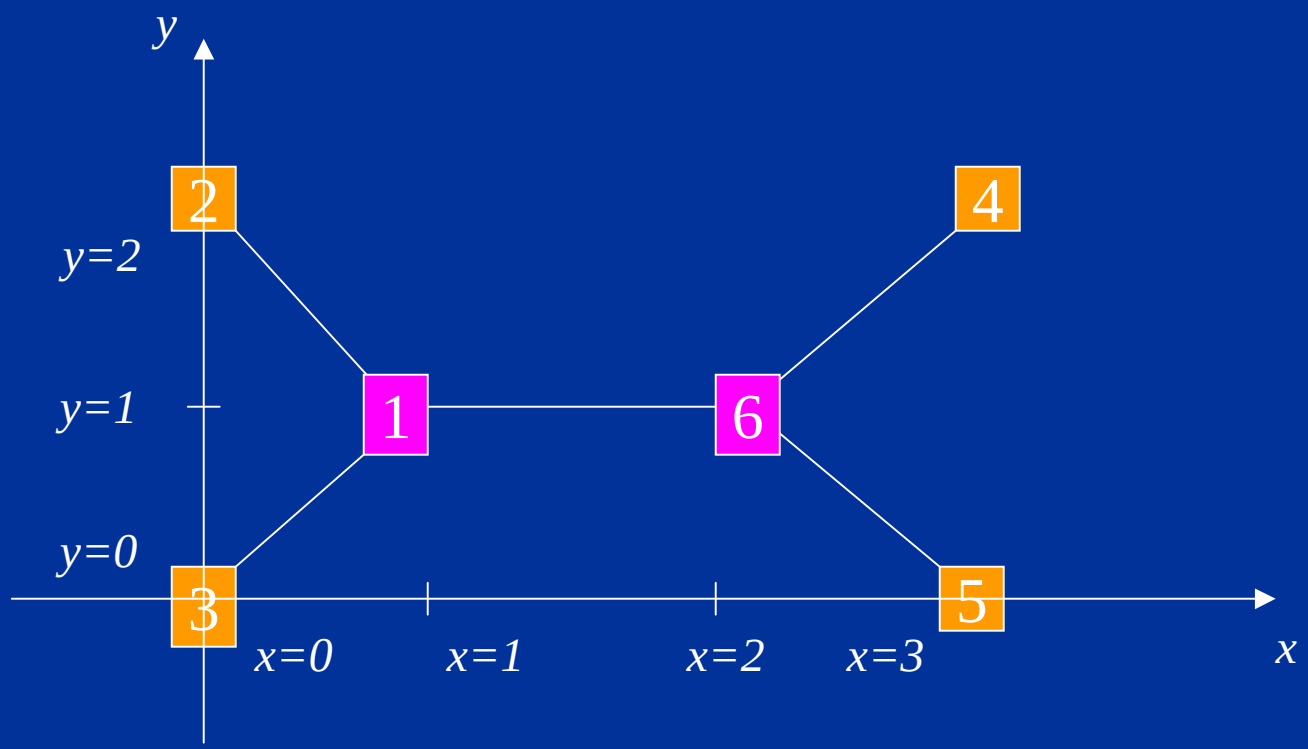
Analytic Placement Details, Contd.

- Laplacian and incidence matrices



Analytic Placement Details, Contd.

- Linear constraints



Analytic Placement Details, Contd.

- One-dimensional WL minimization

- GORDIAN-L, Weiszfeld iteration, and regularizations
  Force-directed placement

Detailed Placement

- Detailed placement optimizations
  - EEQ/LEQ substitution
  - module orientation
  - shifting/alignment
- Routability and wiring estimation
  - A priori, on-line and a posteriori wiring estimators for placement

Effective Optimization Strategies For Large-Scale VLSI Placement

- Circuits as hypergraphs
- Circuit placement ≈ hypergraph placement
  - generalizes Weber problem that seeks a location to minimize sum of distances to fixed points
  - first analytical algorithm due to Weiszfeld 1937
- In general: some vertices fixed, others movable
  - minimize a function of vertex locations
  - modern algorithms exploit Newton's method

VLSI Placement

- Large circuit hypergraphs, limited CPU time
- "*Circuit delay*" to be minimized
- *Interconnect delay* dominates device delay
- Wirelength
  - represents interconnect delay and area
  - *x*- and *y*- components often independent
  - minimization formulated as convex optimization

VLSI Analytical Placement

- Analytical algos useful with non-linear terms present
  - more versatile than min-cut placement
  - faster than Simulated Annealing
- BBox: common objective function (wirelength est.)
  - for one hyperedge: half-perimeter of the bounding box of incident vertices
  - sum over all hyperedges
  - not everywhere differentiable
  - can be complemented by other, e.g., non-linear terms
  - typically dominates other terms

Traditional Approach

- ***Main features:***
  - ***conversion*** of hypergraphs to graphs
    - hyperedges → cliques or stars
  - hyperedge-based objective functions substituted by edge-length variants
    - e.g., “quadratic”, “linear” wirelength etc
- Drawbacks:
  - optimal solution not practical; heuristics applied
  - not as fast as leading optimization techniques
  - optimal solutions are not feasible

Quadratic vs Linear Wirelength Minimization

- $\min_{\mathbf{x}} \Sigma_{i>j} a_{ij}(x_i - x_j)^2$ subject to $\mathbf{Hx} = \mathbf{b}$
  - **x** = unknown node positions, **H** = linear constraints
  - **Benefits**: objective function is differentiable and convex
    - Fast unique solution (PROUD [Tsay et al. `88])
  - **Drawback**: questionable relevancy
- $\min_{\mathbf{x}} \Sigma_{i>j} a_{ij}|x_i - x_j|$ subject to $\mathbf{Hx} = \mathbf{b}$
  - **Benefits**: better model of routed wirelength
    - *Mahmoud et al.* `94
  - **Drawbacks**: not differentiable, and nonconvex
    - typically *many* minimizers
    - minimized by slow linear programming or heuristically by *GORDIAN-L* (*Sigl* et al, *DAC `91*)

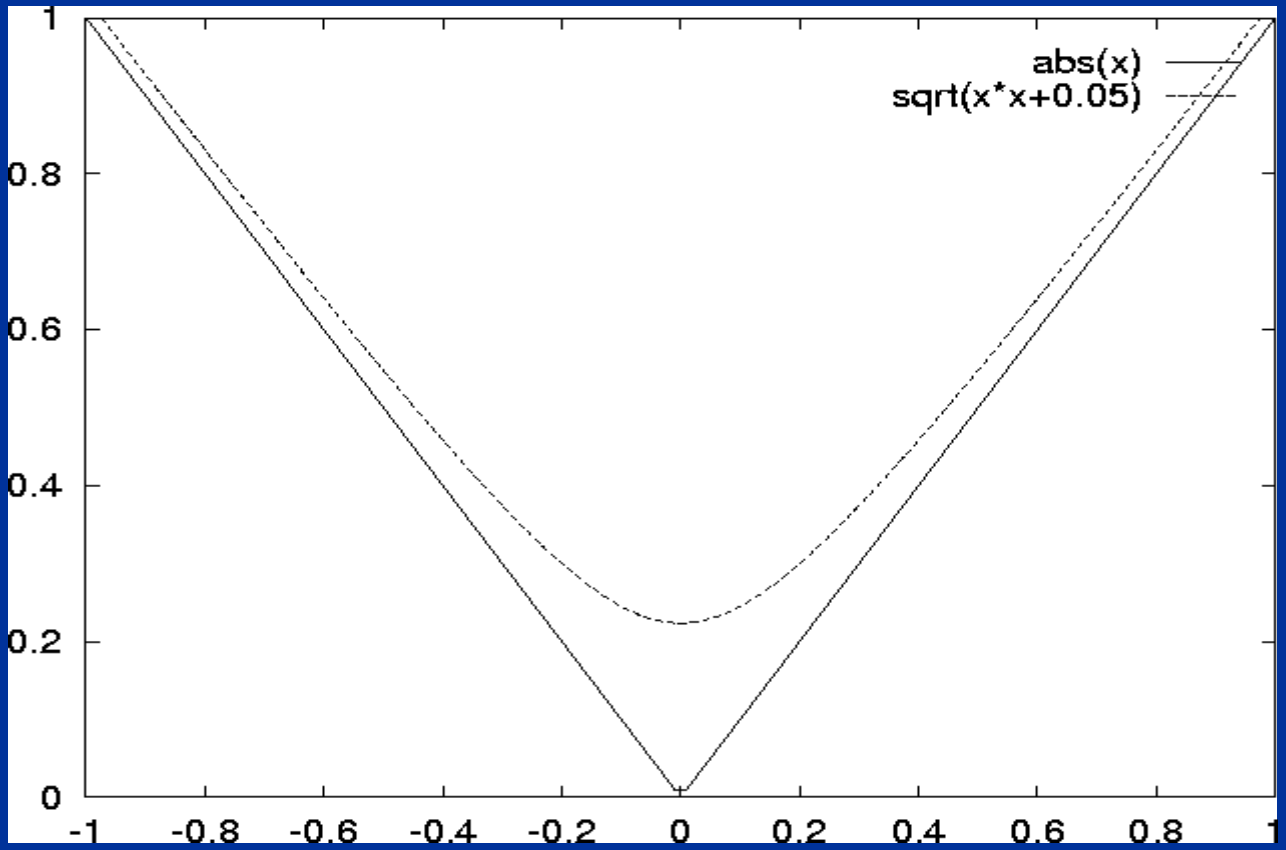
Smooth Approximations

- **Problem**: combine benefits of both objectives
- **Solution**: smooth approximations
  - high accuracy
    - *minimizers* must be "very close"
  - quickly computable (= free of numerical problems)
    - twice continuously differentiable
    - partials not too large
- **Problem**: combine accuracy and speed
- **Solution:** parameterized approximations:
  - trade-offs between approximation quality and runtime

β-regularization and "Weiszfeld method"

- *Alpert, Chan, Kahng, Markov et al., ISPD-97*
- Regularization: $|x| \rightarrow (x^2+\beta)^{1/2}$
- $\beta > 0$ gauges trade-off: quality *vs* run time
- GORDIAN-L a special case $\beta = 0$ of Weiszfeld iteration (*Eckhardt `80*)
- Regularization allows for faster numerical methods (see *Proc. ISPD-97)*

Simple Regularization



Symbolic Regularization

- Look at symbolic representation of objective function
- Find symbolic fragments responsible for singularities
- Relevant fragments often are
  - **univariate** functions
  - absolute value or more general **case analysis**
- Hence our interest in **piece-wise linear functions**
- Approximate ("**regularize**") the fragments
  - e.g., send $|x|$ into $(|x|^p+\beta)^{1/p}$
- Produce new symbolic representation by substituting in approximations of fragments

  e.g., $min(a,b)=(a+b-|a-b|)/2$ by $(a+b-((a-b)^p+\beta)^{1/p})/2$

Symbolic Regularization

- Original function

$$F(x) = \begin{cases} F_1(x - x_0) - C, & \text{if } x \geq x_0; \\ F_2(x - x_0) - C, & \text{if } x < x_0; \end{cases}$$

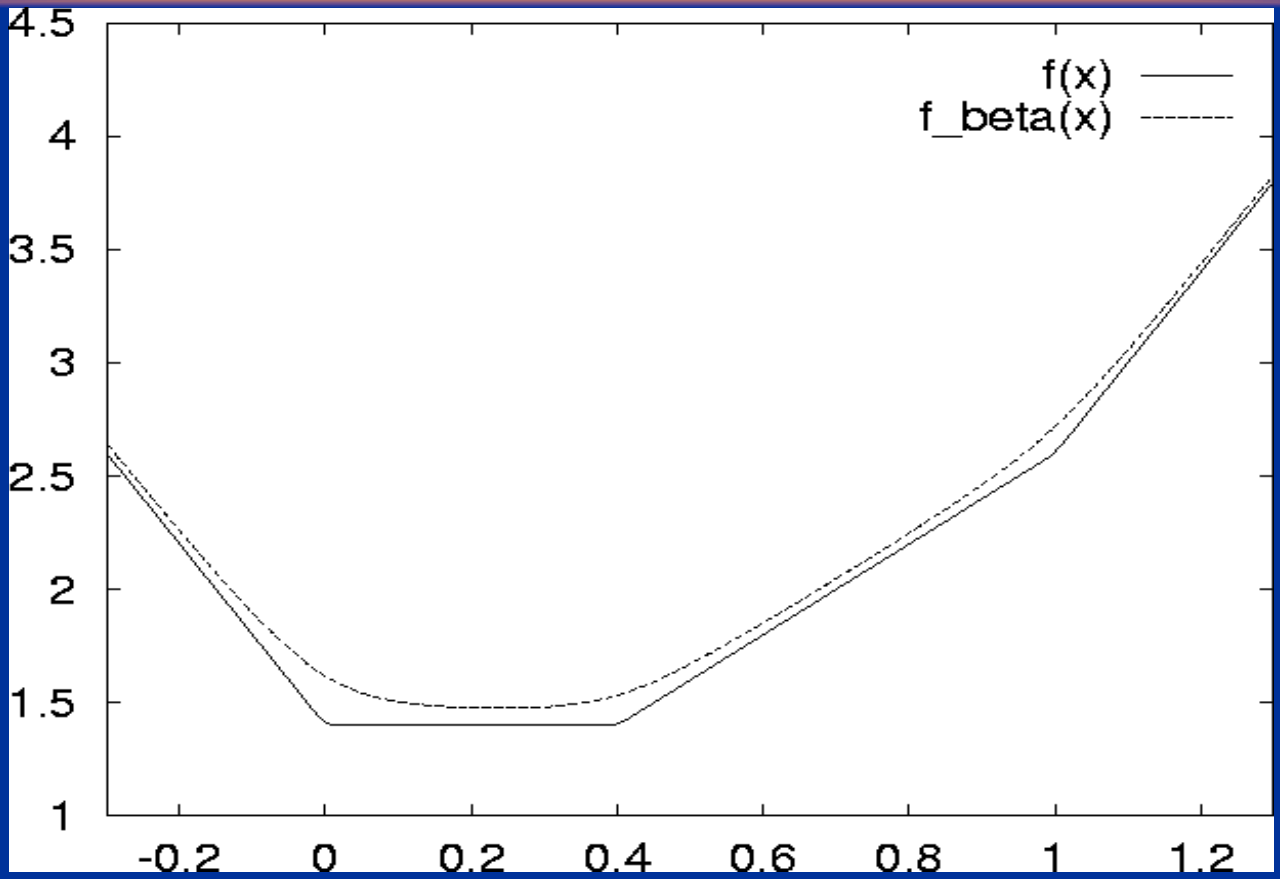
- $F1(t), t \leq 0$ and $F2(t), t \geq 0$
- $F_1(t)$ and $F_1(t)$ continuously differentiable and convex
- $F_1(0) = F_2(0) = 0$, but not necessarily $F'_1(0^+) = F'_2(0^-)$
- arbitrary C

Regularization $F_\beta(x) = C - (|F(x) - C|^p - \beta)^{\frac{1}{\gamma}}$

Univariate Piece-Wise Linear Functions

- Assume orig. function convex (not nec. *strictly* convex!)
- Present original function as sum of *convex PWL* funcs having at most one cusp at $x=x_0$ (simple functions)
- Regularize simple functions one by one
  - send $f(x)$ into $f_\beta(x)=f(x_0)+((f(x)-f(x_0))^p+\beta)^{1/p}$
  - slightly different regularization if $f(x)$ has plateaux
- **Theorem:** regularized func smooth and *strictly* convex
- **Theorem:** overestimates original func by at most $\beta^{1/p}$
- **Theorem:** minimizers converge $lim_{\beta\to 0}\ inf_x\ f_\beta(x) = inf_x\ f(x)$

Example of b-regularization



$f(x)=|1-x|+2|x|+|x-0.4|$

$p=2,\ \beta=0.01$

Multivariate Functions

- Can typically be represented symbolically with univariate functions
- examples of multivariate regularization
  - $max\{a,b\} = \frac{a+b+|a-b|}{2}$ regularized with $\frac{a+b+(|a-b|^p+\beta)^{1/p}}{2}$
  - $min\{a,b\} = \frac{a+b-|a-b|}{2}$ regularized with $\frac{a+b-(|a-b|^p+\beta)^{1/p}}{2}$
  - $max\{a,b,c\}$ regularized via $max\{e, max\{b,e\}\}$ (wirelength)
  - $max\{(a-b)^2, (a-b)^2\} = a^2-b^2-2|ab|$ regularized with $a^2-b^2-2\sqrt{a^2b^2-\beta}$ (delay)
  - $max\{0,(t-t0)\}$ regularized with $\frac{(t-t_\beta)+\sqrt{(t-t_\beta)^2+\beta}}{2}$ (excessive for delay)

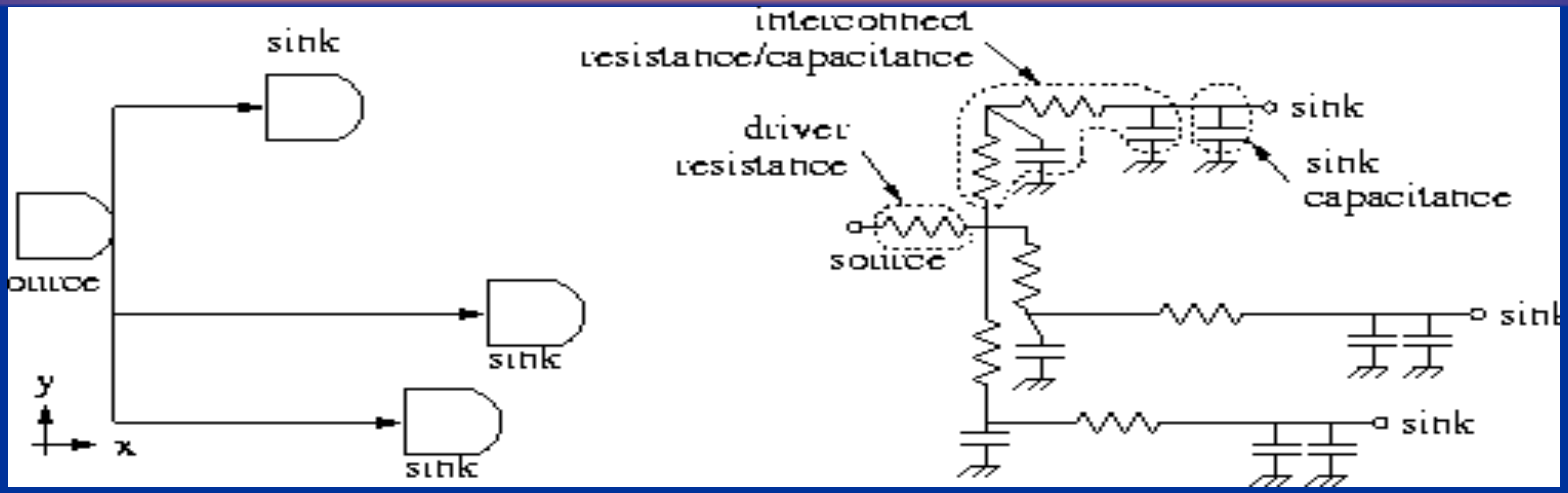
Practical Issues

- Minimizers of the regularization often minimize the original function (no error)
- Different trade-offs in $(x^2+\beta)^{1/2}$ with different scaling of x
  - $\beta = \beta_0 |x|^2$
  - β_0 – scale-independent component
  - 10^{-8} through 10^{-1}

Benefits of New Approach

- Overcome difficulties in objective functions
  - non-smoothness
  - non-convexity
- Overcome difficulties in minimization algorithms
  - no need to create specialized algorithms
  - can apply newton-type methods
    - primal-dual

Lumped-L Elmore Delay



- Three components of Elmore delay
  - source resistance times all downstream capacitance
  - interconnect resistance times sink capacitance
  - interconnect resistance times interconnect delay

$$r_x c_x \chi^2 + r_y c_y \gamma^2 + r_y c_x |\chi\gamma|$$

r_x, r_y - per-unit resistance

c_x, c_y - per-unit capacitance

χ, γ – source-to-sink distances

Delay Approximation

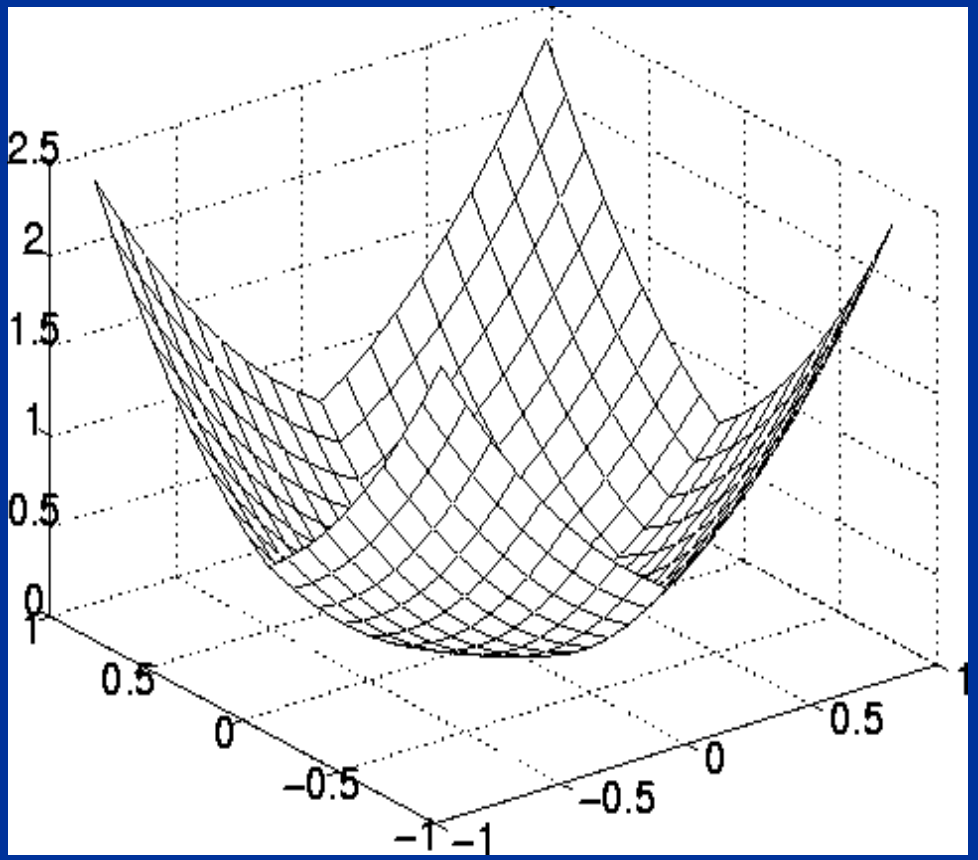
- Third term of Elmore delay neither smooth nor convex:

$$r_x c_x \chi^2 + r_y c_y \gamma^2 + r_y c_x |\chi\gamma|$$

- This is convex when $4c_y/r_y > c_x/r_x$ (!)
  - actually the case for any relevant technology

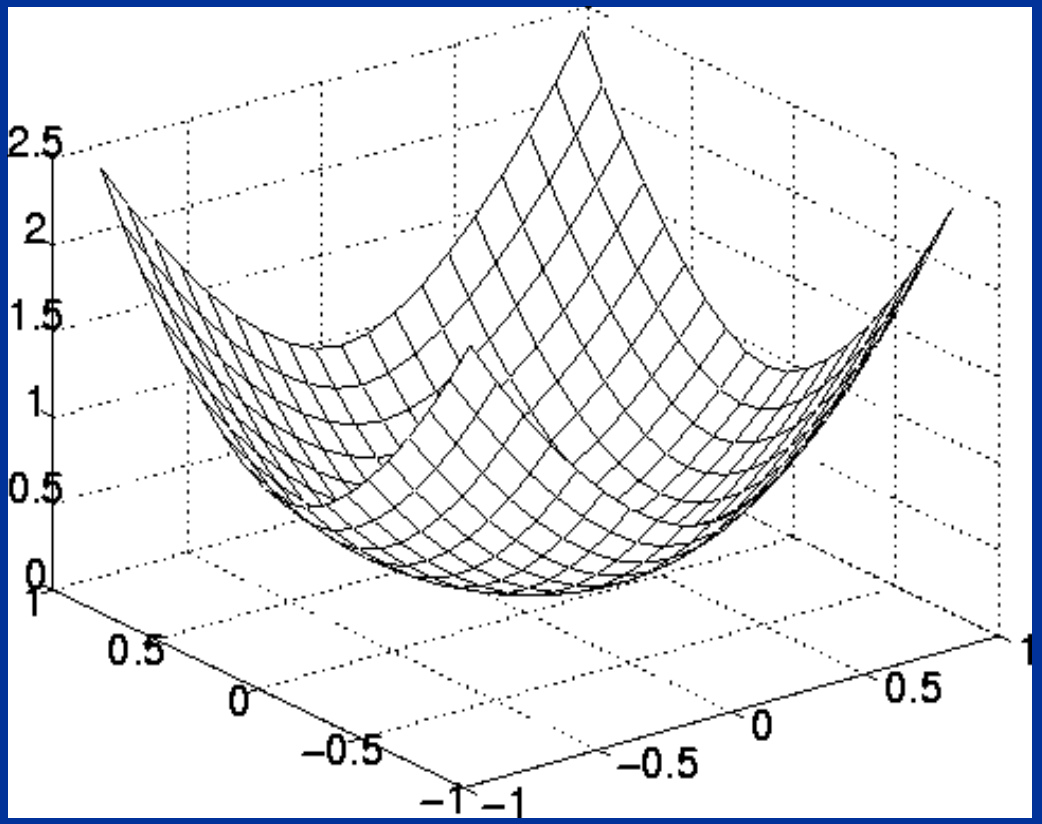
Delay Function Has Cusps

- $r_x c_x \chi^2 + r_y c_y \gamma^2 - r_y c_x |\chi\gamma|$
- Now need to regularize
- Non-linear cusps to be eliminated



Regularization Removes Cusps

- $r_x c_x \chi^2 + r_y c_y \gamma^2 - r_y c_x ((\chi\gamma)^2+\beta)^{1/2}$
- Regularized



- No cusps

Placement Directions

- Global placement
  - engines (analytic, top-down partitioning based, (iterative annealing based) remain the same
  - becomes more hierarchical
    - block placement, latch placement before “cell placement”
  - support placement of partially/probabilistically specified design
- Detailed placement
  - LEQ/EEQ substitution
  - shifting, spacing and alignment for routability
  - ECOs for timing, signal integrity, reliability
  - closely tied to performance analysis backplane (STA/PV)
  - support incremental “construct by correction” use model

Taxonomy of Routing Approaches

- Gridded vs. gridless
- Area-based vs. channel-based
- Full-chip vs. switchbox
- Many details
  - search: BFS (A* or maze) vs. DFS (line probe) vs. pattern-based
  - metaheuristic: iterative (recost/ripup/reroute) vs. combinatorial (multicommodity flow, LP+rounding)
  - resouce model: right-way vs. wrong-way
- High-capacity batch ASIC
  - gridded, area-based, N-layer, symbolic, switchbox, global+detailed, A* search, iterative ripup/reroute
- Lower-capacity, auto-interactive, full-custom/CA/PCB
  - gridless, shape-based, full-chip

Global Routing Overview

- Problem Statement: Given a netlist and placement, connect the pins subject to constraints, design rules, etc.
- Reduction to point-to-point connections
- Overview of search methods
- Graph models on which to search

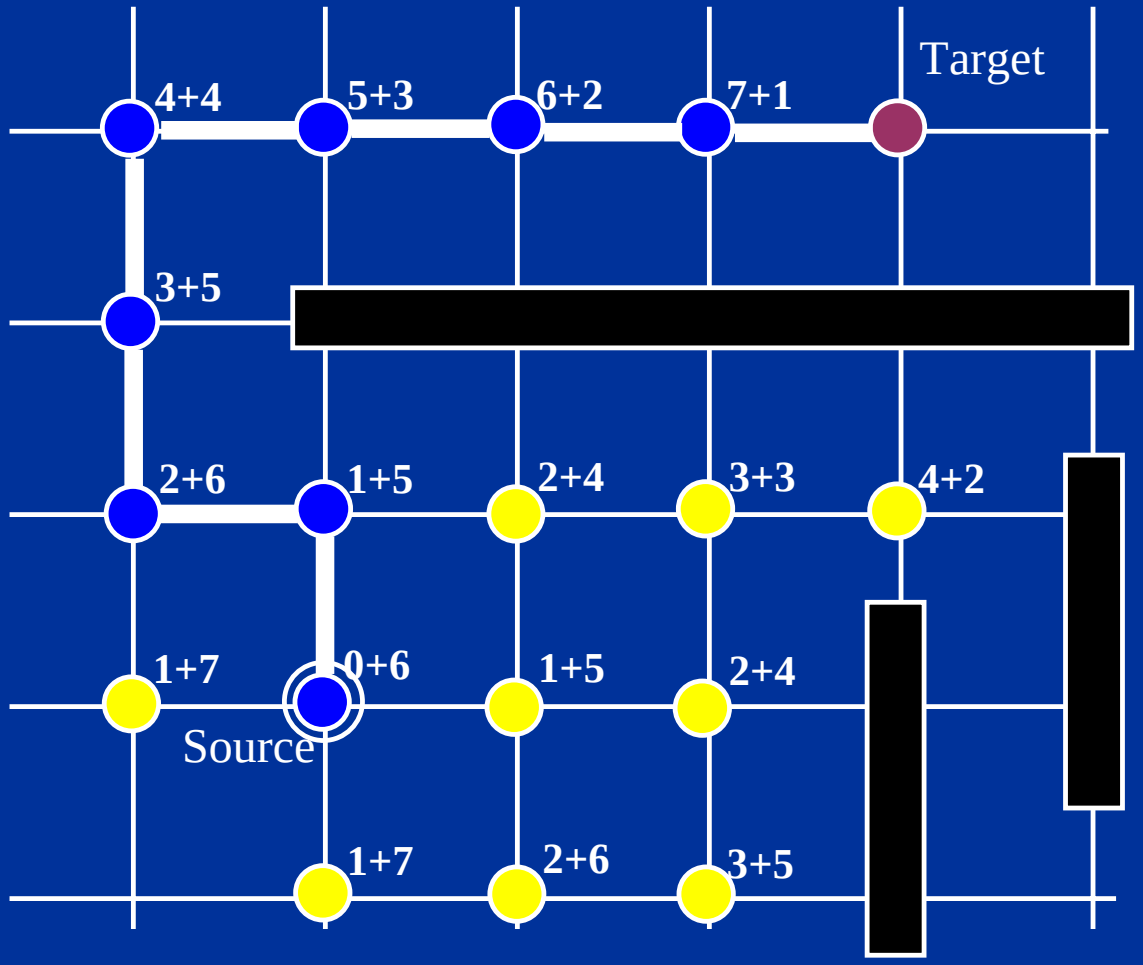
Point-to-Point Connections: A*

- Heuristic search (uses knowledge about the search space)
- Expands nodes in (cost+estimate of remaining cost) order
- cost(n) = g(n) + h(n)
  - g(n) = cost of path from source to n
  - h(n) = estimate (lower bound) on remaining cost from n to target
  - BFS = A* with h(n) = 0
  - if h(n) = actual cost from n to target A* will expand only nodes in the shortest path
- Typical heuristic might be Manhattan distance (assuming regular unit-weighted grid)

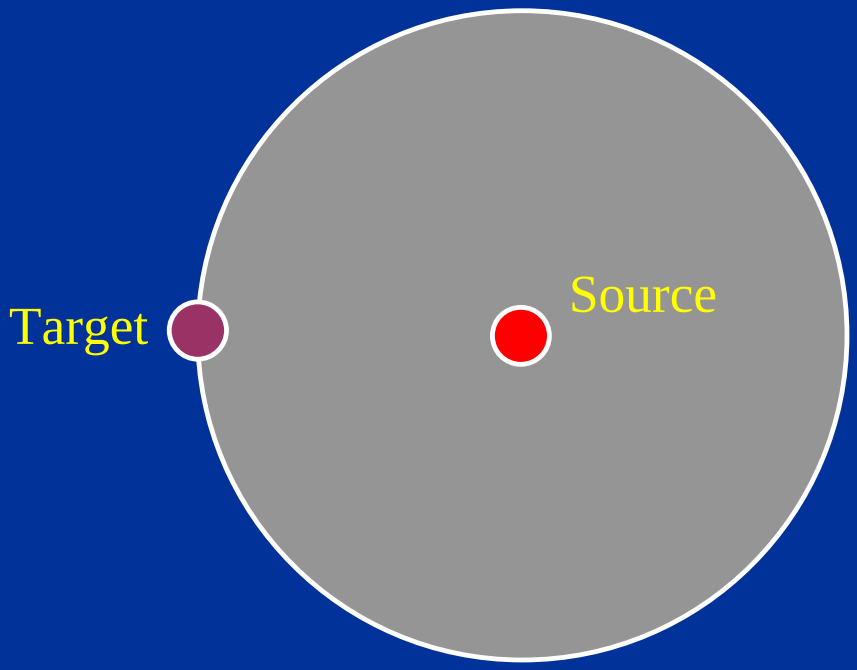
Point-to-Point Connections: A* cont.

- Advantages:
  - much faster than BFS
  - will always find a lowest cost path (if one exists), given that h(n) is 'non-overestimating' (a lower bound)
- Disadvantages:
  - high storage complexity
  - more complex to implement than BFS

A* Example

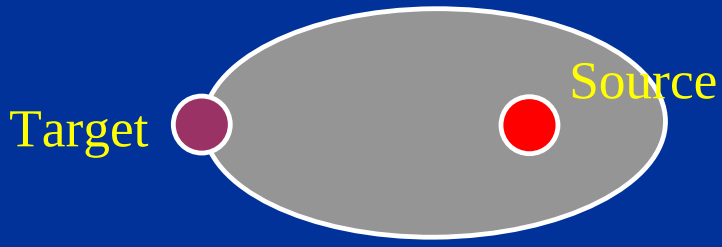


BFS vs. A*



Breadth First Search

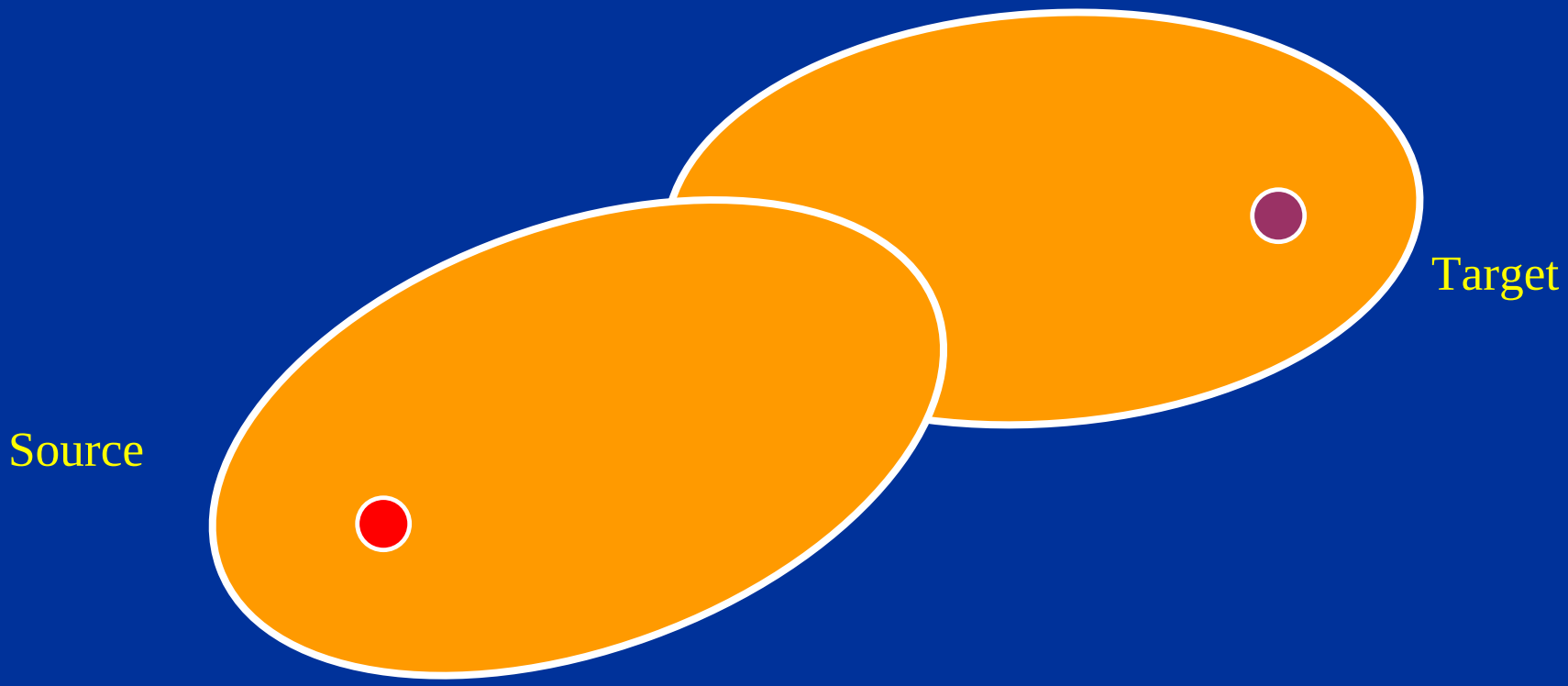
A* Search



Bi-directional A*

- 'Grow' an A* search from both the source and the target
- Terminates when a node has been selected for *expansion* by both search fronts (*not* when they first meet)
- Search fronts CAN miss
  - if there are multiple min-cost paths
  - result is doubling of # nodes expanded
- Advantage
  - Potentially a large savings in number of nodes expanded(runtime)
- Disadvantage
  - Implementation less clear

Bi-directional A*

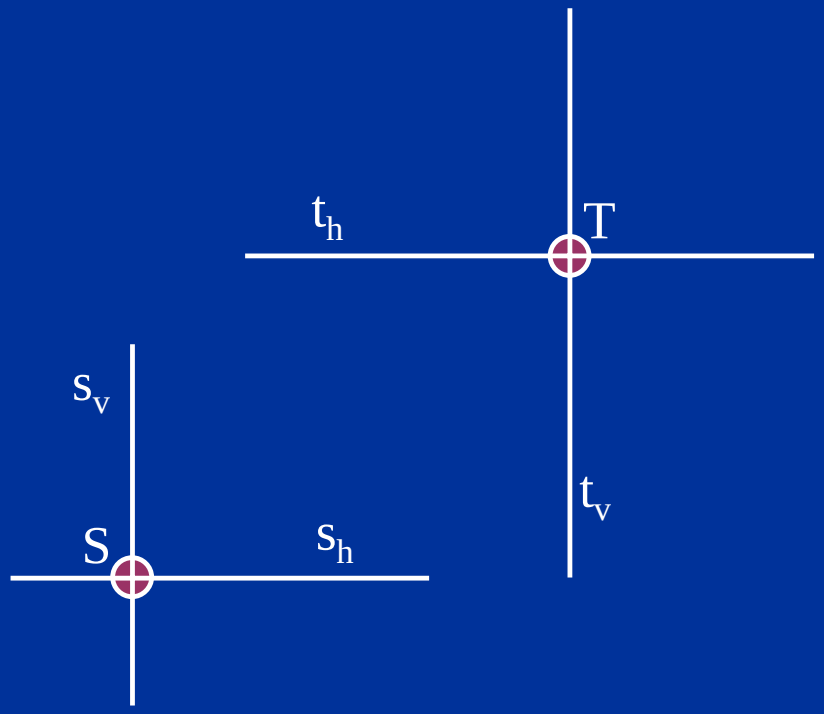


Point-to-Point Connection: Line Expansion

- Reference: Hightower 1969.
- Line Expansion = DFS
- Advantages
  - lower memory usage: O(L) (L is number of line segments)
  - fast: O(L)
  - works well in a gridless environment
- Disadvantages
  - not guaranteed to find a path, even if one exists
  - there are line-expansion methods that will, however

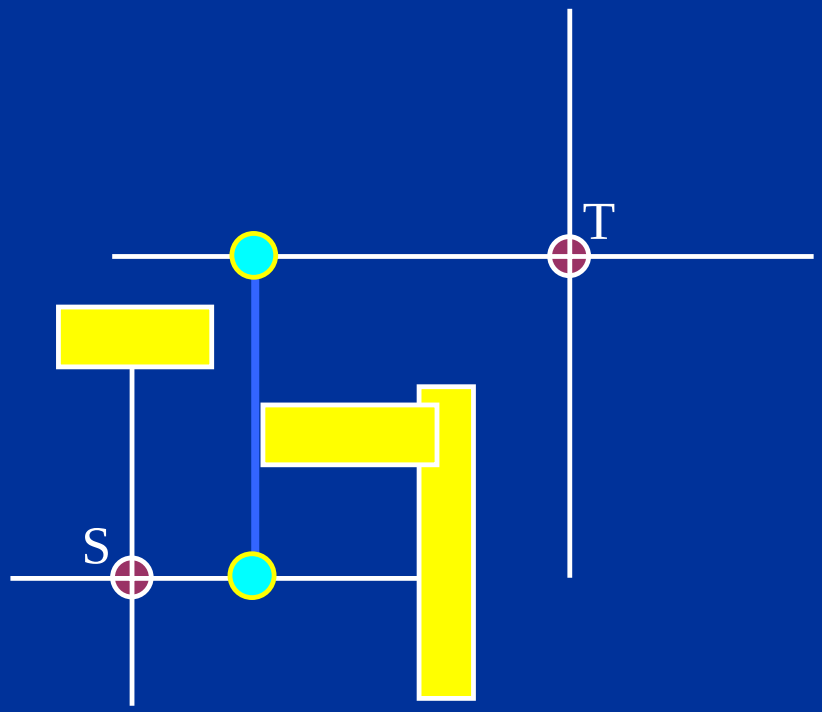
Line Expansion Cont'd

- Observe that a rectilinear path from S to T must include segments from either s_h or s_v, and t_h or t_v.
  - The second segment from S must then be perpendicular to s_h or s_v.
  - the (2) shortest possible paths (assuming no obstacles) are S_h+T_v and S_v+T_h.



Line Expansion Cont'd

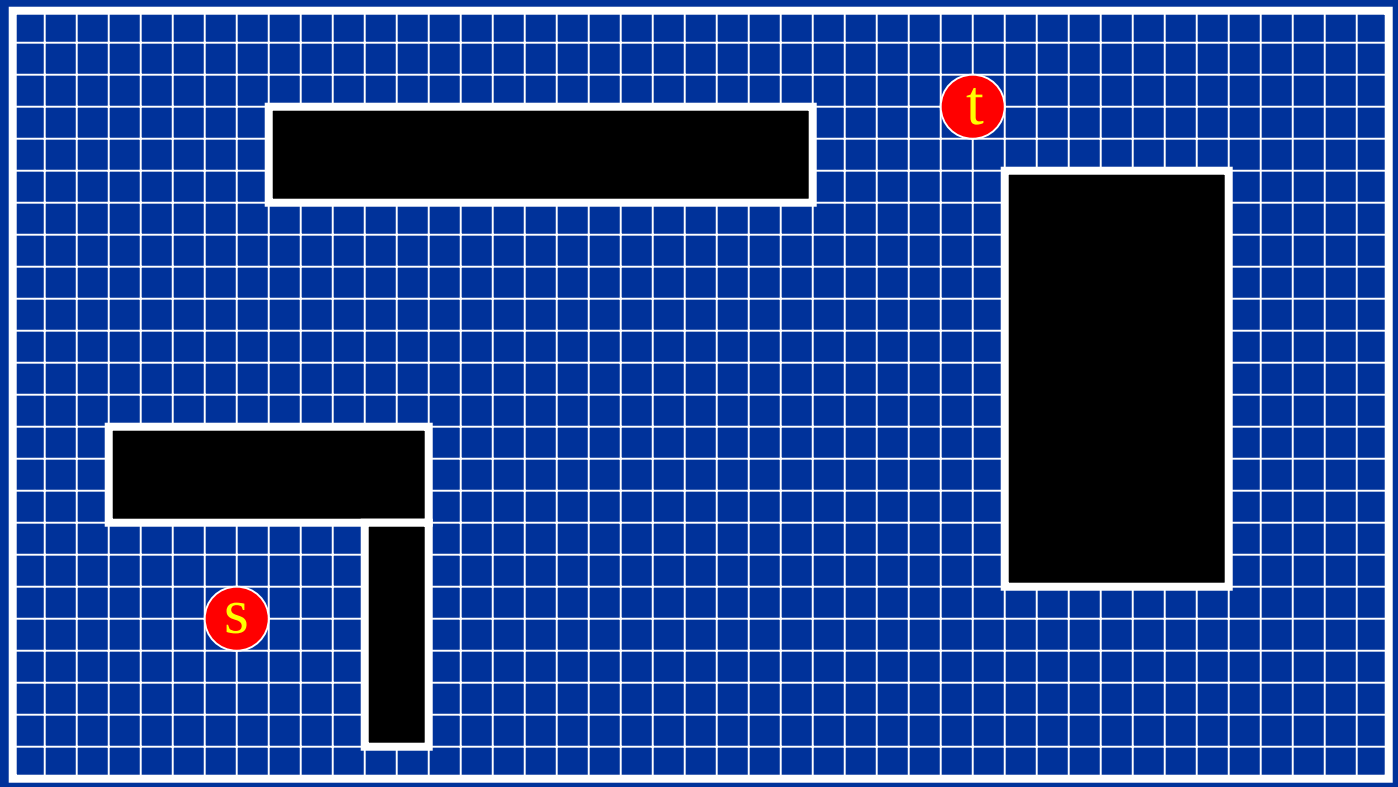
- Finding expansion route around obstacles



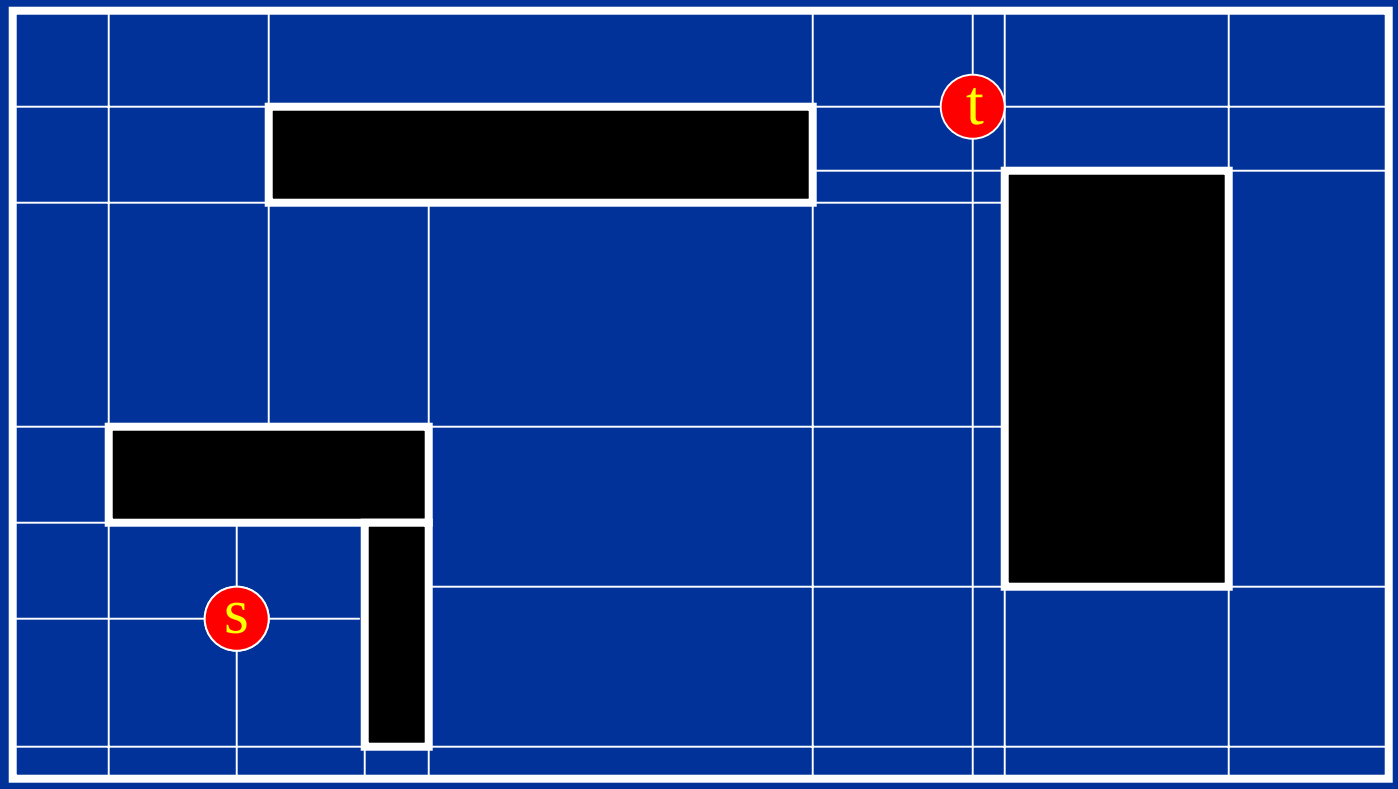
How To Model Resources?

- Complete (unit) Grid
  - store all possible paths a route could take
  - high memory overhead
  - simple model
- Connection Grid
  - only some gridlines need be stored/searched
  - a ‘strong connection graph’ guarantees that the shortest path can be made using only lines in the graph
  - lower memory overhead
- Implicit Connection Grid (‘Gridless’)
  - connection grid can be generated on the fly, as needed
  - lowest memory overhead
  - improves runtime for some algorithms!!

Complete (Unit) Grid



Connection Grid



Implicit Connection Graph

- S.Q. Zheng, et. al TCAD 1996
- Generates the connection grid ‘on the fly’
  - a.k.a. Gridless
  - saves memory - avoids storing large graph for short nets
- Key operation - find adjacent nodes
  - given: a node *n* in the connection graph
  - produce: all neighboring nodes to n
- Operation find_neighbors
  - Lu, Lv be the set of all vertical, horizontal line segments
  - find the (at most 2) members of each set intersecting *n*
  - trace each segment, starting at *n*, looking for the next intersection with a member of the other set
  - using balanced binary tree, can be done in O(log *e*)

Out-of-Box Uses of Routing Results

- Modify floorplan
  - floorplan compaction, pin assignments derived from top-level route planning
- Determine synthesis constraints
  - budgets for intra-block delay, block input/output boundary conditions
- Modify netlist
  - driver sizing, repeater insertion, buffer clustering
- Placement directives for block layout
  - over-block route planning affects utilization factors within blocks
- Performance-driven routing directives
  - wire tapering/spacing/shielding choices, assumed layer assignments, etc.

Routing Directions

- Cost functions and constraints
  - rich vocabulary, powerful mechanisms to capture, translate, enforce
- Degrees of freedom
  - wire widths/spacings, shielding/interleaving, driver/repeater sizing
  - router empowered to perform small logic resyntheses
- "Methodology"
  - carefully delineated scopes of router application
  - instance complexities remain tractable due to hierarchy and restrictions (e.g., layer assignment rules) that are part of the methodology
- Change in search mechanisms
  - iterative ripup/reroute replaced by "atomic topology synthesis utilities": construct entire topologies to satisfy constraints in arbitrary contexts
- Closer alignment with full-/automated-custom view
  - "peephole optimizations" of layout are the natural extensions of Motorola CELLERITY, IBM CM5, etc. methodologies

Session Overview

- New issues and problems arising in UDSM technology
  - catastrophic yield: critical area, antennas
  - parametric yield: density control (filling) for CMP
  - parametric yield: subwavelength lithography implications
    - optical proximity correction (OPC)
    - phase-shifting mask design (PSM)
  - signal integrity
    - crosstalk and delay uncertainty
    - DC electromigration
    - AC self-heat
    - hot electrons
- Current context: cell-based place-and-route methodology
  - placement and routing formulations, basic technologies
  - methodology contexts

P&R Methodology

- Centered on logic design
  - wire-planning methodology with block/cell global placement
  - global routing directives passed forward to chip finishing
  - constant-delay methodology may be used to guide sizing
- Centered on physical design
  - placement-driven or placement-knowledgeable logic synthesis
- Buffer between logic and layout synthesis
  - placement, timing, sizing optimization tools
- Centered on SOC, chip-level planning
  - interface synthesis between blocks
  - communications protocol, protocol implementation decisions guide logic and physical implementation