Interconnect Optimization Considering Multiple Critical Paths

Jiang Hu
Department of ECE
Texas A&M University

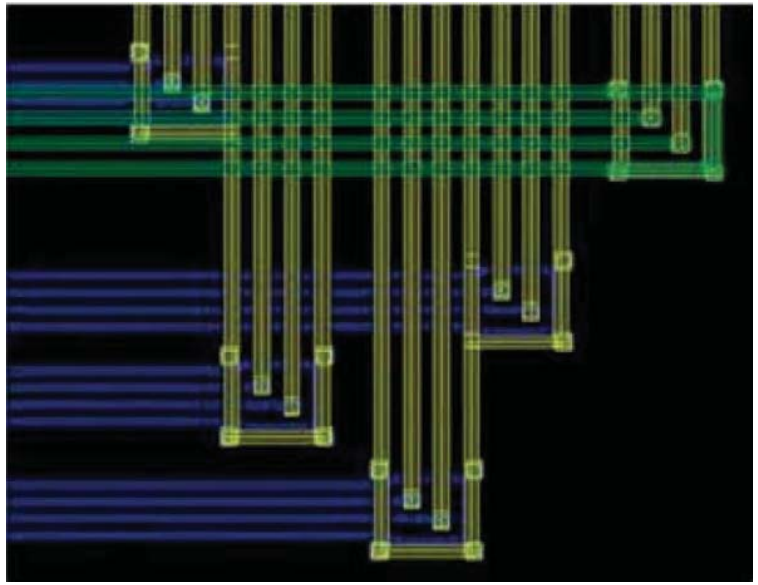
Ying Zhou, Yaoguang Wei, Steve Quay
IBM Corporation

Lakshmi Reddy, Gustavo Tellez, Gi-Joon Nam
IBM Research

Outline

- Conventional Interconnect Optimization
- Motivation Example
- P-Norm-Based Figure of Merit
- PFOM-Based Buffer Insertion
- Interaction with Timing Driven Steiner Tree
- Experimental Results
- Conclusions

Interconnect Challenge and Solutions



Performance Bottleneck

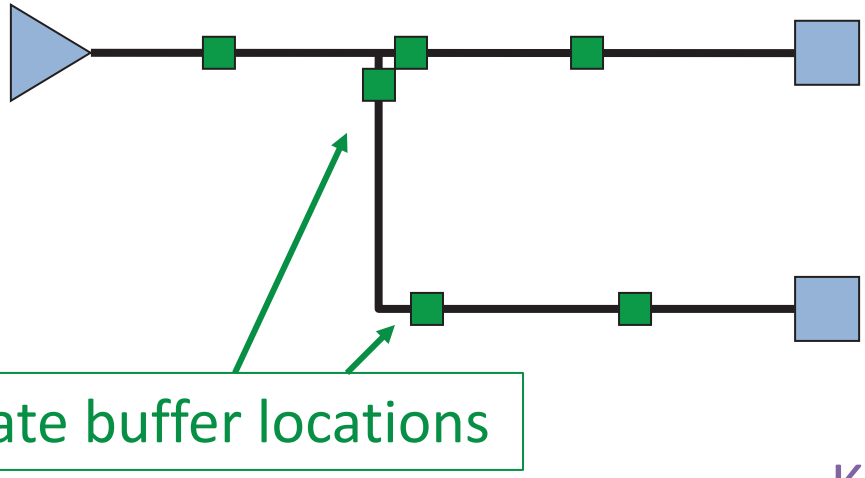
Steiner tree

Buffer insertion

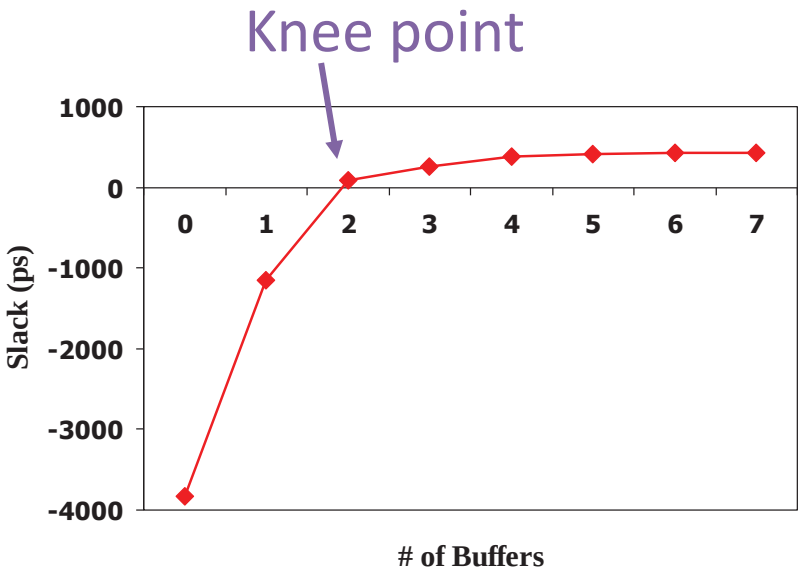
Wire sizing

Layer assignment

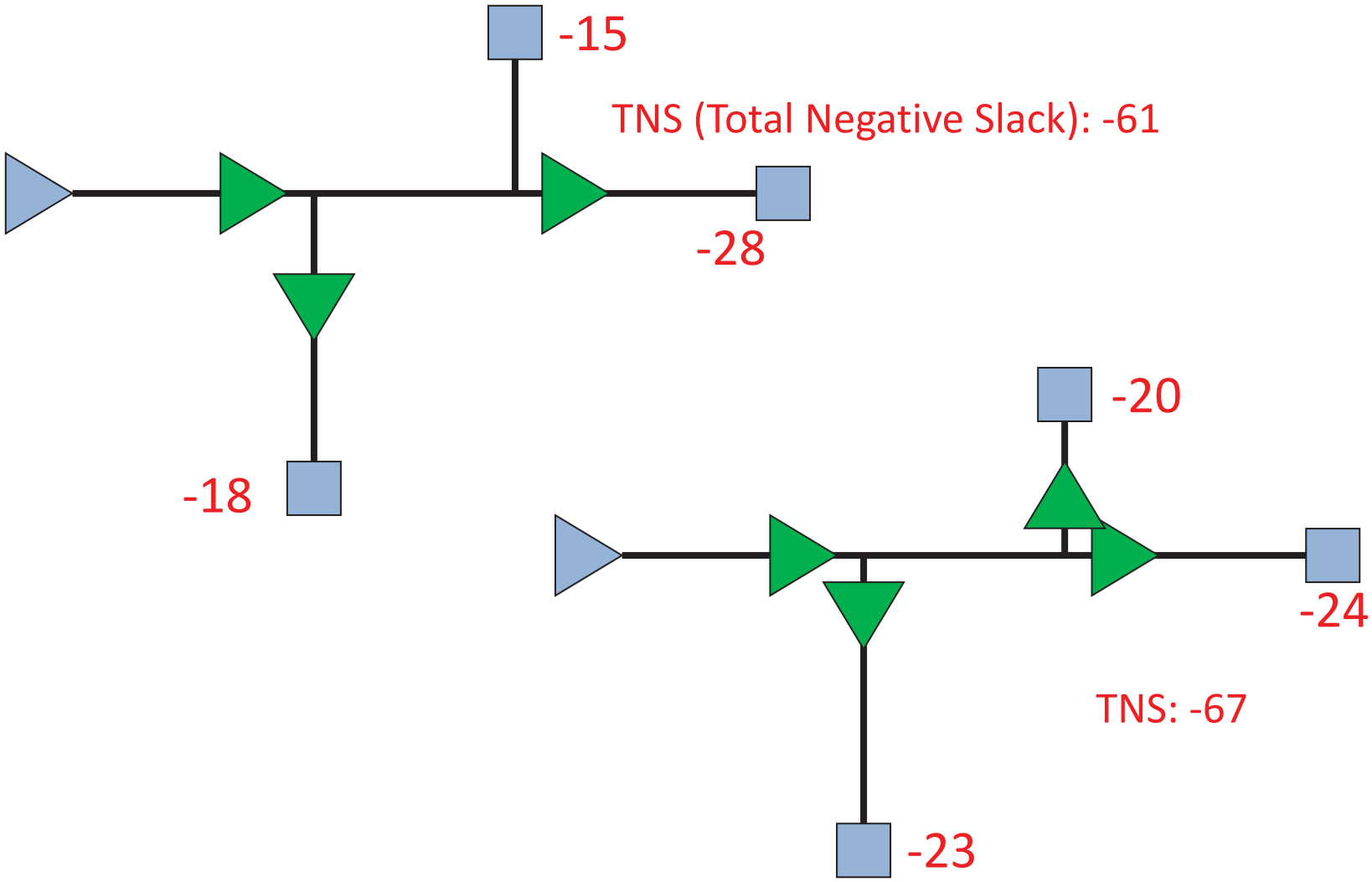
Buffer Insertion



- Worst Slack: min slack among all sinks
- Maximize slack
- Slack vs. buffer area tradeoff
- Placement-buffering-routing



Motivation Example



Does TNS Matter?

- Yes
  - Delay model in buffering is not accurate, a near critical path may actually be the most critical path
  - In later design steps, less critical paths would be easier to handle
- Worst slack vs. TNS, which is more important?

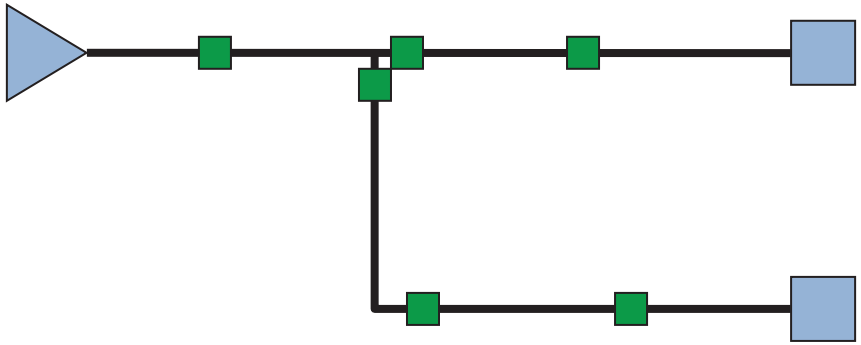
P-Norm-Based Figure Of Merit (FOM)

- $\vec{x} = (x_1, x_2, \ldots, x_m)$,
- p-norm $||\vec{x}||_p = (\sum_{i=1}^{m} |x_i|^p)^{\frac{1}{p}}$
- Slack deficit $s = \max(-slack, 0)$
- For m sinks: $pFOM = ||\vec{s}||_p = (\sum_{i=1}^{m} {s_i}^p)^{\frac{1}{p}}$

Worst Slack vs. TNS Tradeoff

- For $p = \infty$, $pFOM = \max(s_1, s_2, \ldots, s_m)$, worst slack
- For $p = 1$, $pFOM = \sum_{i=1}^{m} s_i$, TNS
- *p* indicates worst slack vs. TNS tradeoff
- pFOM-driven buffer insertion

Conventional Buffering Algorithm



- Candidate buffering solutions are generated at sinks, propagated toward the source
- New solutions are added during traversal
- Solution is characterized by *(c, q, w)*: load cap, required arrival time, buffer cost
- Inferior solutions, worse on all of *c, q, w*, are pruned
- Replace *q* by *pFOM*?

The Main Difficulty

In the past, we care only the slowest person, do not even care who the person is

Now we need to track all people who MAY miss deadline

Difficulties and Handling

- Slack prediction: upstream buffered delay

$$d(l) = \left(R_b C + R C_b + \sqrt{2 R_b C_b R C}\right) \cdot l$$

Pathlength; Buffer R, C; Wire R, C

- Storage: keep track all sinks in a subtree
- Numerical: too large p => overflow

$$\left(\sum_{i=1}^{m} {s_i}^p\right)^{\frac{1}{p}}$$

Explicitly track ∞FOM == worst slack

Solution Pruning

- Conventional pruning: efficiency from the min/max operation on timing
- Neither TNS nor pFOM has this advantage
- Fast pruning techniques, e.g., squeeze pruning, cannot provide solution guarantee any more
- pFOM buffering is quite expensive!

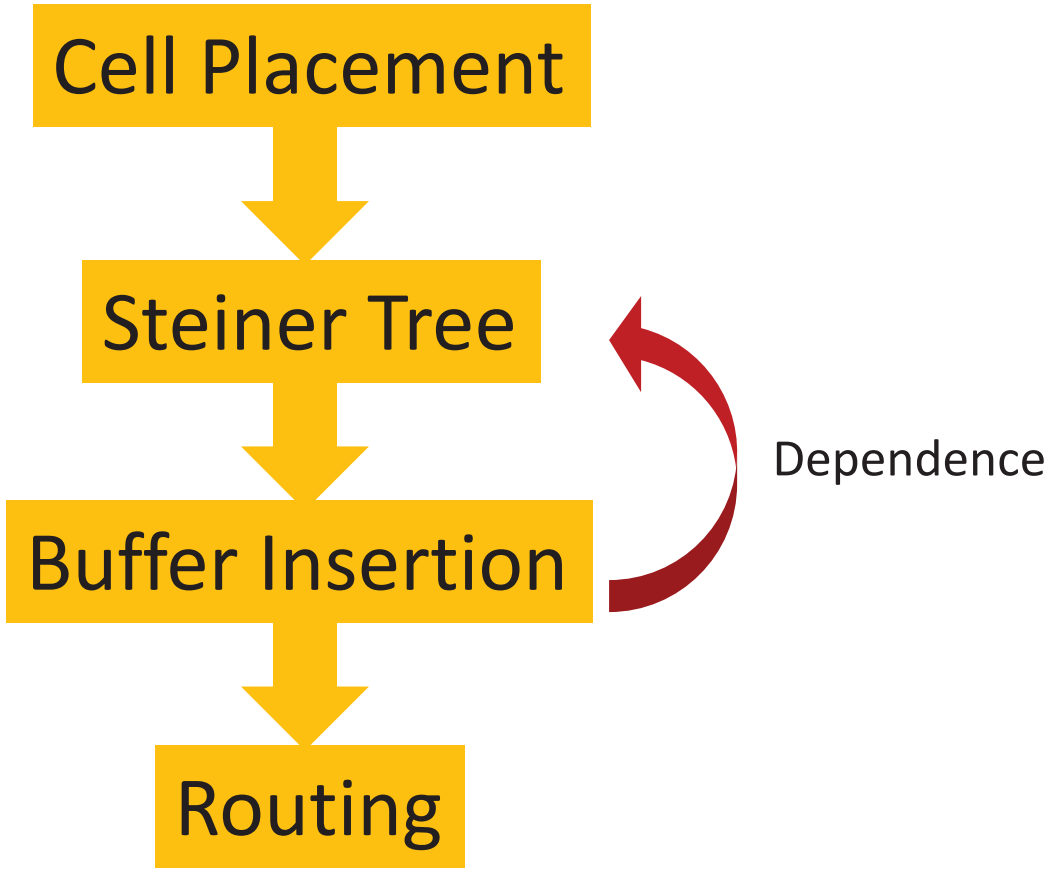
Apply to Which Nets?

- Not good to apply for all nets
- Apply for nets with relatively large number of negative sink slacks

Solution Selection at the Root

- Multiple solutions at the root
- Conventional: slack-cost tradeoff, knee point
- Now: slack-TNS-cost tradeoff
  - Slacks are tracked in bottom-up
  - Pick solution S_k, the knee of slack-cost tradeoff
  - Prune solutions with slack much worse than S_k
  - Pick the knee of pFOM-cost tradeoff

Interactions with Steiner Tree



History of Steiner Tree

- Wirelength driven: Steiner minimum tree
- Wirelength-radius tradeoff: AHHK, A-Tree
- Timing-driven: SERT, P-Tree
- Simultaneous buffering and Steiner tree construction: buffered P-Tree, buffered A-Tree
- Buffering-aware Steiner tree: C-Tree, S-Tree

Overview of C-Tree

- Cluster sinks according to
  - timing criticality
  - spatial proximity
  - signal polarity
- AHHK tree for each cluster, and connecting clusters
  - For each subtree, try 5 AHHK options with different wirelength-radius tradeoff, pick the one with the best timing

Extensions

- Additional option
  - SERT: greedy timing-driven Steiner tree
  - B-Tree: Bartoschek, et al, ISPD'06, Bonn Univ, greedy timing-driven Steiner tree with pre-sorted order
- Timing evaluation of tree options
  - Wire delay: predicted buffered delay
  - Driver delay: $R_{driver}\sqrt{C_{buf}C_{tree}}$

Experiment Setup

- 1112 nets from 14nm industrial design, single-sink nets not included
- Baseline: worst slack buffering + C-Tree
- Buffering
  - pFOM buffering
  - P-Lite: bottom-up the same as worst slack buffering + TNS solution, pick pFOM solution at the root
- Steiner tree: C-Tree, SERT, B-Tree

905 Small Nets (2-5 Sinks)

Buf	Steiner	ΔTNS (%)		ΔWS (ps)		$\Delta BufArea$ (%)	
		Ave	Max	Ave	Max	Ave	Max
Lite	CTree	0.67	54	0.06	9.5	0.91	69
	SERT	0.62	54	0.06	9.5	0.96	80
	BTree	0.60	54	0.05	9.5	0.97	80
p1	CTree	0.32	62	-0.04	40.2	0.35	50
	SERT	0.27	62	-0.06	40.2	0.37	50
	BTree	0.28	62	-0.06	40.2	0.44	50
p2	CTree	0.40	62	-0.01	40.2	0.05	75
	SERT	0.38	62	-0.03	40.2	0.16	75
	BTree	0.39	62	-0.03	40.2	0.20	75
p4	CTree	0.63	65	0.05	40.2	0.00	50
	SERT	0.56	65	0.04	40.2	0.09	50
	BTree	0.52	65	0.03	40.2	0.08	50
p6	CTree	0.63	65	0.07	40.2	-0.07	50
	SERT	0.66	65	0.06	40.2	0.07	50
	BTree	0.64	65	0.06	40.2	0.09	50
p8	CTree	0.73	65	0.06	40.2	-0.08	54
	SERT	0.68	65	0.04	40.2	-0.01	54
	BTree	0.66	65	0.04	40.2	0.01	54
p10	CTree	0.75	62	0.09	40.2	-0.03	50
	SERT	0.69	62	0.07	40.2	0.05	50
	BTree	0.67	62	0.07	40.2	0.07	50
p12	CTree	0.62	62	0.04	40.2	-0.07	50
	SERT	0.58	62	0.04	40.2	-0.03	50
	BTree	0.56	62	0.03	40.2	-0.01	50

$p1$: pFOM buffering with *p=1*

ΔTNS: TNS reduction

ΔWS: Worst Slack reduction

$\Delta BufArea$: Area increase

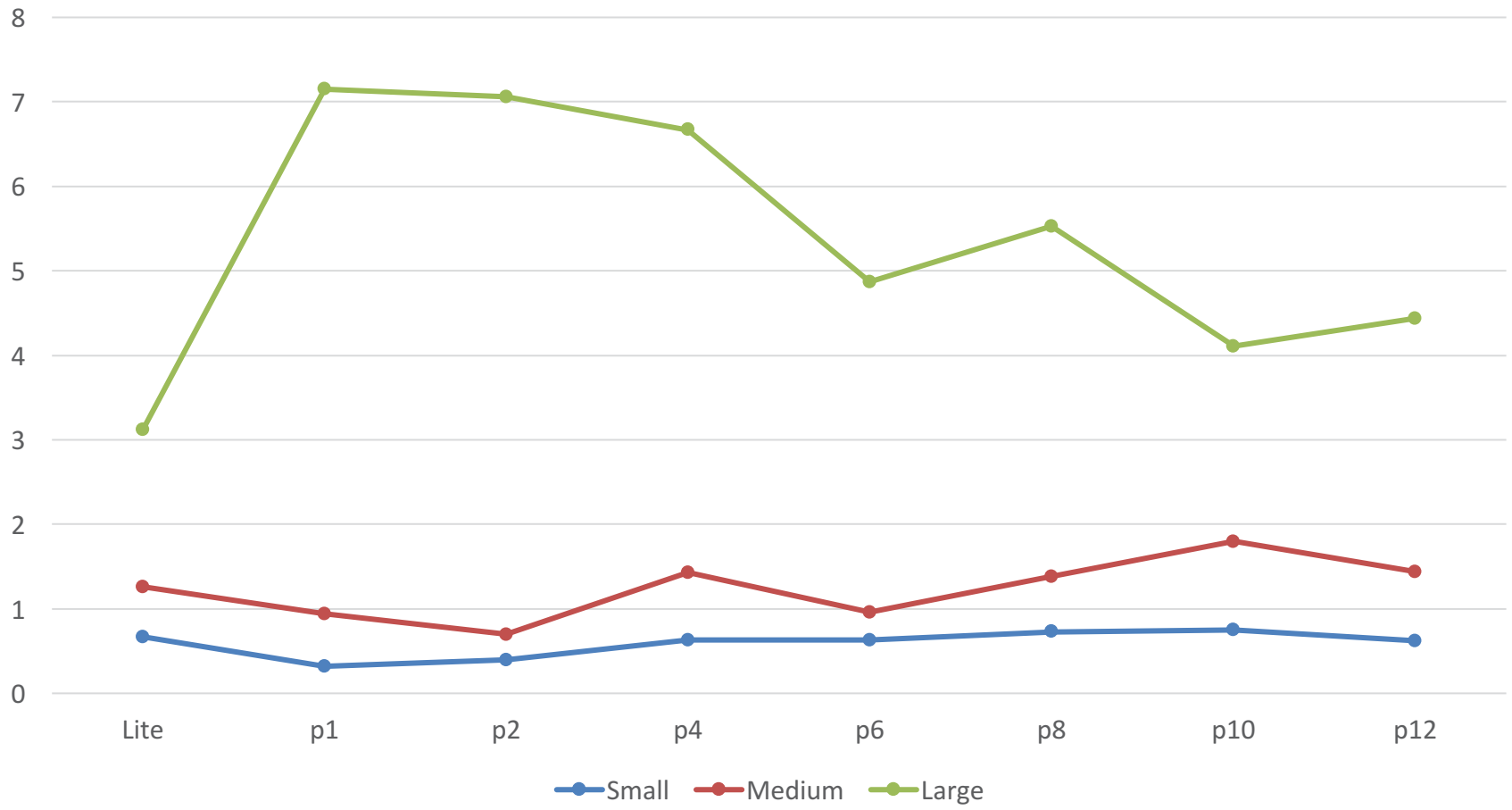
147 Medium Nets (6-15 Sinks)

Buf	Steiner	ΔTNS (%)		ΔWS (ps)		$\Delta BufArea$ (%)	
		Ave	Max	Ave	Max	Ave	Max
Lite	CTree	1.26	27	0.41	18.0	2.61	44
	SERT	0.55	52	0.93	32.6	3.40	86
	BTree	1.66	46	0.76	31.6	3.55	83
p1	CTree	0.94	32	-0.30	18.3	1.47	54
	SERT	2.16	53	0.26	32.6	3.11	59
	BTree	2.60	52	0.11	29.0	1.97	62
p2	CTree	0.70	32	-0.26	18.3	0.72	54
	SERT	1.68	53	0.31	32.6	2.26	59
	BTree	2.04	52	0.16	31.8	1.18	80
p4	CTree	1.43	32	-0.05	18.7	0.99	54
	SERT	2.08	52	0.48	32.6	1.84	59
	BTree	2.80	52	0.39	31.8	1.53	78
p6	CTree	0.96	33	0.02	18.7	1.57	54
	SERT	1.64	52	0.40	32.6	1.49	56
	BTree	2.43	52	0.34	31.7	1.07	93
p8	CTree	1.38	44	0.20	19.3	1.68	47
	SERT	2.10	52	0.39	32.6	1.00	56
	BTree	2.60	52	0.37	32.5	0.32	62
p10	CTree	1.80	44	0.35	19.3	2.04	65
	SERT	2.33	52	0.53	32.6	1.52	65
	BTree	2.88	52	0.56	32.5	1.16	65
p12	CTree	1.44	44	0.24	19.3	1.45	47
	SERT	1.86	52	0.55	32.6	0.91	56
	BTree	2.45	52	0.62	32.5	0.97	62

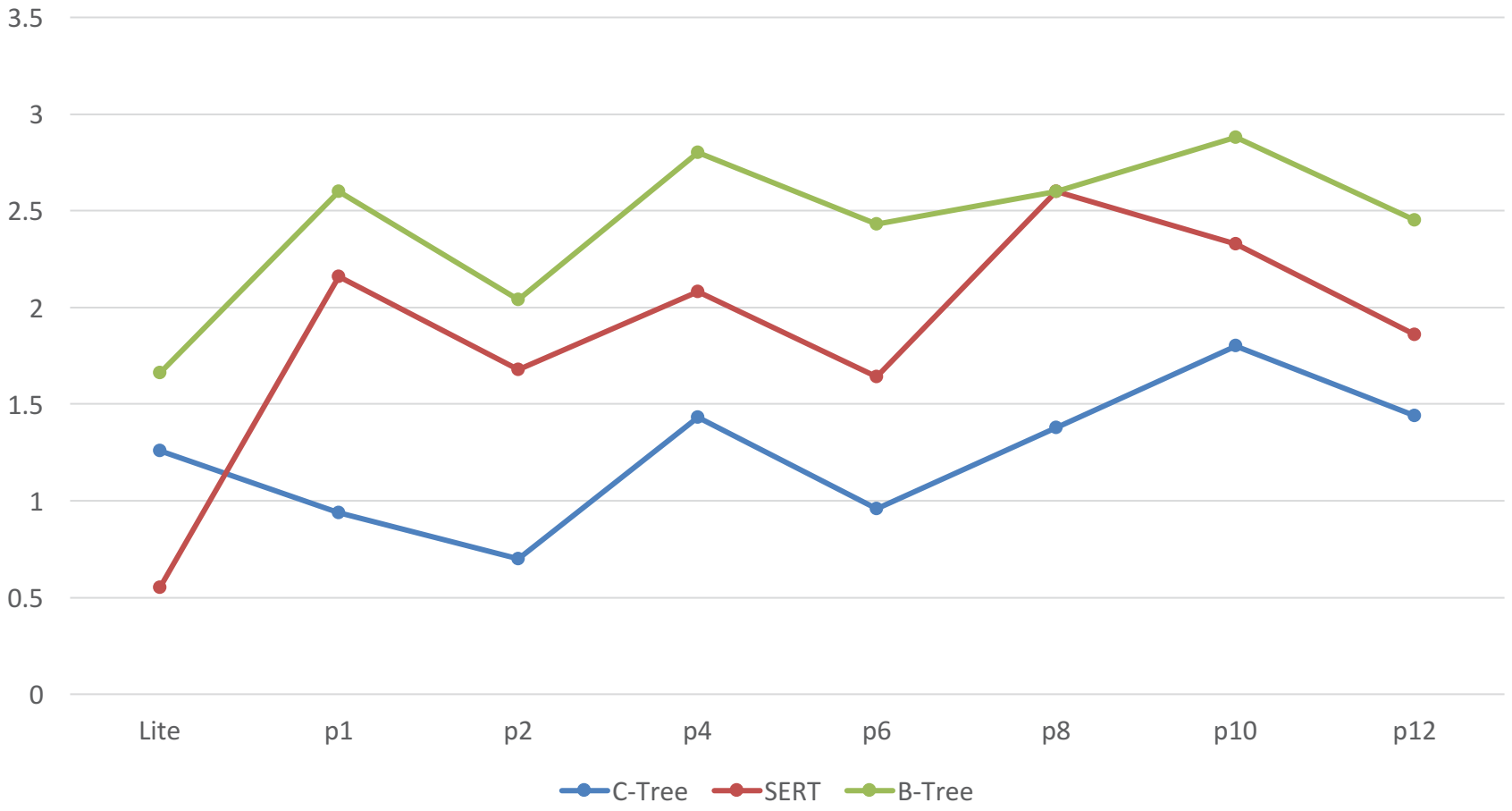
60 Large Nets (> 15 Sinks)

Buf	Steiner	ΔTNS (%)		ΔWS (ps)		$\Delta BufArea$ (%)	
		Avg	Max	Avg	Max	Avg	Max
Lite	CTree	3.12	69	1.47	32.8	8.22	88
	SERT	7.20	65	-2.07	60.4	7.53	98
	BTree	3.87	65	2.50	60.4	9.24	71
p1	CTree	7.15	65	0.50	42.8	5.58	85
	SERT	6.62	65	-5.37	35.5	4.55	57
	BTree	4.63	65	0.40	35.5	11.69	96
p2	CTree	7.06	65	0.38	39.5	5.16	82
	SERT	6.96	65	-4.92	35.5	2.76	57
	BTree	5.60	65	0.78	35.5	11.70	96
p4	CTree	6.67	65	1.03	39.5	4.18	57
	SERT	7.02	65	-4.46	40.5	3.60	57
	BTree	5.10	65	1.01	40.5	10.44	97
p6	CTree	4.87	70	0.50	39.5	5.26	59
	SERT	6.98	65	-4.06	40.5	4.39	92
	BTree	4.38	65	1.33	40.5	7.79	93
p8	CTree	5.53	70	0.90	39.5	6.71	66
	SERT	5.98	65	-4.18	40.5	3.32	57
	BTree	4.29	65	1.28	40.5	7.87	93
p10	CTree	4.11	65	0.74	39.5	3.57	70
	SERT	6.32	65	-4.08	41.8	4.06	99
	BTree	4.02	65	1.22	41.8	6.10	93
p12	CTree	4.44	65	0.84	39.5	3.98	70
	SERT	5.59	65	-4.16	41.8	4.09	99
	BTree	3.64	65	1.22	41.8	5.75	81

TNS Reduction vs pFOM



TNS Reduction of Different Steiner Trees on Medium Nets



A 64-Sink Net

Steiner	Buf	$TNS(ps)$	$WS(ps)$	#Buffers	Buf Area	Runtime (s)
CTree	WS	-6610	-140	8	63	0.2
	Lite	-6444	-141	10	73	0.3
	p1	-6446	-141	11	77	1.1
	p4	-6465	-134	15	97	2.8
	p8	-6291	-134	20	121	2.8
	p16	-6311	-134	17	103	2.8
SERT	WS	-6152	-139	12	102	0.2
	Lite	-6152	-139	12	102	0.4
	p1	-6217	-141	10	94	2.7
	p4	-6059	-126	23	154	5.5
	p8	-6265	-142	8	86	4.7
	p16	-5764	-103	30	163	3.9
BTree	WS	-6083	-117	27	128	0.2
	Lite	-6260	-143	15	95	0.3
	p1	-6009	-121	22	108	5.7
	p4	-6280	-143	16	101	7.7
	p8	-5950	-113	35	168	5.4
	p16	-5899	-110	32	154	4.3

Overall Flow Results

Circuit	$\#nets$	ΔTNS	ΔWS	$\Delta Area$	$\Delta Power$
1	0.3M	23.6%	-8.1%	0.01%	-0.97%
2	0.6M	6.1%	2.4%	-0.17%	-0.16%
3	0.5M	6.7%	-0.7%	0.42%	-0.19%
4	0.5M	8.1%	6.2%	0.06%	0.27%
5	0.6M	8.6%	5.3%	0.16%	-0.26%
6	1M	4.4%	0.2%	0.15%	0.02%
Ave		10%	0.9%	0.11%	-0.22%

Conclusions

- pFOM buffering is proposed for optimizing total negative slack besides the worst slack
- Buffering aware Steiner tree is enhanced
- Industrial design results show significant timing benefit
- The slack-TNS tradeoff control needs to be improved

Thank You!