Investigation of Performance Metrics for Interconnect Stack Architectures

Puneet Gupta[1] , Andrew B. Kahng[1] ,

Youngmin Kim[2], Dennis Sylvester[2]

[1]ECE Department, University of California at San Diego

[2]EECS Department, University of Michigan at Ann Arbor

Outline

- Motivation
- Delay and Bandwidth
  - Via Blockage
  - WLD and Wire Assignment (Avg. wire length for each layer)
  - Bandwidth Metrics
- Energy-Driven Metrics

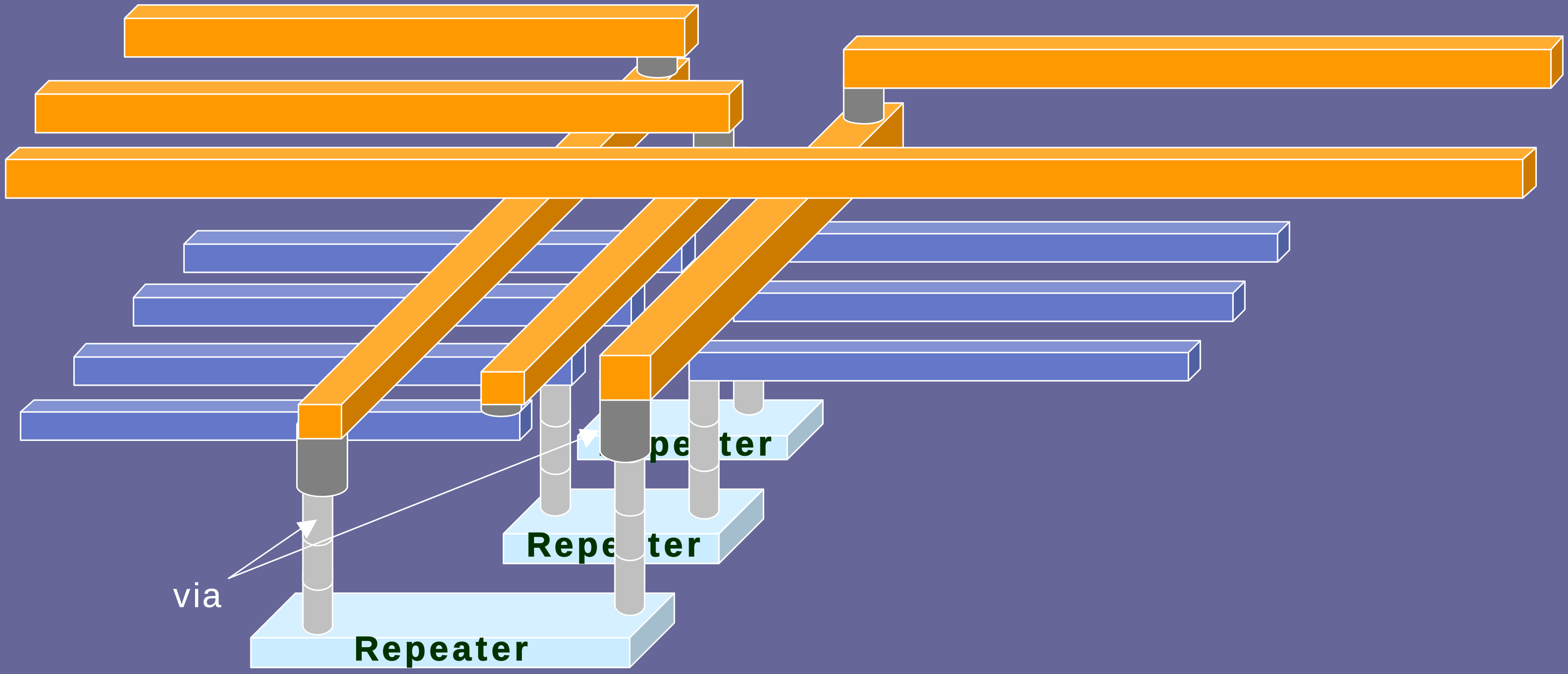
Motivation

- Front-end dimensions set by lithography restrictions
- Back-end dimensions are often area/performance driven
  - Especially intermediate and global metal levels
- Front-end performance quantified with known metrics:
  - FO4 or ring oscillator delays
  - Ioff, Ion
- No comparable metrics for back-end
  - RC per µm ignores many issues
- Via blockage factor is important to consider

Delay and Bandwidth

- Bandwidth or throughput-driven approach for the interconnect recently proposed
  - Ho 01, Young 01, Lin 02
- All approaches are applied to one layer only (i.e., a top level)
- Ignore factors due to multilayer interconnect
  - via blockage
  - repeater insertion
- Bandwidth and energy metrics *considering the entire multilayer interconnect stack* are required

Interconnect Stack



Source: Muddu

- *We seek to develop metrics that can be used to compare back-end dimensions, and eventually to drive the selection of such dimensions*

Standard Interconnect Delay Models

- Worst-case 50% delay of a minimum sized inverter driving an interconnect is (Pamunuwa 03)

$$t_{0.5} = 0.7R_{drv}(C_g + 4.4C_c + C_{drv}) + R_w(0.4C_g + 1.5C_c + 0.7C_{drv})$$

- The number of repeaters is k and the size of each repeater is h, then

$$t_{0.5} = k\left[0.7\frac{R_{drv}}{h}\left(\frac{C_g}{k} + 4.4\frac{C_c}{k} + hC_{drv}\right) + \frac{R_w}{k}\left(0.4\frac{C_g}{k} + 1.5\frac{C_c}{k} + 0.7hC_{drv}\right)\right]$$

Via Blockage; Previous Work

- Via blockage factor (v_i) :
  - Fraction of total available space that is not available due to via blockage effects on layer *i*
- A layer blocks 12% ~ 15% of the wiring capacity of every layer underneath it at constant pitch (Sai-Halasz)
- Via blockage is only severe on the lower metal layers (Chen et al.)

$$N_{v\_wire} = 2\left[I(L_{\max}) - I(L_n)\right] \qquad when \ \ n \neq 0$$

$$N_{v\_wire} = 2I(L_{\max}) - I(L_n) \qquad when \ \ n = 0$$

Where N_{v_wire} is the number of vias by wires, *I(l)* is cumulative interconnect density

Via Blockage, continued

- Repeater insertion for long wires in the semi-global and global interconnect layers is necessary for delay and slew constraints

$$N_{v\_rep} = 2 \sum_{i=n+1}^{top\_layer} \#\, of \;\; repeaters \;\; in \; i^{th} \;\; layer$$

- Total via blockage factor in nth layer

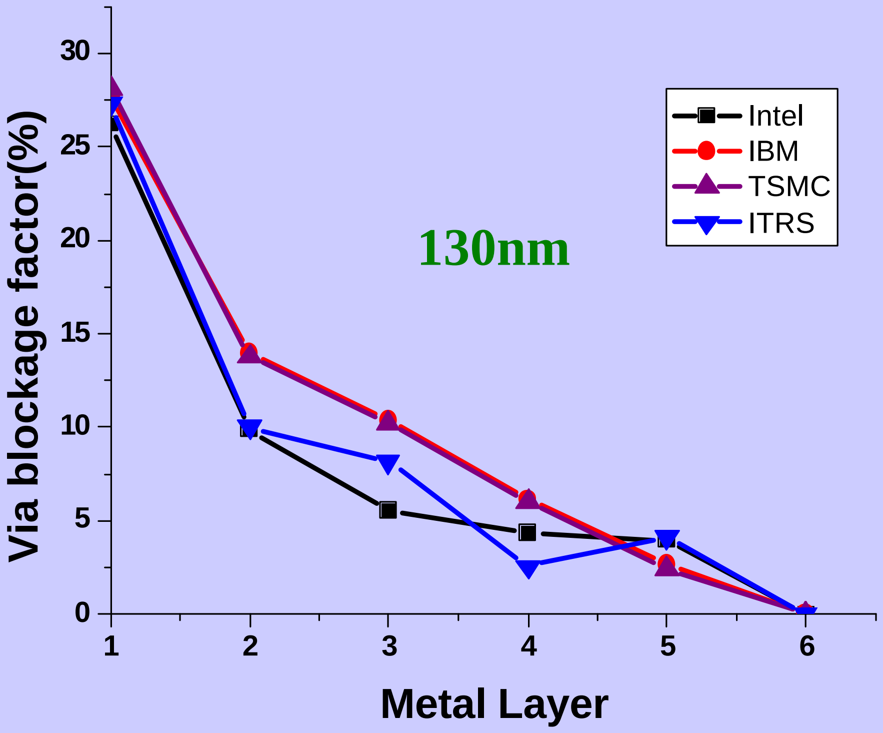
$$B_v(n) = \sqrt{N_{v\_total}\,(2W + s\,l\,)^2 / A_c}$$

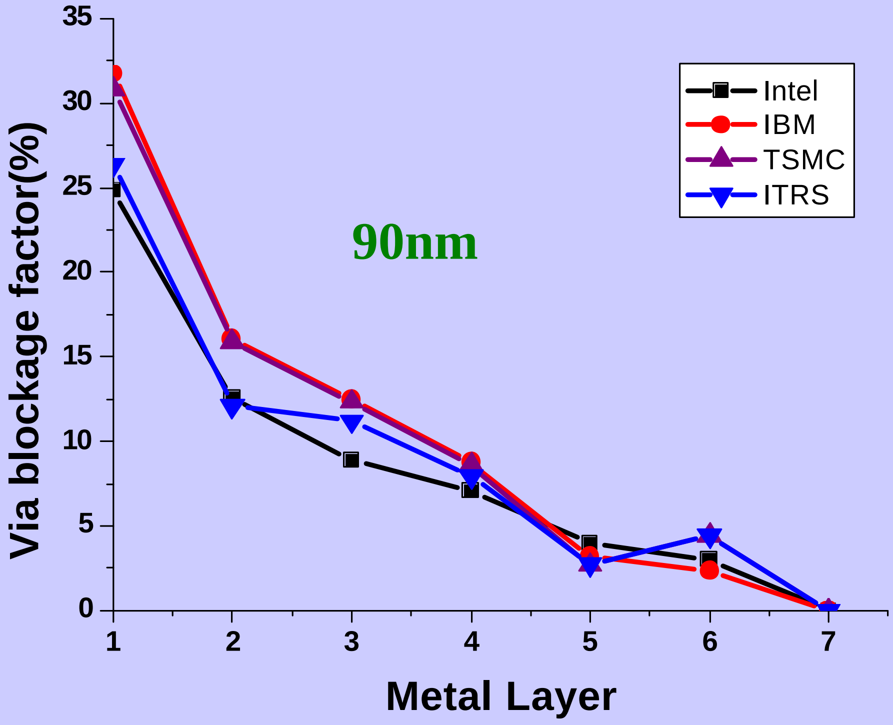
Interconnect Technology Parameters

Parameters	*130nm*	*90nm*
Ac (cm^2)	0.98	0.98
#of gates	6.4M	12.87M
Vdd (V)	1.4	1.2
Id (mA/um)	1.2	1.0
# of nets	29.08M	58.15M
k (ILD)	3.6	2.9
p (Rents exponent)	0.6	0.6

Layer	*130nm technology (nm)*				*90nm technology (nm)*			
	Intel	IBM	TSMC	ITRS	Intel	IBM	TSMC	ITRS
1	350	320	340	350	220	245	240	210
2	448	400	410	350	320	280	280	210
3	448	400	410	450	320	280	280	275
4	756	400	410	450	400	280	280	275
5	1120	400	410	1340	480	280	280	275
6	1204	800	900	1340	720	560	840	820
7	-	-	-	-	1080	1120	840	820

Via Blockage Projections





- 2~5x times larger blockage at the bottom layer
- IBM and TSMC show more significant via blockage → less tapered nature (many iso-pitch layers)

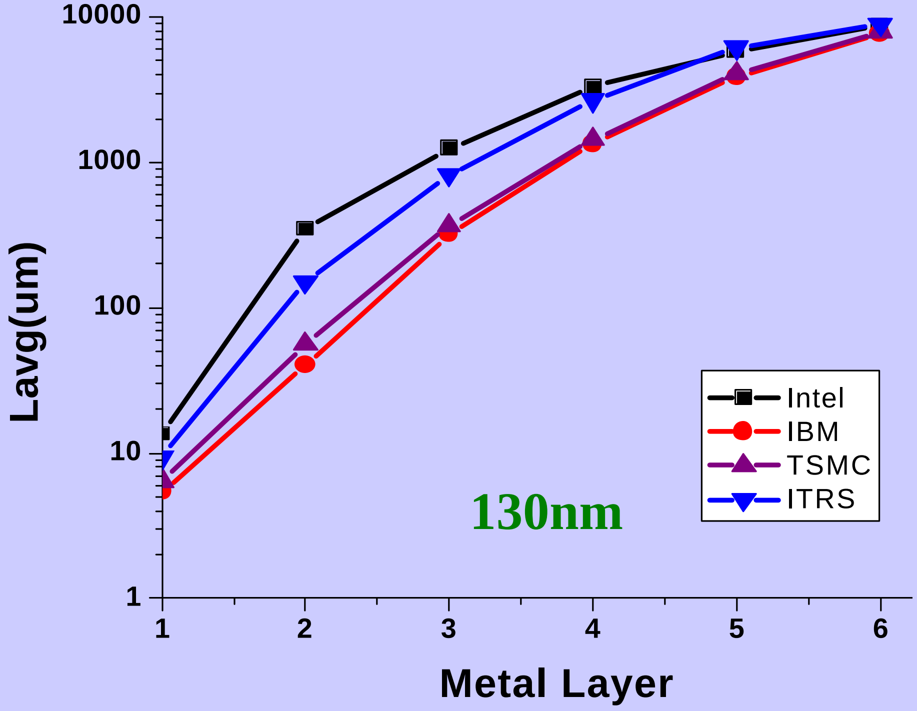
WLD and Wire Assignment

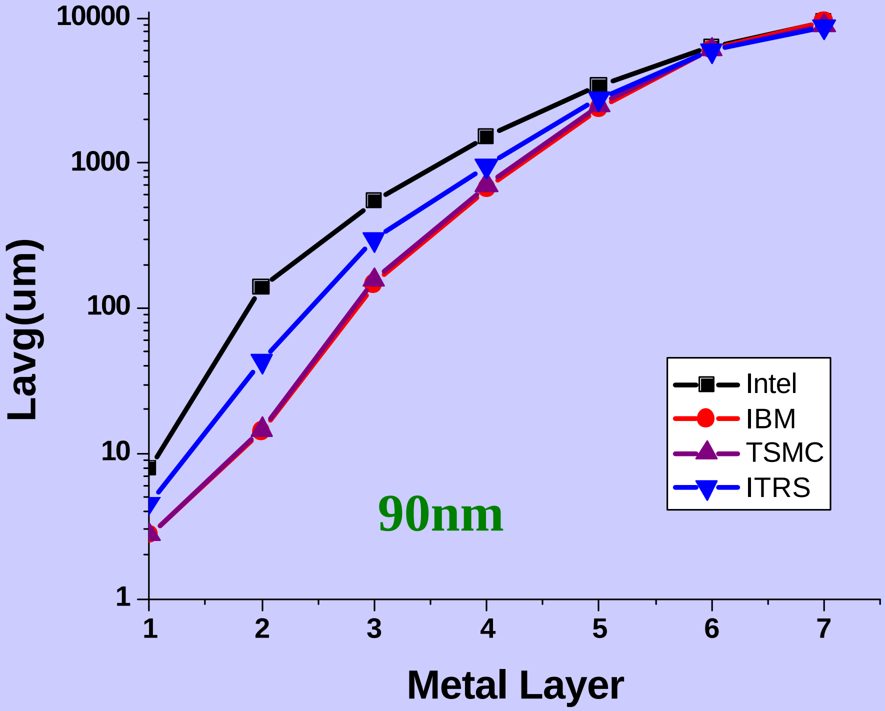
- Typical wire lengths on a given layer can be considerably different → delay as well
- Davis's wirelength model and a top town wire assignment technique (Venkatesan 01) are applied

$$A_{av} = e_w A_c = c\ P_n \sqrt{\frac{A_c}{N_g}} \int_{L_{n-1}}^{L_n} l \cdot i(l)\, dl = A_{req}$$

- ***Average wire length*** on each layer is used as ***typical*** wire length on the layer

Average Wire Length





- Intel and ITRS show much larger average wire length for all layer
  - top down wire assignment
  - Wider pitch at top two layers for Intel/ITRS

Bandwidth Metrics

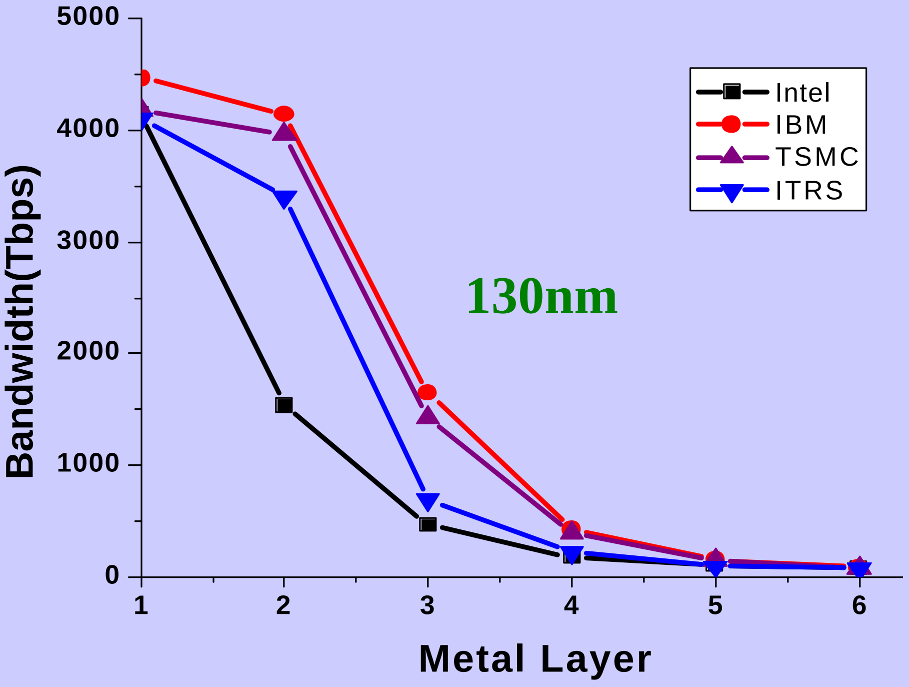
- Represents the rate at which information can be transferred through channel
  - # of wires or bits per second
- Delay_n is the wire delay calculated at the average wire length in a given layer *n*
  - Driver is sized to match wire cap. using optimal repeater as baseline
- N_{wire} is the number of parallel wires

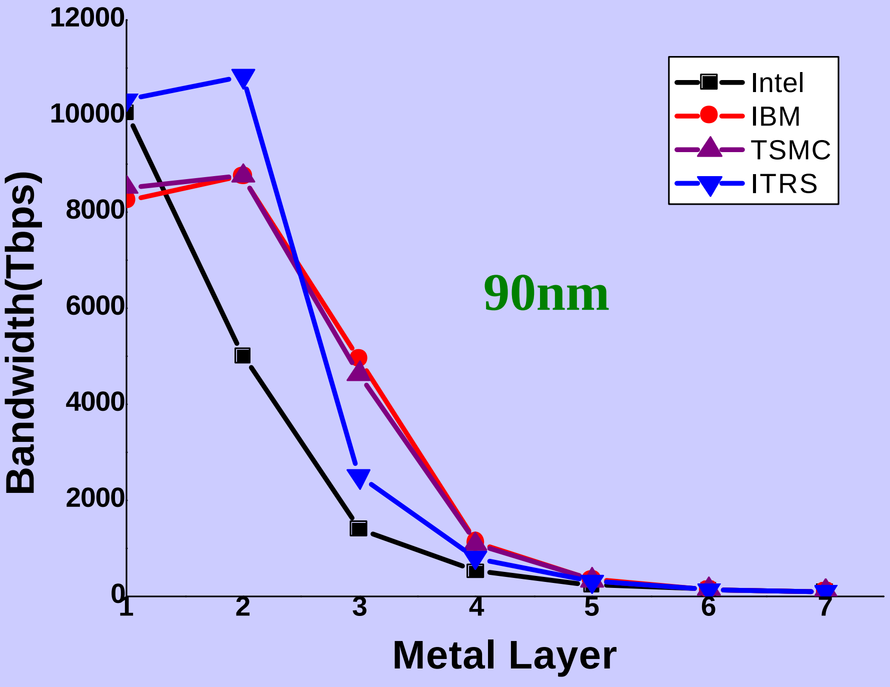
$$BW_n = \left(\frac{1}{Delay_n}\right) \times N_{wire} = \frac{1}{Delay_n} \times \left(\frac{chip\ \ side\ \ length}{pitch_n}\right)$$

- If L_{avg} > maximum allowable distance between repeaters → insert optimal repeaters
- Via blockage reduces the wiring resources directly

$$BW_n = \left(\frac{1}{Delay_n}\right) \times \left(\frac{chip\ side\ length}{pitch_n}\right) \times B_v(n)$$

Bandwidth





- IBM and TSMC show better bandwidth
  - Shorter average wire length and greater wiring density overcome larger RC per unit length
- Bandwidth in lower layers is much higher than in upper layers
  - Shorter wires and greater wiring density

Normalized Bandwidth

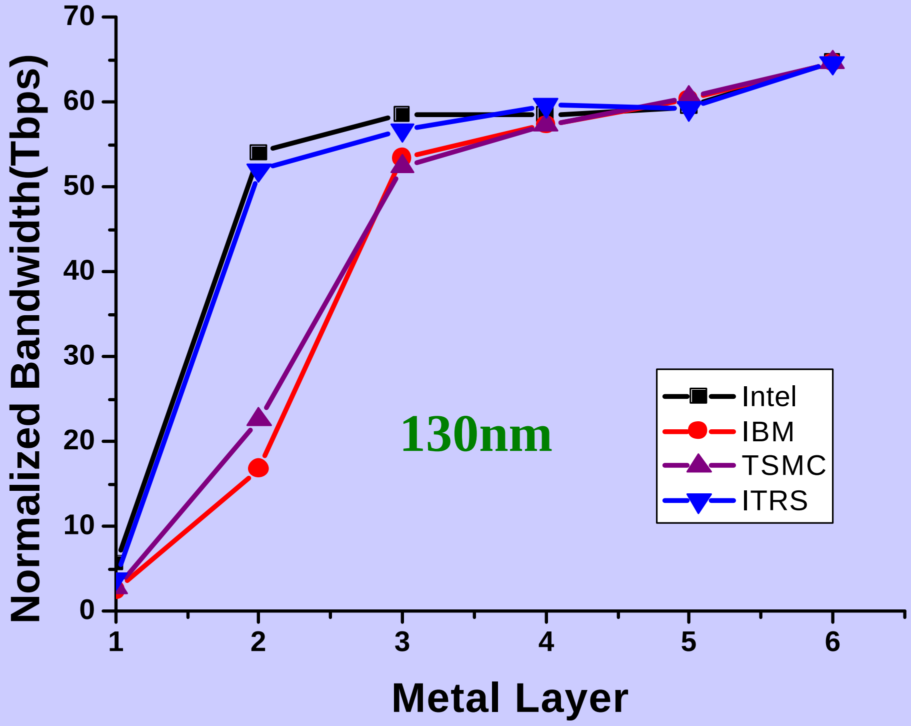
- Simply summing BW of individual layers is not a good way to assess the performance → normalization required
- The number of segments:
  - *"routing demand"* for a given layer

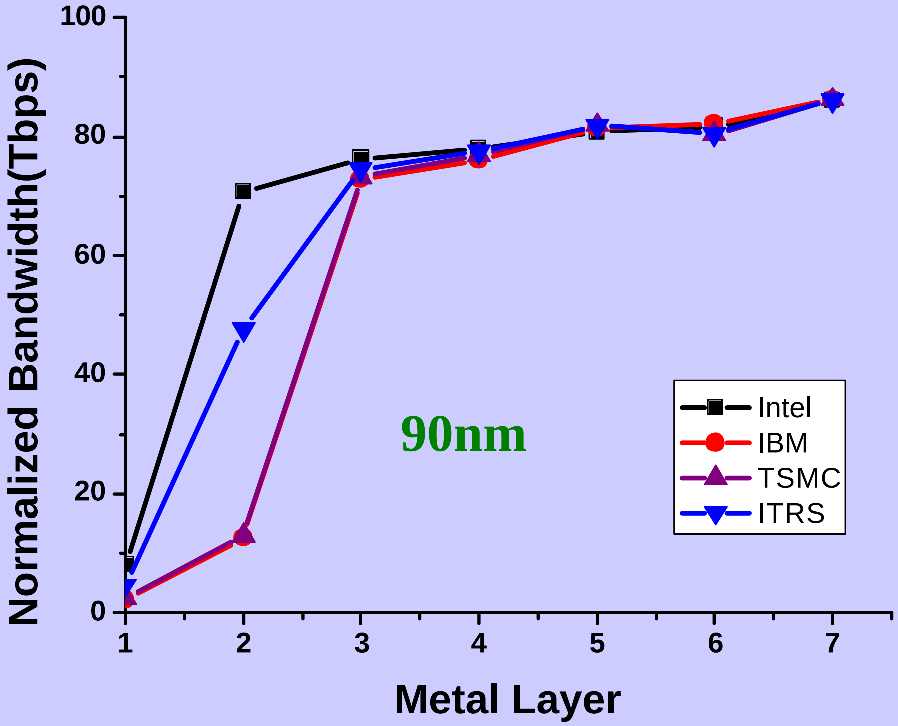
$$N_{seg} = \frac{\sqrt{A_c}}{L_{avg}}$$

→

$$Normalized \quad BW_n = \frac{BW_n}{N_{seg}}$$

Normalized Bandwidth

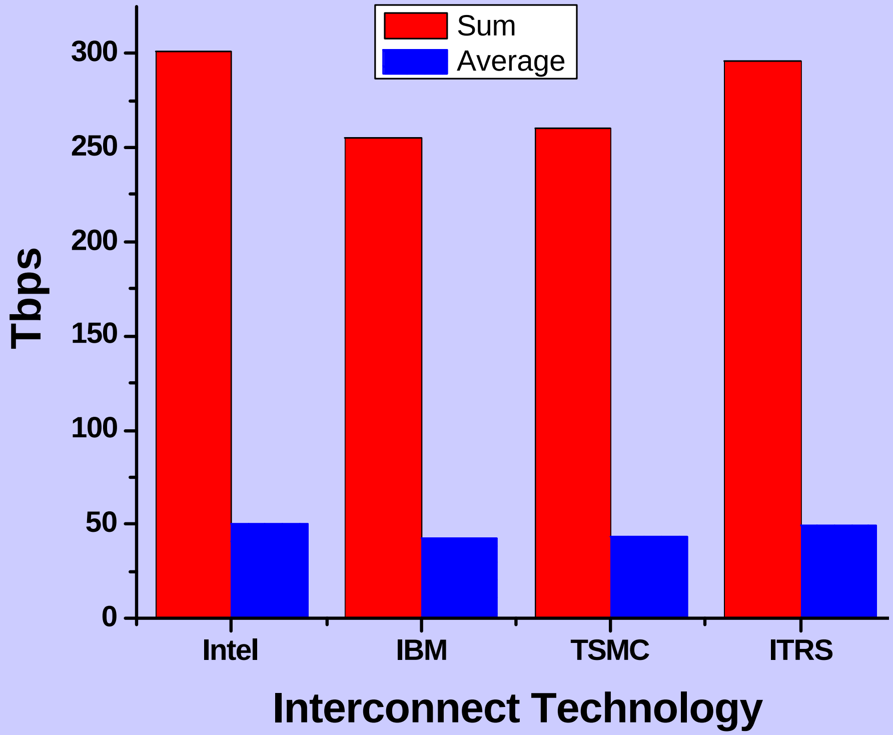




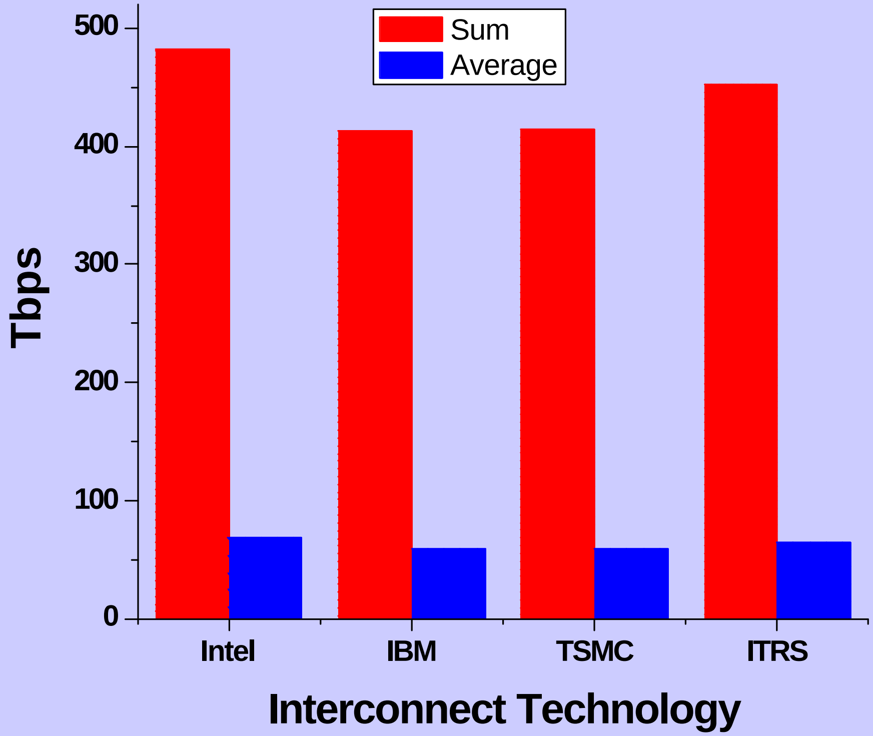
- Intel and ITRS stacks are superior when considering normalization
  - They have fewer segments due to longer average wire length on all layers
  - Their pre-normalized BW was already penalized for longer wirelengths
- Normalized bandwidth is more consistent across the stack (ignoring metal 1 where density is main criteria)

Statistics of Normalized BW

- ~15% variation in total normalized BW across technologies
  - Somewhat smaller than spread of front-end metrics like FO4 and Ion



130nm

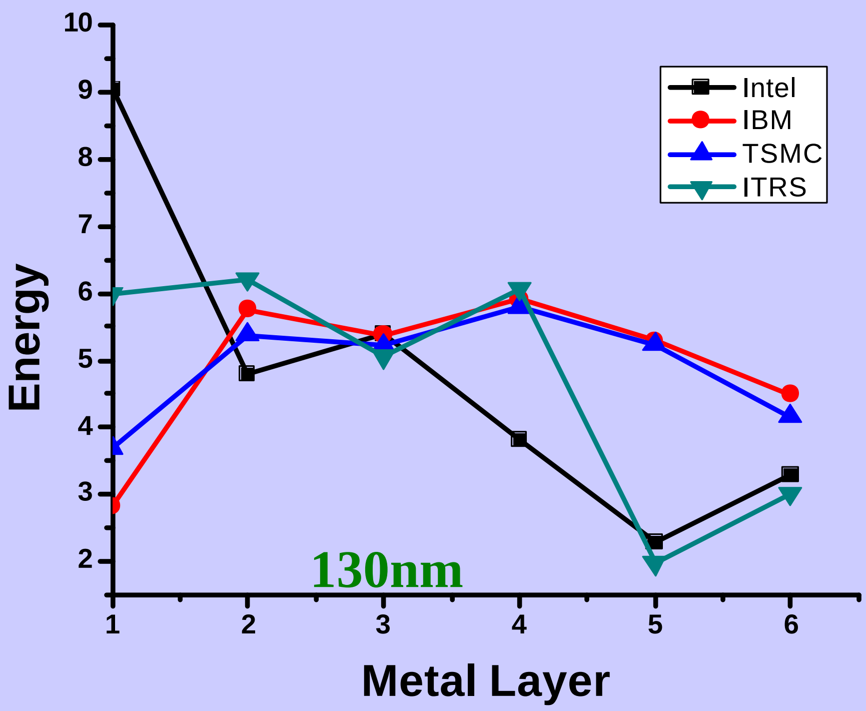


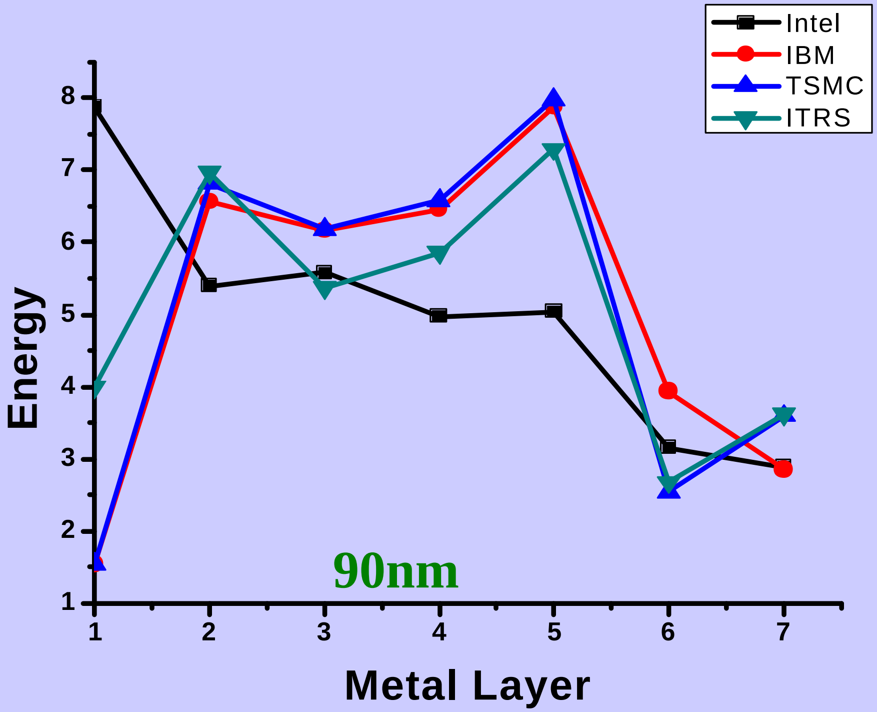
90nm

Energy-Driven Metrics

- The increasing use of repeaters on global and even intermediate layers
  - Reduces delay and maintains good signal integrity
  - But: Increases power consumption dramatically
- Repeater capacitance : $C_{rep} = khC_{drv}$
  - Drivers other than repeaters also considered
- Wire capacitance : $C_{wire} = 2(C_g + C_c)$
- Energy is then $= N_{wire}(C_{rep} + C_{wire})$
  - Ignoring operation frequency and supply voltage which are not varying

Energy in 130nm, 90nm

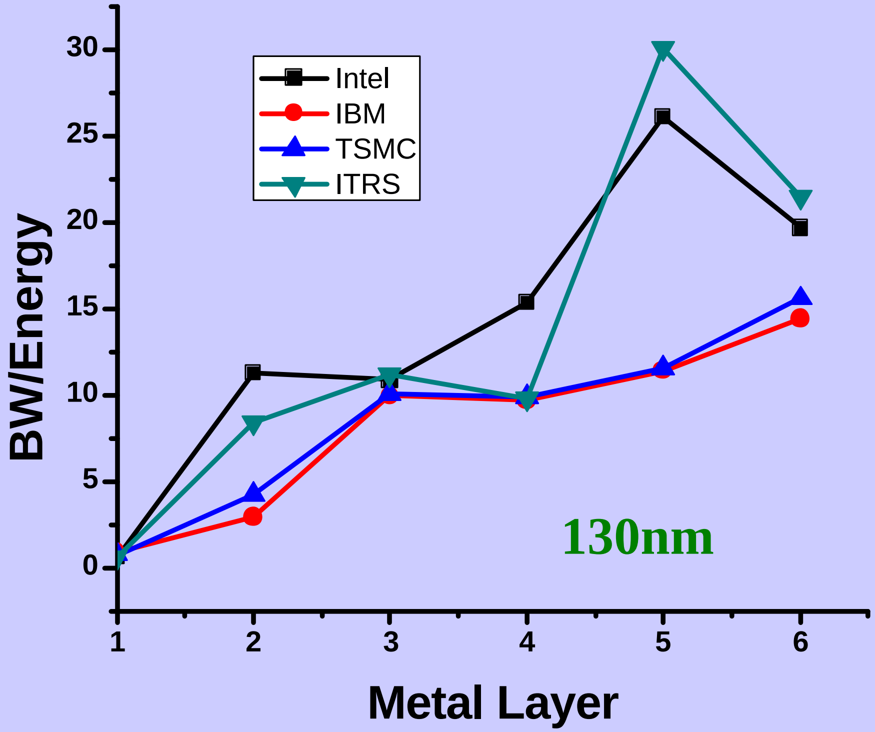


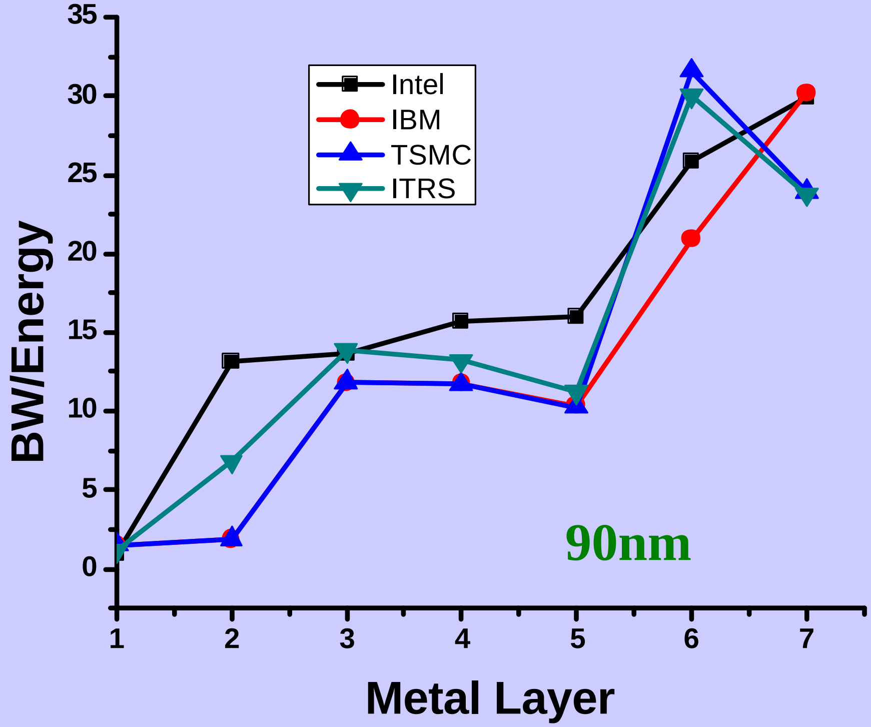


- Total energy on a layer is fairly constant across metal levels
  - Large differences on layer 1 are due to top-down layer assignment, different utilization factors
- Larger pitches at top levels allows for smaller energy consumption (fewer repeaters) in Intel and ITRS
  - With growing # of repeaters in future technologies (Saxena, ISPD03), it becomes critical to choose wiring pitches (*reverse scale*) with energy/repeaters in mind

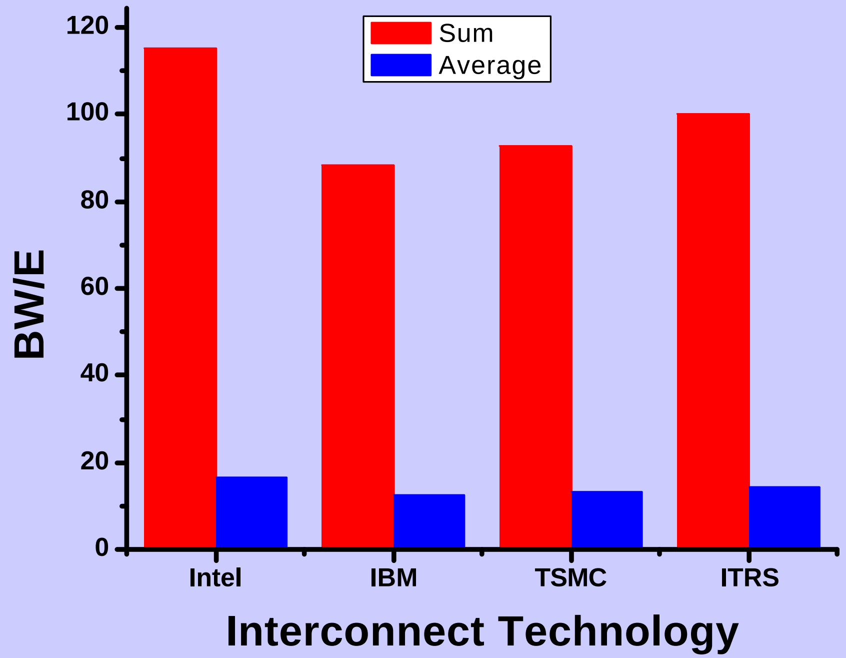
Bandwidth per unit Energy

- Bandwidth and energy can be combined to provide a complete interconnect performance metric → Bandwidth (normalized) per unit Energy

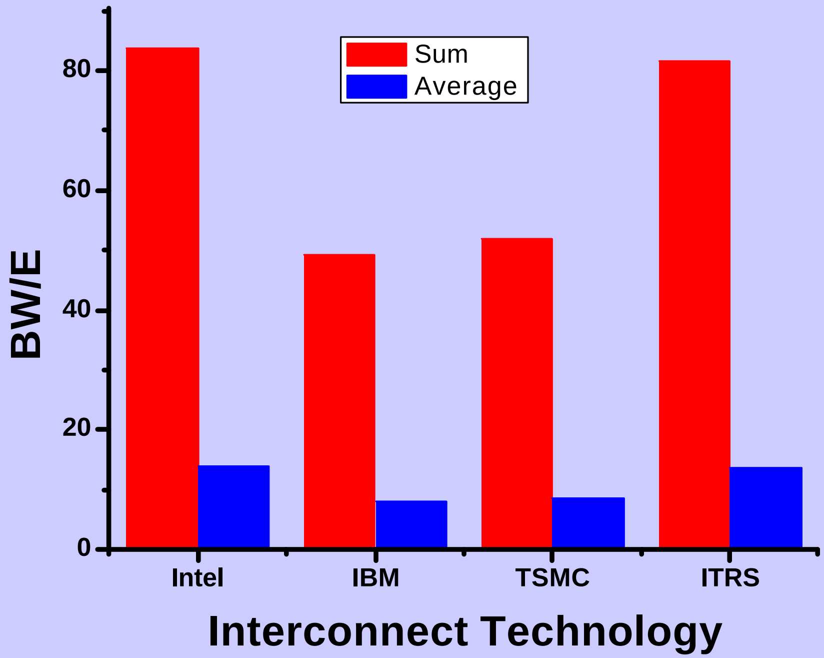




Sum and Avg. of BW/Energy



130nm



90nm

- Intel/ITRS remain appealing in terms of BW/Energy
  - Spread is now larger than in case of just BW
  - Gap increases from 130 to 90nm

Conclusions

- Bandwidth and energy metrics for complete interconnect stacks identified
- Growing impact of repeaters on via blockage
- Normalized bandwidth metrics for comparison of bandwidth across layers
- Intel and ITRS tend to show better results in terms of normalized bandwidth
  - Wider pitches, question of routability (?)
- Energy-based metrics indicate that top-level pitch choice has a large impact on BW/Energy
  - As repeaters become common on intermediate metallization layers, more layers must consider reverse scaling
- A gradually tapered interconnect stack provides best performance but somewhat more manufacturing complexity

Thank you