Performance Prediction of Throughput-Centric Pipelined Global Interconnects with Voltage Scaling

Yulei Zhang[1], James F. Buckwalter[1], and Chung-Kuan Cheng[2]
[1]Dept. of ECE, [2]Dept. of CSE, UC San Diego, La Jolla, CA

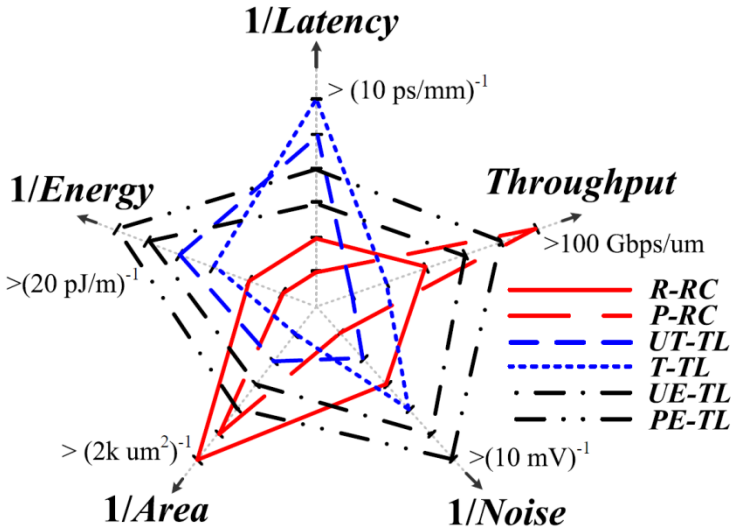
12th International Workshop on System Level Interconnect Prediction
June 13, 2010 Anaheim, USA

Outline

- Introduction
- Pipelined Global Interconnects
  - Overview
  - Glossary
  - Assumption and Modeling
- Design Objectives and Metrics
  - Design objectives
  - Performance metrics
- Performance Evaluation Flow
- Experimental Results
  - Pipelining effect
  - Voltage/ Technology scaling
  - Design example
- Conclusion

Wire Scaling Issues and Design Criteria

- On-chip global wires become barrier for achieving
  - High-performance:
    - *542ps (1mm wire) vs. 161ps (10 FO4 inverter)* [ITRS 2008]
  - Low-power:
    - *Contribution for 50% dynamic power.* [Magen 2004]
- Various interconnect schemes proposed
  - RC wires
  - On-chip T-lines
    - *transceiver design, equalization, etc*
  - Design criteria
    - ***minimum latency***

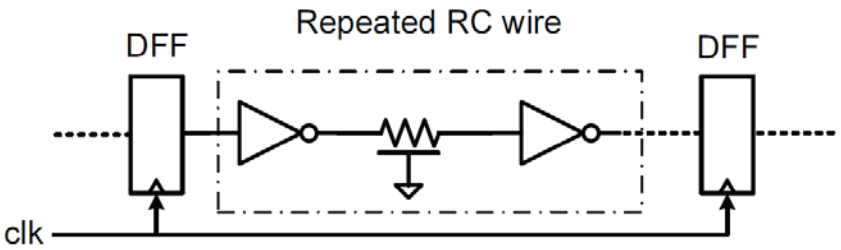


[Zhang 2009]

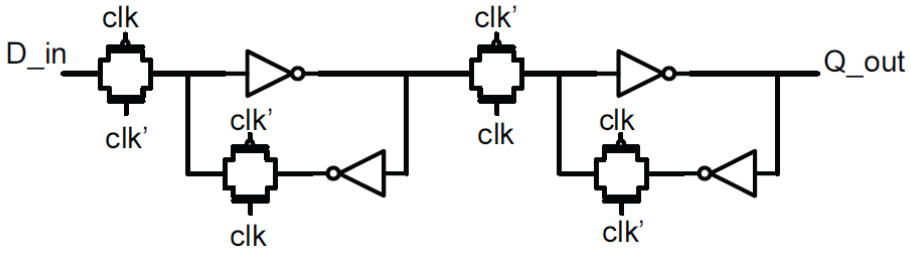
Throughput-Centric Interconnect Design

- Throughput-centric interconnect design [Shah 2002] become necessary because
  - Increasing demand for computing capacity
  - Emerging parallel computing architectures
  - More stringent ***throughput*** requirement of on-chip interconnects
    - Wires in the NoCs (Networks-on-Chips) [Jantsch 2003]
- Our work
  - Wires are *pipelined* to meet required clock period (throughput)
  - Explore the power-saving of pipelined interconnects with more design freedoms
  - Optimize for different applications
    - High-Performance / Low-Power / Moderate Cost

Overview of Pipelined Global Interconnects



One stage of pipelined interconnect.



Schematic of a latch-based D flip-flop.

- Adopt flip-flop based pipelining structure [Heo 2005]
  - Flip-flop inserted to meet throughput
  - Repeaters inserted for delay optimization
- Two-stage latch-based D flip-flop
- Knobs for manipulating pipelined interconnects
  - Wire geometries / repeater placement
  - Pipelining depth / supply voltage

Glossary for Pipelined Global Interconnects

Table 1: Symbols used for variables and parameters of pipelined global interconnects.

f_{clock}	Target clock frequency [3]
l	Total wire length
N	Number of pipelined stages
V_{dd}	Supply voltage
w	Wire width
$pitch$	Wire pitch
s_{inv}	The scaled size of the repeater
l_{inv}	The repeater interval
t	Wire thickness
h	Dielectric height
ρ	The copper resistivity
r_w	Wire resistance per unit length
c_w	Wire capacitance per unit length
r_0	Output resistance of a min-sized repeater
c_{nmos}	Min-sized NMOS gate capacitance
I_{leak}	The leakage current for one min-sized repeater
η_{leak}	The ratio between leakage and dynamic power
C_{FF}	Effective capacitance of a flip-flop
d_{FF}	Delay of a flip-flop at nominal V_{dd}
g=1.34	P/N ratio of transistor width
f	The diffusion to gate capacitance ratio
a=0.4,b=0.7	Constants related to transistor switching model [13]
d_{seg}	The delay of each repeater-wire segment
e_{seg}	The energy dissipation of each repeater-wire segment

Table 2: Design parameters for global pipelined interconnect based on ITRS Roadmap 2008 and predictive SPICE models.

Year	2007	2010	2013	2016
Technology node (nm)	65	45	32	22
Target clock freq. (GHz)	5.06	5.88	7.34	9.18
Supply voltage (V)	1.1	1.0	0.9	0.8
Interlayer dielectric constant	2.9	2.6	2.4	2.1
Copper resistivity[1] ($\mu\Omega \cdot cm$)	2.73	3.10	3.52	3.93
Min global pitch (nm)	210	135	96	75
Aspect ratio (A/R)	2.3	2.4	2.5	2.6
Resistance r_0 of min-repeater[2] ($k\Omega$)	19.3	16.2	23.6	37.5
Leakage current at 100°C[2,3] (nA/nm)	0.22	0.085	0.18	0.38
Flip-flop capacitance[2] (fF)	16.4	10.2	6.94	4.78
Flip-flop delay[2,3] (ps)	90.3	63.2	58.4	57.3

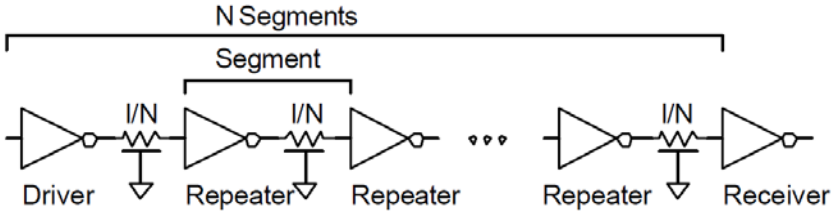
[1] The copper resistivity includes scattering and barrier effect.
[2] Data are obtained by simulation using predictive models [14].
[3] Data are measured under nominal supply voltages.

- Physical parameters of interconnects/repeaters
  - Calculated from ITRS data or based on SPICE characterization
- Define delay/energy dissipation of one repeater-wire segment

Assumptions and Modeling

- Assumptions
  - Repeaters/flip-flops are inserted evenly along the wire.
  - Repeaters/flip-flops are equally sized.
  - The size of flip-flop is fixed and optimized for the average-sized repeater loading.
- Repeated wire modeling [Zhang 2007]
  - Wire delay: *Elmore delay*

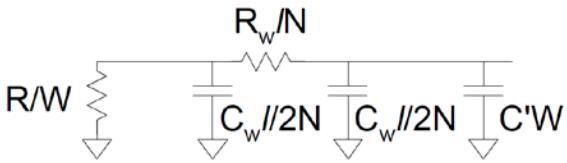
$$d_{seg} = b(1+f)(1+g)r_0 c_{nmos} + a r_w c_w l_{inv}^2 + b r_0 c_w l_{inv}/s_{inv} + b(1+g) r_w c_{nmos} l_{inv} s_{inv}$$



  - Wire energy: *dynamic + leakage*

$$e_{seg} = \alpha_{sw} C_{eff} V_{dd}^2$$

$$C_{eff} = c_w l_{inv} + (1.1 + \eta_{leak})(1+f)(1+g) c_{nmos} s_{inv}$$



    - α_{sw} *is the data activity.*

Modeling of Pipelined Interconnects

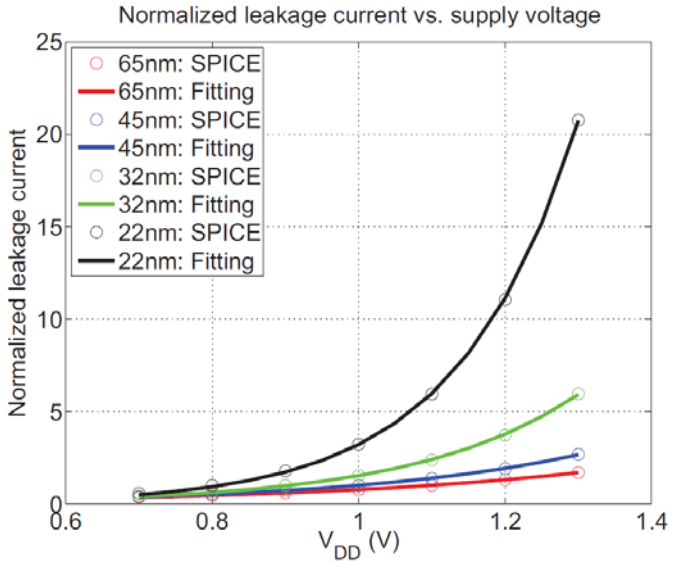
- Considering delay/energy overhead of flip-flops
  - Effective capacitance: C_{FF}
  - Delay of flip-flop: d_{FF}
- Performance modeling
  - Delay

    $$d_{total} = (l/l_{inv})d_{seg} + Nd_{FF}$$
  - Energy

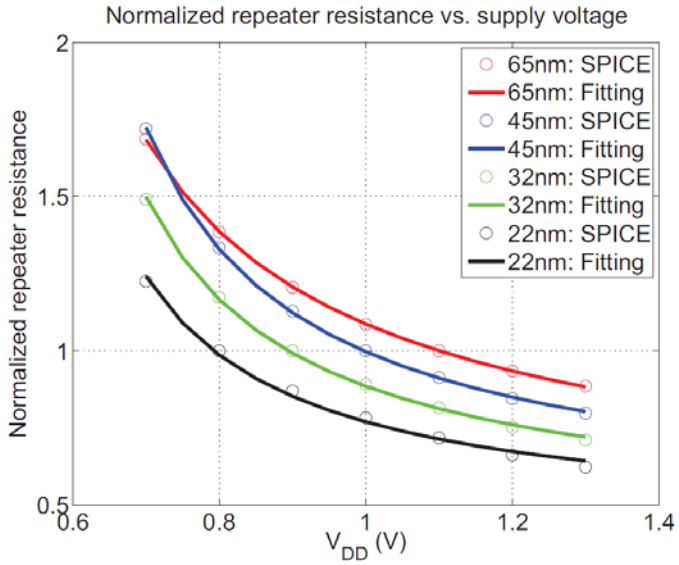
    $$e_{total} = (l/l_{inv})e_{seg} + N\alpha_{sw}C_{FF}V_{dd}^2$$
  - Throughput

    $$f_{bw} = N/d_{total} = \frac{1}{(l/N)(d_{seg}/l_{inv}) + d_{FF}}$$
- Observations
  - Throughput improved with more FFs but larger delay/energy.
  - With the constraint of target throughput, cost of adding FFs can be minimized.

Voltage Scaling Modeling



(a) Modeling leakage current with voltage scaling.



(b) Modeling repeater resistance with voltage scaling.

- Leakage current
  - *Exponential function* [Rabaey 2009]

$$I_{leak}(V_{dd}) = K_1 e^{K_2 V_{dd}}$$

$$\eta^n_{leak}(V_{dd}) = \frac{\eta_{leak}(V_{dd})}{\eta_{leak}(V^{nom}_{dd})} = \frac{V^{nom}_{dd}}{V_{dd}} I^n_{leak}(V_{dd})$$

$$\eta_{leak}(V_{dd}) = \eta^n_{leak}(V_{dd}) \times \eta_{leak}(V^{nom}_{dd})$$

- Repeater/FF delay
  - *alpha-power current law* [Rabaey 2009]

$$r_0(V_{dd}) = K_1 \frac{V_{dd}}{(V_{dd} - V_{th})^{K_2}}$$

$$r_0(V_{dd}) = r^n_0(V_{dd}) \times r_0(V^{nom}_{dd})$$

Design Objectives

- Min-Latency
  - For conventional low-latency repeated wire design
  - Fewer FFs but larger energy/area overhead
- Throughput-Centric Designs [Deodhar 2005]
  - Max-TPE (*low-power* application)
    - Optimize *throughput-per-bit-energy* for single pipelined wire
    - Reduce *total energy* for set of parallel wires
  - Max-TPA (*high-performance* application)
    - Optimize *throughput-per-area* for single pipelined wire
    - Reduce *total area* for set of parallel wires
  - Max-TPEA (*moderate-cost* application)
    - Optimize *throughput-per-energy-area* for single pipelined wire
    - Reduce the *total power-area product* for set of parallel wires

Performance Metrics

- Throughput
  - Maximum clock frequency (unit: *GHz* or *Gbps*)
- Latency
  - Normalized latency (unit: *ps/mm*)

$$latency_n = \frac{\text{total latency}}{\text{wire length}}$$

- Energy per Bit
  - Normalized energy per bit (unit: *pJ/mm*)

$$energy_n = \frac{\text{energy per bit}}{\text{wire length}} = \frac{\text{total power}}{\text{throughput} \times \text{wire length}}$$

- TPEA
  - Throughput-per-energy-area (unit: *Gbps/um/pJ*)

$$TPEA = \frac{\text{throughput}}{\text{energy per bit} \times \text{effective pitch}}$$

    - Effective pitch is defined as total area divided by wire length

Performance Evaluation Flow

Algorithm 1 Pipelined Wire Optimization Algorithm

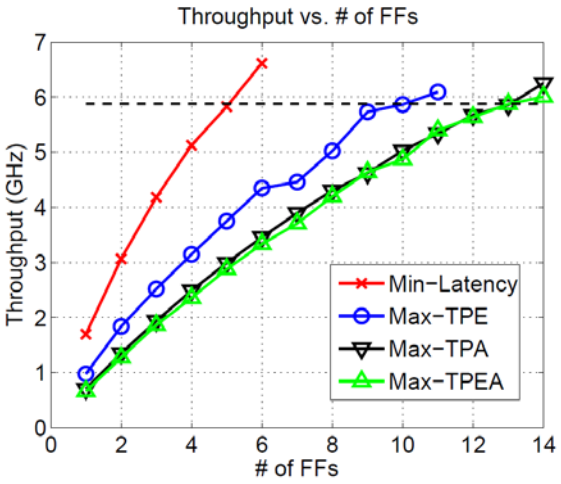
1: Define global and technology parameters
2: Define design *objective*
3: **for** $V_{dd} = V_{dd}^{min}$ to V_{dd}^{max} **do**
4: Compute η_{leak}, r_0, d_{FF}
5: $N \leftarrow 1$
6: **repeat**
7: **for** $pitch = pitch_{min}$ to $pitch_{max}$ **do**
8: **for** $w = w_{min}$ to $pitch$ **do**
9: Compute $r_w(pitch,w)$, $c_w(pitch,w)$
10: s_{inv}, l_{inv}= fminsearch($objective$, r_w, c_w)
11: Compute cost function f
12: **end for**
13: **end for**
14: Search minimum cost $f(N)$
15: Estimate throughput(N), delay(N), and energy(N)
16: $N \leftarrow N + 1$
17: **until** Throughput reaches the target frequency
18: **end for**
19: **return** Optimal design variables: $pitch$, w, s_{inv}, l_{inv}
performance: $f(V_{dd}, N)$, delay(V_{dd}, N), energy(V_{dd}, N)

- Simply the problem by [Nagpal 2007]
  - Limiting the range of wire geometries (pitch, width)
  - Optimize repeater for given wire geometry
- Support different objectives
- FFs are added incrementally until reaching the throughput constraint
- Return performance metrics for given supply voltage (V_{DD}) and pipelining stage (N) and corresponding optimal design.

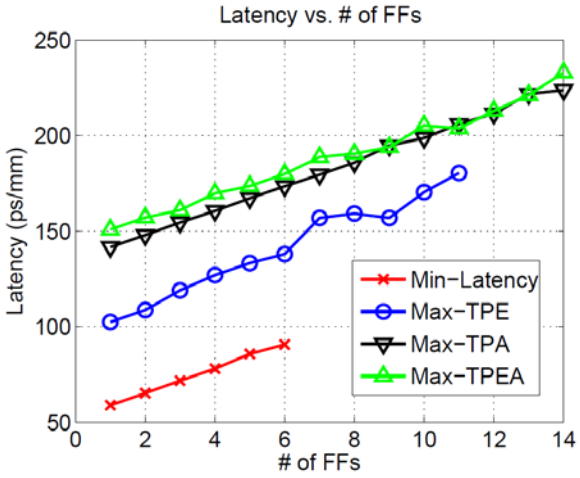
Experimental Settings

- Transistor Models
  - ASU predictive technology models
  - Level 54 BSIM3v3 MOSFET models
- Repeater/Flip-Flop Characterization
  - HSPICE timing/power simulation
  - MATLAB curve regression and whole flow implementation
- Global Wire Parameters
  - Wire length: 10mm
  - Switching factor: 0.2
  - Upper bound of wire pitch: 1um
- Voltage/Technology Scaling
  - Supply voltage: 0.7V → 1.3V (50mV step)
  - Technology: 65nm, 45nm, 32nm, 22nm

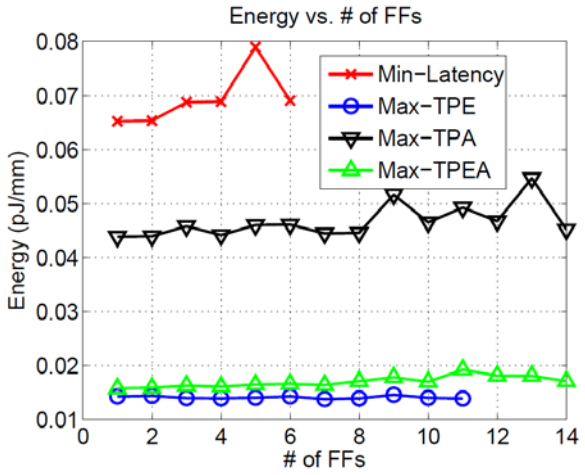
Pipelining Effect



(a) Throughput.



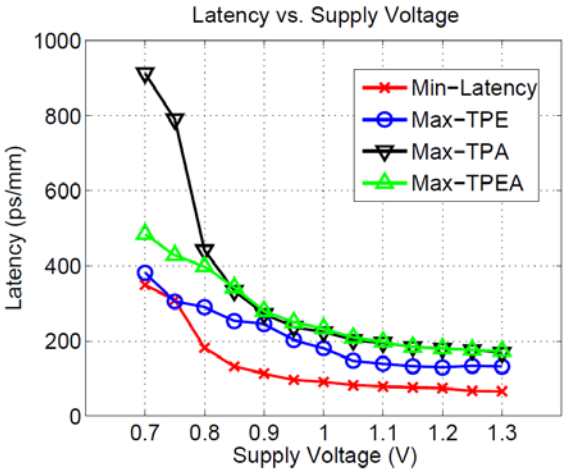
(b) Latency.



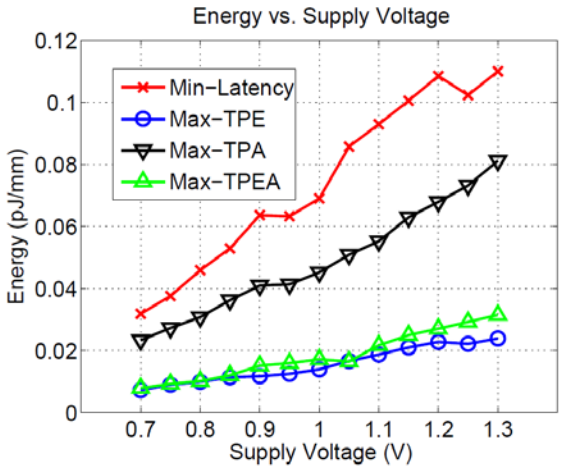
(c) Energy per bit.

- Study impact of pipelining using 45nm under nominal V_{DD}=*1V*
- Throughput is improved with deeper pipelining
  - *Throughput-centric design* uses more FFs
- Latency/Energy increases with deeper pipelining
  - *Min-latency* achieves lowest delay but largest energy
  - *Throughput-centric design* reduces energy greatly (~4x) with delay overhead (~2.5x)

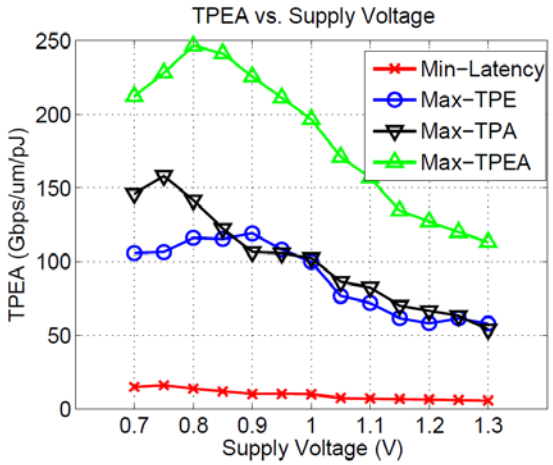
Voltage Scaling Effect



(a) Latency.



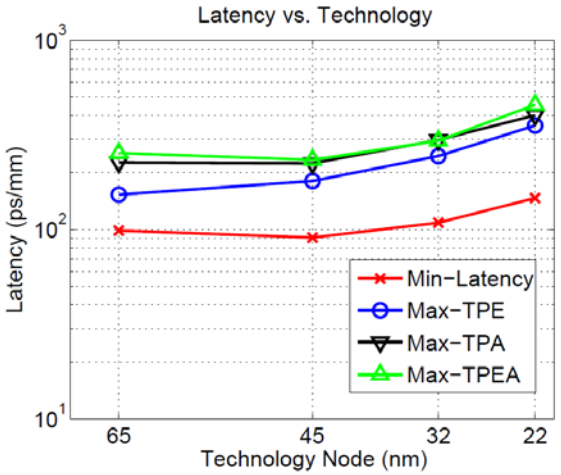
(b) Energy per bit.



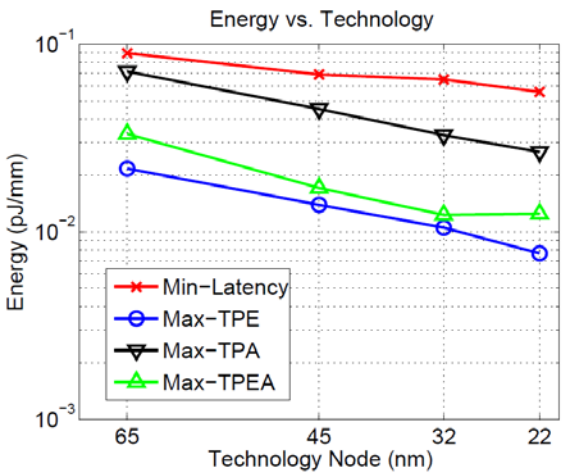
(c) TPEA metric.

- Study impact of voltage scaling using 45nm under throughput constraint
- Latency decreases as V_{DD} increases
  - Tend to saturate when V_{DD} is larger than the nominal value
  - Latency increases more quickly for *Min-Latency/TPA* as V_{DD} goes smaller
- Energy increases as V_{DD} increases
  - Similarly, energy of *Min-Latency/TPA* increases more quickly
- Optimal V_{DD} for TPEA metric
  - Reducing V_{DD} improves TPEA for *throughput-centric designs*.

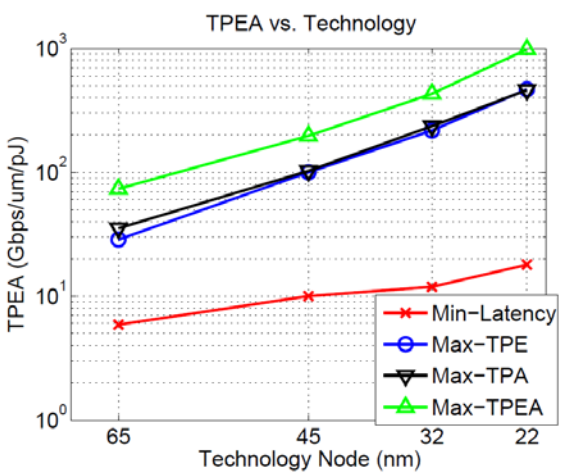
Technology Scaling Effect



(a) Latency.



(b) Energy per bit.



(c) TPEA metric.

- Study impact of technology scaling under nominal V_{DD} and throughput constraint.
- Latency increases nearly exponentially (1.2-1.4x per generation)
  - Drop from 65nm to 45nm due to improved process.
- Energy decreases nearly exponentially (~0.7x per generation)
- TPEA improves with process scaling
  - 2.4x per generation for *throughput-centric designs*
  - 1.5x per generation for *min-latency design*

Design Example

Table 3: Performance comparison of Nominal V_{dd} Design (Min-Latency) and Voltage Scaling Design (Max-TPEA) using 45 nm CMOS process.

Design Variables and Performance Metrics	Nominal V_{dd} Design (Min-Latency)	Voltage Scaling Design (Max-TPEA)
Supply voltage ($V_{dd} : V$) / # of Flip-Flops (N)	1.0 / 6	0.8 / 22
Wire pitch ($pitch : \mu m$) / Wire width ($w : \mu m$)	0.957 / 0.735	0.222 / 0.055
Repeater size (s_{inv}) / Repeater interval ($l_{inv} : mm$)	260x / 0.417	26x / 0.455
Latency (ps/mm)	90.7 (1x)	397.9 (4.4x)
Energy per Bit (pJ/mm)	0.069 (1x)	0.010 (0.14x)
Throughput Density ($Gbps/\mu m$)	6.91 (1x)	24.96 (3.6x)
TPEA ($Gbps/\mu m/pJ$)	10.01 (1x)	246.52 (24.7x)

- Two design criterions are compared using 45nm for the same throughput constraint.
- Design variables
  - Max-TPEA uses deeper pipelining and lower voltage (1.0V → 0.8V)
  - Max-TPEA uses narrower wire (0.07x) and weak repeater (0.1x)
- Performance metrics
  - Latency increases
  - But, energy/area reduces
  - 25x improvement on overall TPEA

Conclusion

- We study the performance of *pipelined* global interconnects with *voltage/process scaling* for different applications.
- ***Throughput-centric*** designs are introduced and compared with ***min-latency*** design:
  - Deeper pipelining to alleviate timing slack and therefore reduce energy/area.
  - 20%-50% overall TPEA improvement by supply voltage scaling.
  - Max-TPEA w/ voltage scaling can improve TPEA by 25x w/ only 4x latency overhead.

Thank You!