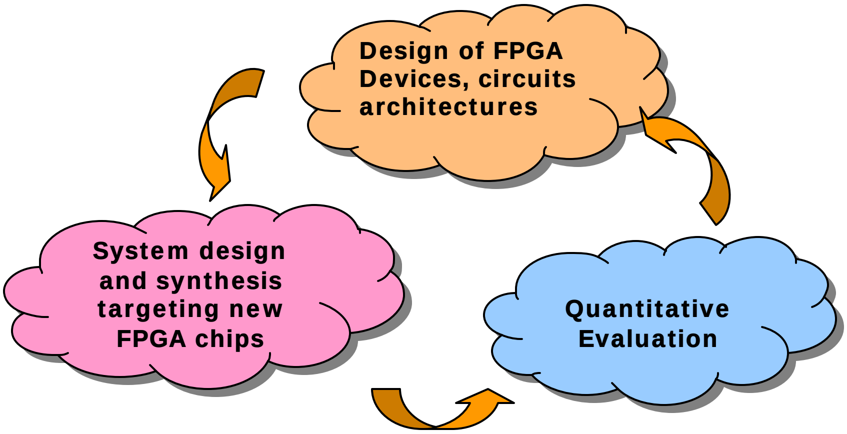
Vdd Programmable and Variation Tolerant FPGA Circuits and Architectures

Prof. Lei He
EE Department, UCLA
LHE@ee.ucla.edu

Partially supported by NSF.

Pathway to Power Efficiency and Variation Tolerance

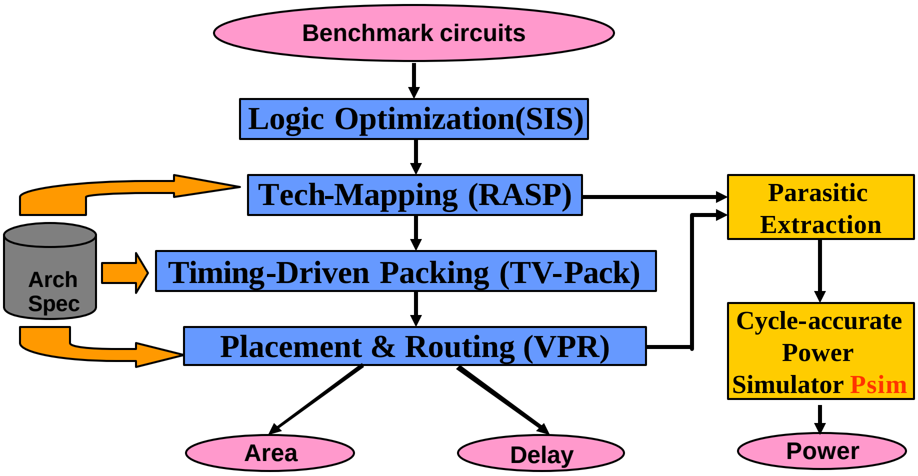
- FPGA = microprocessor


Design of FPGA Devices, circuits architectures
Quantitative Evaluation
System design and synthesis targeting new FPGA chips

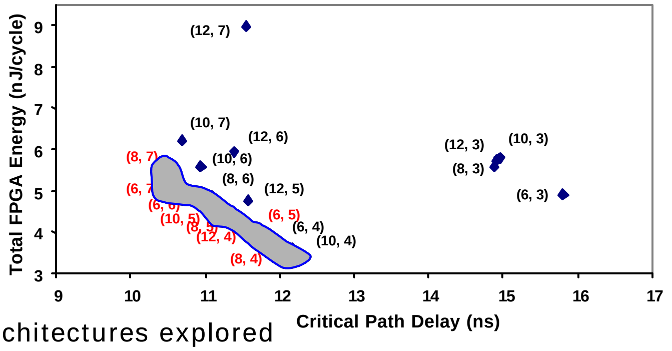
Outline

- Motivation for field programmability of dual-Vdd
- Circuit and architecture for Vdd programmability
- Device and architecture co-optimization
- Impact of Process variations
- References

FPGA Architecture Evaluation



Evaluation of Single-Vdd FPGAs

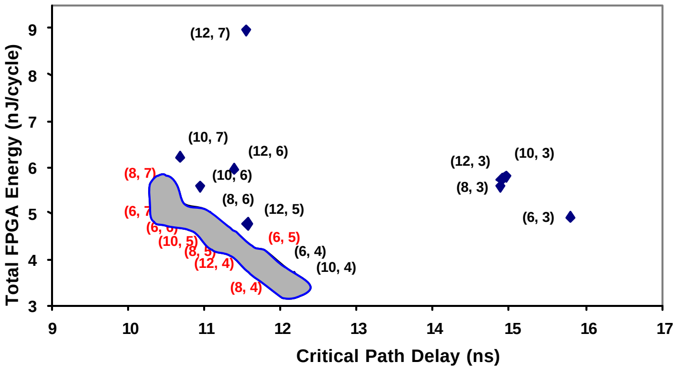


- Architectures explored
  - Cluster size N = {6, 8, 10, 12}
  - LUT size k = {3, 4, 5, 6, 7}
- Energy-delay (ED) dominant architectures
  - Architecture with smaller delay or less energy (compared to any other architecture)
- Relaxed ED dominant set may be also valuable

Energy versus Delay

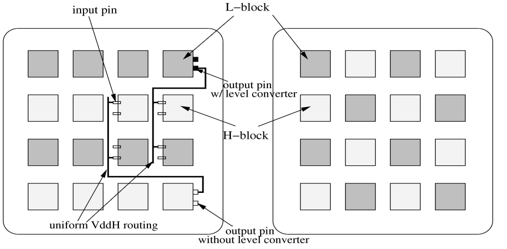
Current commercial architecture

- For 100nm ITRS technology
  - Min-Energy arch (N,k)=(10,4) or (8,4)
  - Min-Delay arch (N,k)=(8,7) ⬄ 0.8x delay but 1.7x power



Idea from ASIC: Pre-Defined Dual-Vdd Fabric

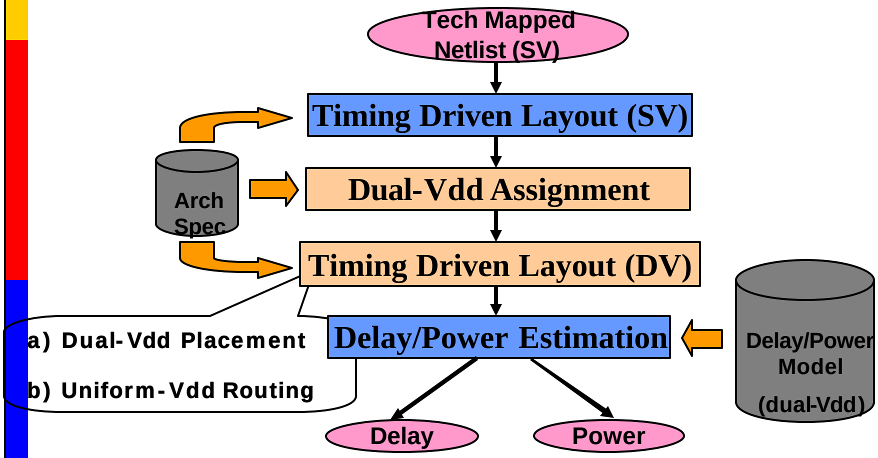
- Each logic block has a pre-defined Vdd level
  - L-block: slot of VddL logic block in the fabric
  - H-block: slot of VddH logic block in the fabric
- Physical locations of logic blocks define layout patterns
  - Row-based (Ratio VddL row/VddH row)
  - Interleaved (Ratio VddL cluster/VddH cluster)
- For time being, interconnect uses uniform VddH



(a) Row–based dual–Vdd layout pattern (Ratio VddL Row/VddH Row = 1:1)

(b) Interleaved dual Vdd layout pattern (Ratio VddL Block/ VddH Block =1:1

Simple yet Practical Design Flow



Tech Mapped Netlist (SV)

Timing Driven Layout (SV)

Dual-Vdd Assignment

Timing Driven Layout (DV)

Delay/Power Estimation

Arch Spec

Delay/Power Model (dual-Vdd)

a) Dual-Vdd Placement

b) Uniform-Vdd Routing

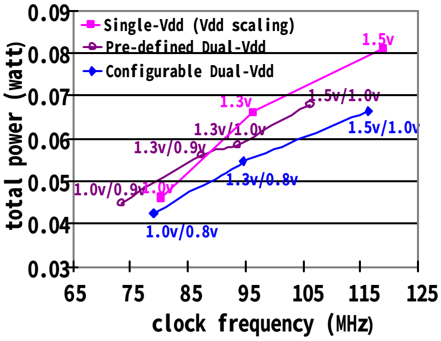
Delay

Power

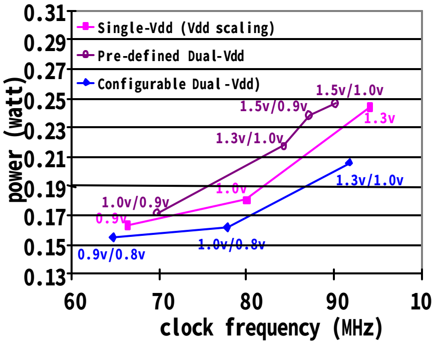
Experiments on Pre-Defined Dual-Vdd Fabric

- Pre-Defined dual-Vdd fabric is not always effective to reduce power
  - Layout constraint in placement => excessive delay increase

circuit: alu4



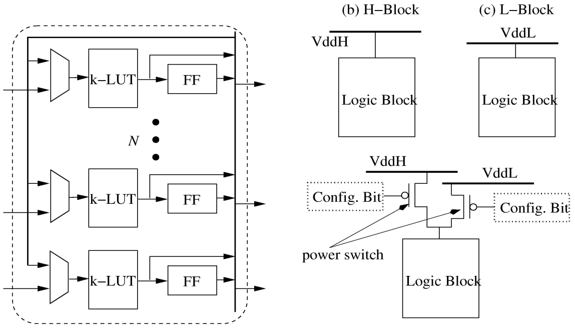
circuit: s38584



Solutions: Configurable dual-Vdd fabric

Logic Blocks with Vdd Programmability

- H-block (VddH), L-block (VddL)
- **P-block** (programmable Vdd)
  - Support Vdd selection and power gating

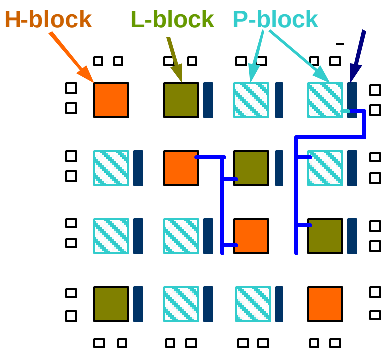


(a) Logic Block

(d) P–Block

Configurable Dual-Vdd Fabrics

- Interleaved dual-vdd layout patterns
  - Interleaved sequence in each row (H-block → L-block → P-block)
  - Ratio H-block/L-block/P-block is pre-determined
  - Vdd-level converters are inserted
- 100% P-blocks
  - With more device overhead and more power reduction

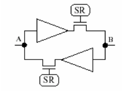


Outline

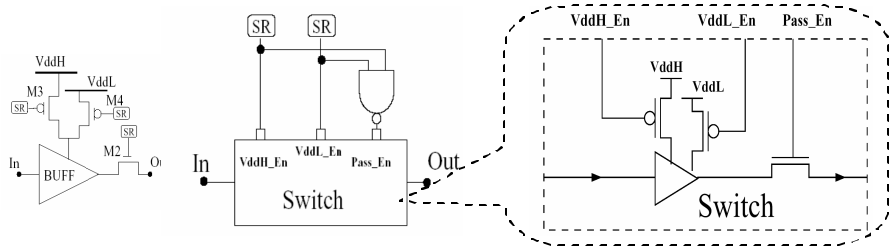
- Motivation for field programmability of dual-Vdd
- Circuit and architecture for Vdd programmability
- Device and architecture co-optimization
- Impact of Process variations
- References

Vdd-programmable Routing Switch

- Routing switch is of paramount importance



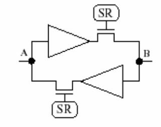
- Vdd-programmable routing switch
  - Brute-force design
    - Two extra SRAM cells for each routing switch



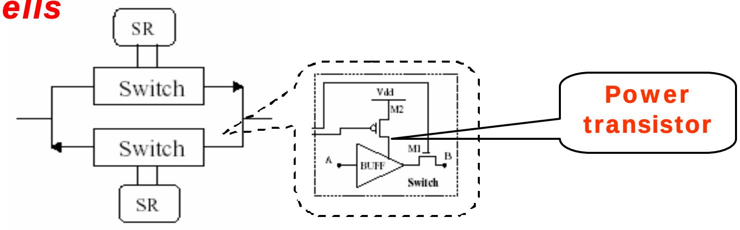
  - New design
    - ***One extra SRAM cell***
    - NAND2 gate — minimum size & high-Vt transistor

Vdd-gateable Routing Switch

- Conventional



- Vdd-gateable
  - two states ⇔ Normal Vdd or Power-gating
- Enable power-gating capability ***w/o extra SRAM cells***



- Can be replaced by tri-state buffer
- Can be extended to MUX

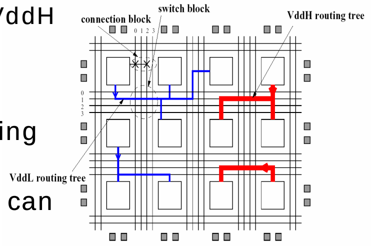
FPGA Architecture Classes

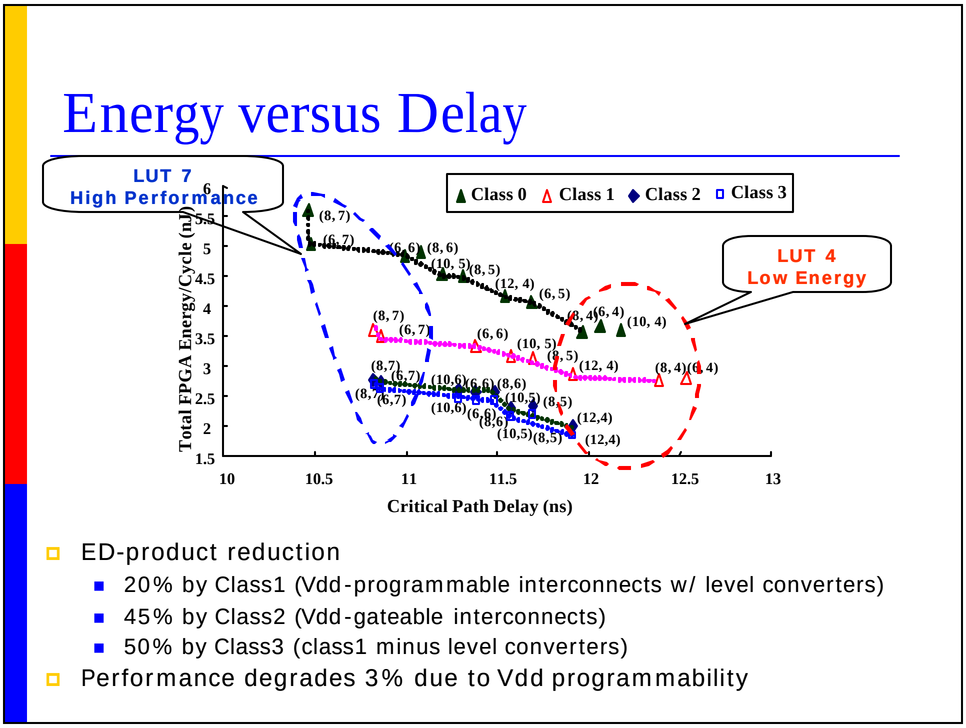
Architecture Class	Logic Block	Interconnect
Class0 (baseline)	single-Vdd	single-Vdd
Class1	programmable dual-Vdd	programmable dual-Vdd, level converters in routing
Class2	programmable dual-Vdd	VddH and Vdd-gateable
Class3	programmable dual-Vdd	Class 1, but no level converters in routing

- High-Vt is applied to configuration SRAM cells for all the classes

Vdd-level Converters

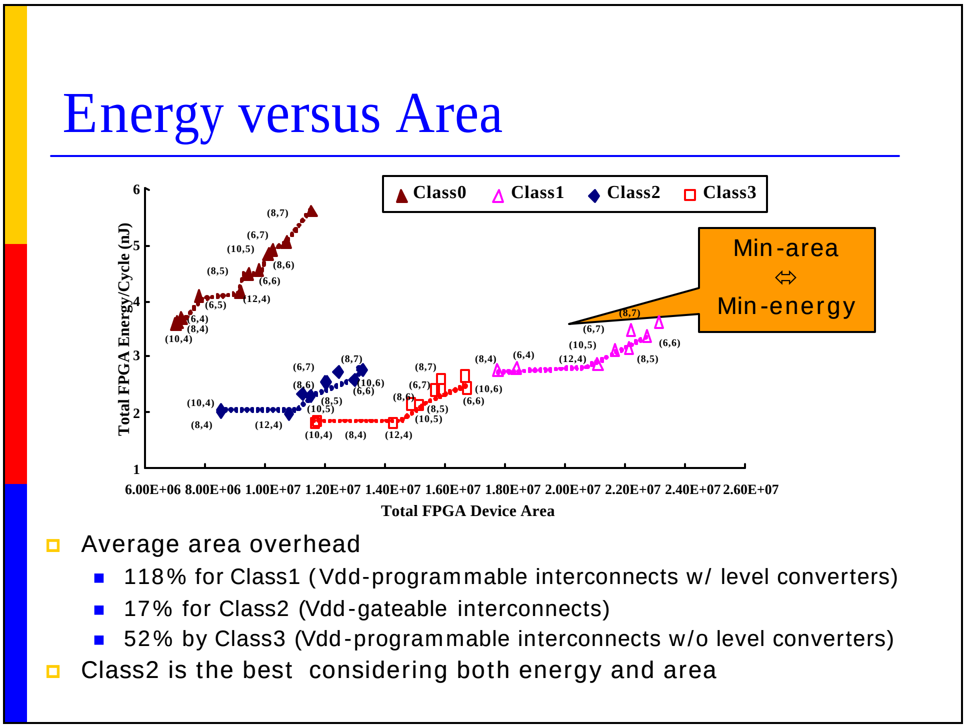
- Class3 removes Vdd-level converters from interconnects in Class1
  - With constraints that no VddL drives VddH
- We developed a routing that one routing tree has a single Vdd level
  - But trees with different Vdd-levels can share the same wire track





Energy versus Delay

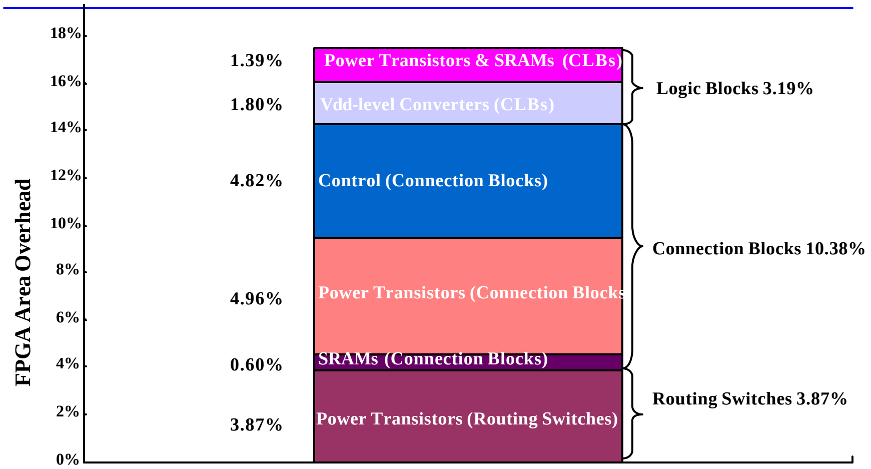
- ED-product reduction
  - 20% by Class1 (Vdd-programmable interconnects w/ level converters)
  - 45% by Class2 (Vdd-gateable interconnects)
  - 50% by Class3 (class1 minus level converters)
- Performance degrades 3% due to Vdd programmability



Energy versus Area

- Average area overhead
  - 118% for Class1 (Vdd-programmable interconnects w/ level converters)
  - 17% for Class2 (Vdd-gateable interconnects)
  - 52% by Class3 (Vdd-programmable interconnects w/o level converters)
- Class2 is the best considering both energy and area

Area Overhead

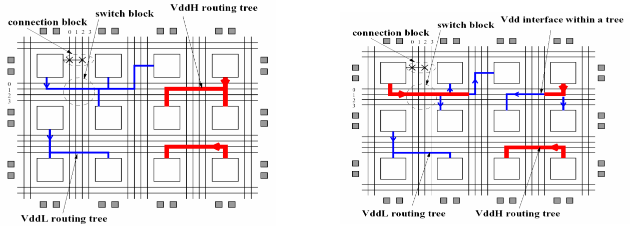


Class2: Vdd-gateable interconnects + Vdd-programmable CLBs(12, 4)

- 17% = 9% for power transistors + 5% for control + 2% for SRAM

More on Dual-Vdd Interconnects

- Tree based LC insertion ***(TLC)***
  - allows one type of Vdd-level within one routing tree
- Dual-Vdd tree based LC insertion ***(dTLC)***
  - allows high-Vdd switch drives low-Vdd switches, but not vice versa



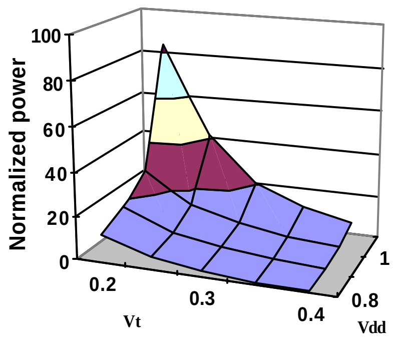
- dTLC achieves ~5% more power reduction compared to TLC
  - A chip level time slack allocation is able to better use the time slack between different routing trees

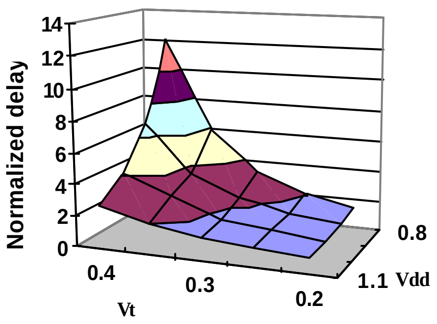
Outline

- Motivation for field programmability of dual-Vdd
- Circuit and architecture for Vdd programmability
- Device and architecture co-optimization
- Impact of Process variations
- References

Impact of Device Tuning

- Device tuning leads to 84X power difference and 12X delay difference
- It is necessary to perform device tuning and architecture tuning simultaneously



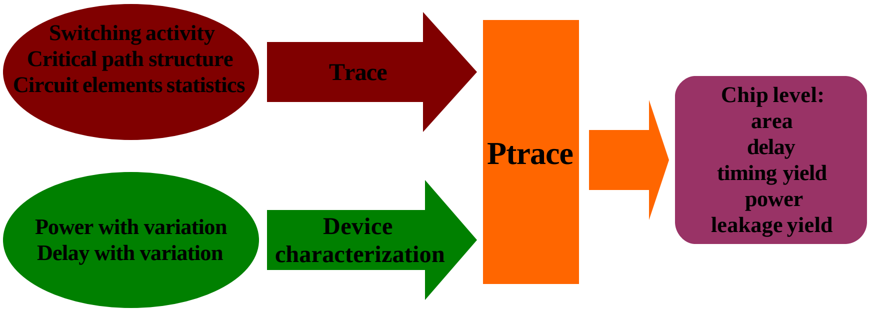


Challenge of Device and Architecture Co-Optimization

- We consider the following architecture and device parameters during our co-optimization:
  - Architecture parameters:
    - Cluster size (N)
    - LUT size (K)
  - Device parameters:
    - Supply voltage (Vdd)
    - Threshold voltage (Vt)
- Hyper-architecture (hyper-arch) is the combination of the device and architecture parameters.
- Large number of hyper-arch combinations
- VPR and Psim are too slow to deal with such large number of experiments
- Need fast yet accurate power and delay estimation

Ptrace:
Trace-Based Power and Timing Models

- **Ptrace is 1000x faster than simulation (VPR and Psim) based evaluation**



Switching activity
Critical path structure
Circuit elements statistics

Trace

Power with variation
Delay with variation

Device characterization

Ptrace

Chip level:
area
delay
timing yield
power
leakage yield

Architectures Classes to be Evaluated

- Hyper-architecture classes

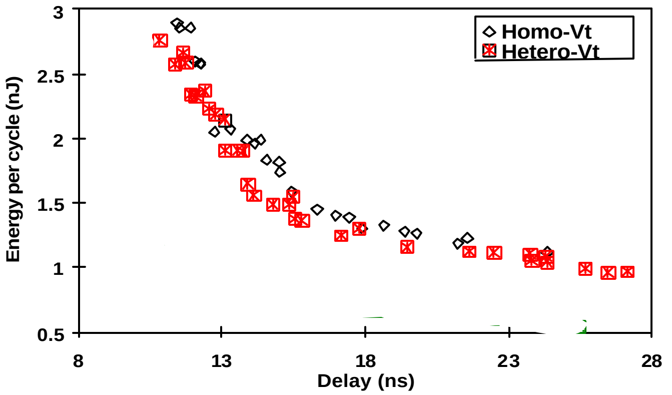
Hyper-arch classes	Vt
Homo-Vt	Homogeneous Vt
Hetero-Vt	Heterogeneous Vt
Homo-Vt+G	Homogeneous Vt + Power Gating
Hetero-Vt+G	Heterogeneous Vt + Power Gating

- Baseline case

Vdd	Vt	LUT size (K)	Cluster size (N)
0.9	0.3	4	8

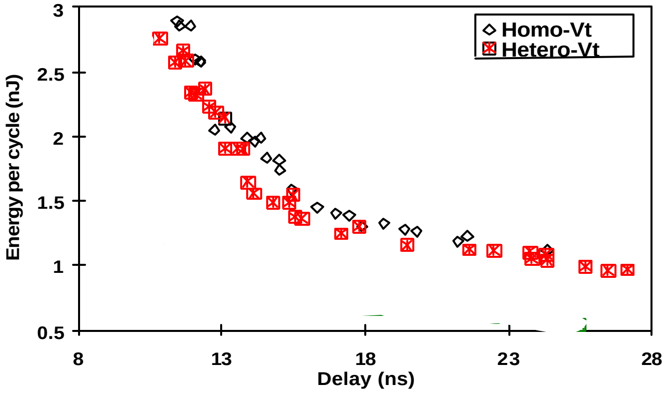
  - Vdd suggested by ITRS
  - Architecture same as Xilinx Virtex-II™.
  - Vt optimized by our method with respect to the above architecture and Vdd

Energy and Delay Tradeoff



- Dominant hyper-arch
  - Hyper-arch B is *inferior* to A if A has less energy and smaller delay than B.
  - Dominant hyper-archs (dom-arch) are the hyper-archs that are NOT inferior to any other hyper-archs.

Energy and Delay Tradeoff



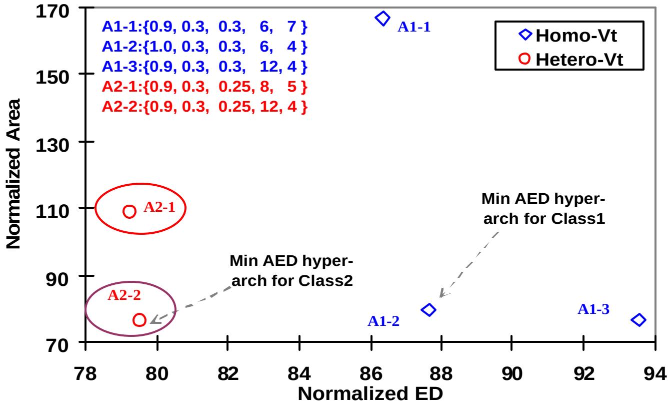
- Hetero-Vt can reduce power
- Power gating reduces more leakage power than hetero-Vt
- Hetero-Vt has less impact when power gating is applied

Min-ED Hyper-Arch

Hyper-arch classes	Vdd (V)	CVt (V)	IVt (V)	(N, K)	ED (nJ·ns)	ED reduction %
Baseline	0.9	0.3	0.3	(8,4)	26.9	-
Homo-Vt	0.9	0.3	0.3	(6,7)	23.3	13.4
Hetero-Vt	0.9	0.2	0.25	(8,4)	21.4	20.5
Homo-Vt+G	0.9	0.25	0.25	(12,4)	11.1	58.9
Hetero-Vt+G	0.9	0.2	0.25	(8,4)	11	59.0

- To achieve the best energy and delay tradeoff, we find out the hyper-arch with the minimum energy and delay product (ED)
  - Compared to the baseline, the min-ED hyper-arch of the conventional FPGA (Homo-Vt) reduces ED by 13.4%
  - For the Hetero-Vt class, ED is reduced by 20.5%
  - If power gating is applied, ED can be reduced by up to 59.0%

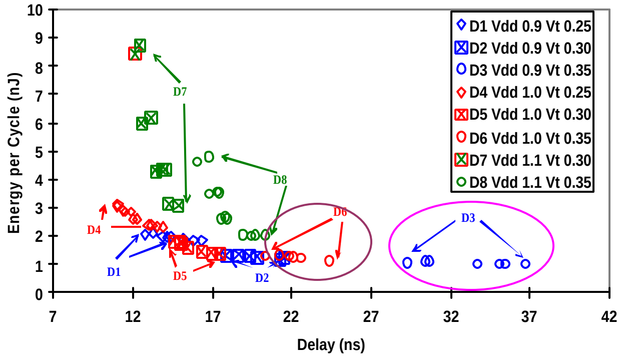
ED and Area without Power Gating



- Compared to the min-ED hyper arch, the min-AED hyper-arch significantly reduce area with a small ED increase

Dom-Archs under Different Device Settings

- For a given device setting architecture tuning changes delay and energy in a smaller range
- Device tuning has a much more impact on delay and energy



Outline

- Motivation for field programmability of dual-Vdd
- Circuit and architecture for Vdd programmability
- Device and architecture co-optimization
- Impact of Process variations
- References

Models of Variations

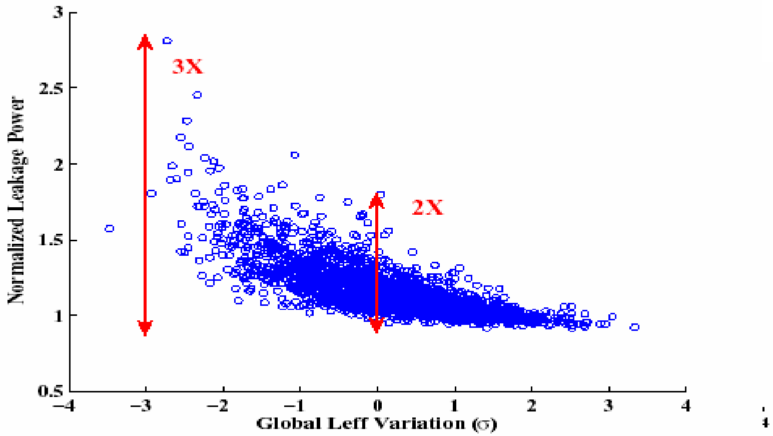
- Source of variations
  - Threshold Voltage (V_{th}) due to doping variation
  - Effective channel length (L_{eff})
  - Oxide thickness (T_{ox})
- Types of variations
  - Die-to-die (global variations)
  - Within-die (local variations)
- Amount of variations
  - 10% of nominal value as 3σ

Methodologies

Variations(σ)			Mean(W)		SD(%)	
(L_g, L_l)	(V_g, V_l)	(T_g, T_l)	M-C	Model	M-C	Model
(±3,0)	(±3,0)	(±3,0)	1.24	1.2	14	13
(±3,±1)	(±3,±1)	(±3,±1)	1.41	1.37	14	13
(±3,±2)	(±3,±2)	(±3,±2)	2.07	2	13	12

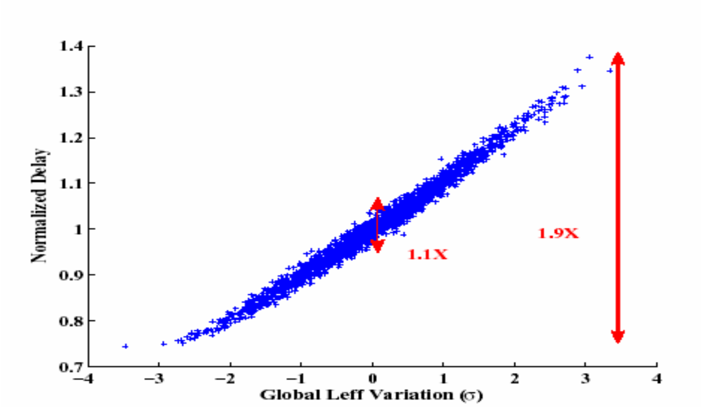
- A set of closed-form formula for leakage and timing variations
- Verified by Monte Carlo Simulation
  - 3% error for mean, and 1% error for standard deviation
- Integrated with Ptrace

Leakage Distribution



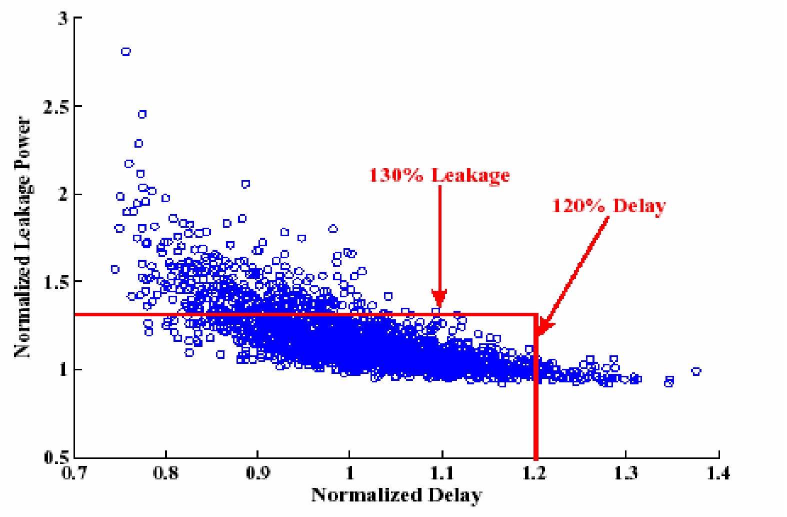
- Leakage is more sensitive to within-die variations

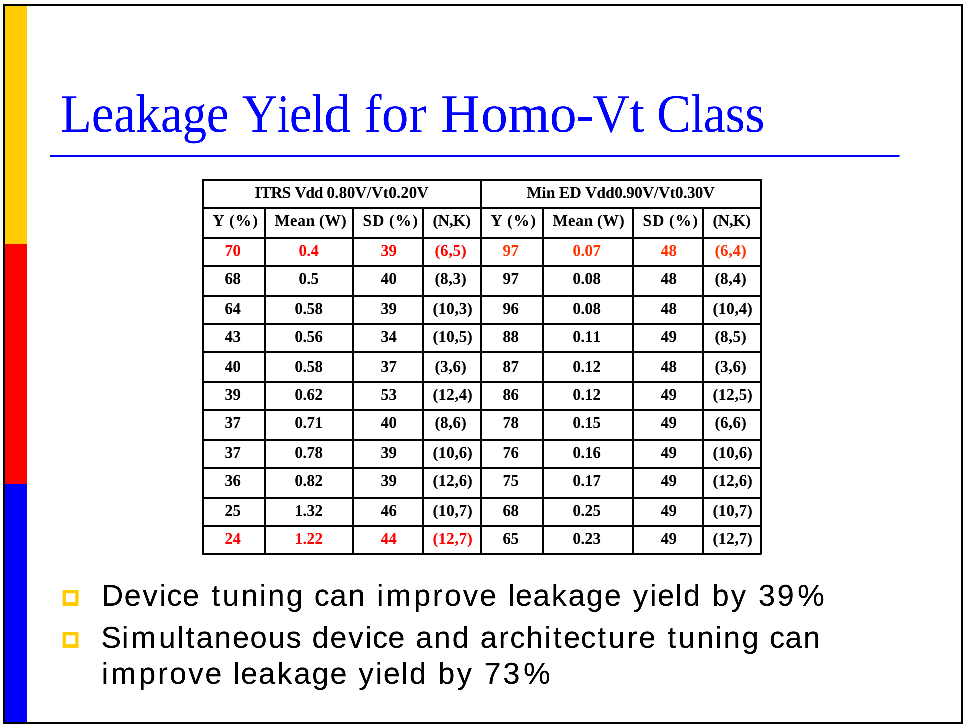
Timing Distribution



- Timing is more sensitive to die-to-die variations

Leakage and Timing Yield Analysis

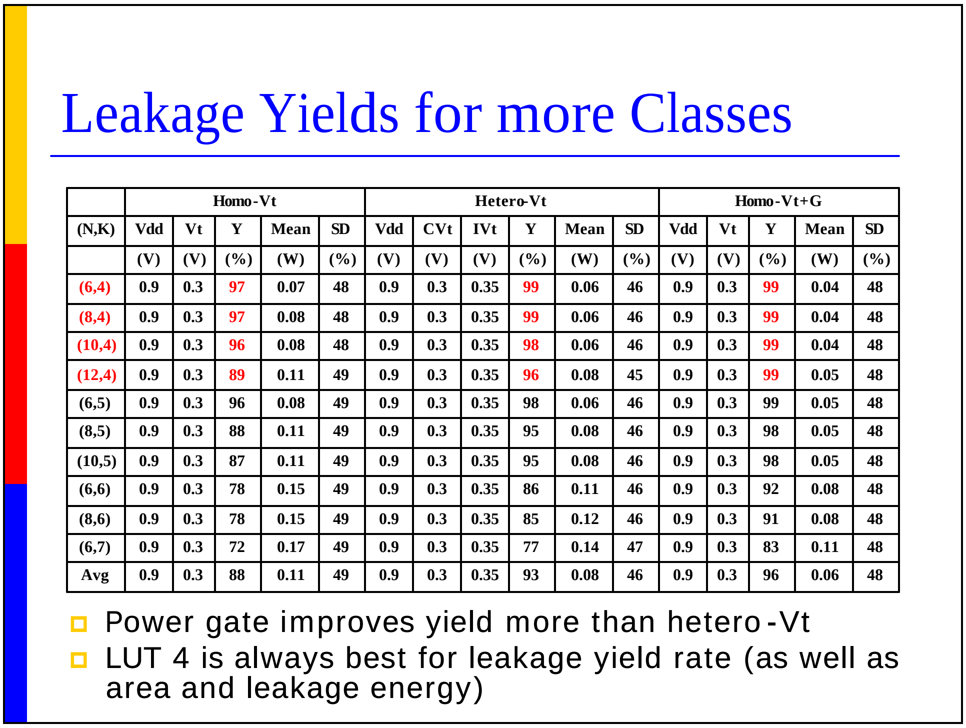




Leakage Yield for Homo-Vt Class

ITRS Vdd 0.80V/Vt0.20V				Min ED Vdd0.90V/Vt0.30V			
Y (%)	Mean (W)	SD (%)	(N,K)	Y (%)	Mean (W)	SD (%)	(N,K)
70	0.4	39	(6,5)	97	0.07	48	(6,4)
68	0.5	40	(8,3)	97	0.08	48	(8,4)
64	0.58	39	(10,3)	96	0.08	48	(10,4)
43	0.56	34	(10,5)	88	0.11	49	(8,5)
40	0.58	37	(3,6)	87	0.12	48	(3,6)
39	0.62	53	(12,4)	86	0.12	49	(12,5)
37	0.71	40	(8,6)	78	0.15	49	(6,6)
37	0.78	39	(10,6)	76	0.16	49	(10,6)
36	0.82	39	(12,6)	75	0.17	49	(12,6)
25	1.32	46	(10,7)	68	0.25	49	(10,7)
24	1.22	44	(12,7)	65	0.23	49	(12,7)

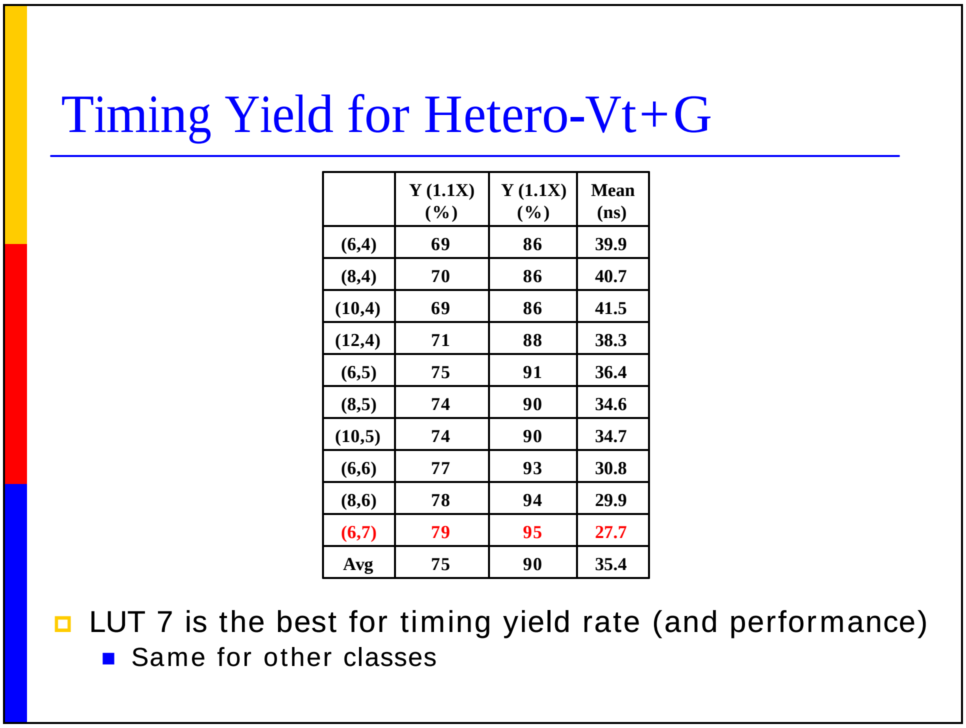
- Device tuning can improve leakage yield by 39%
- Simultaneous device and architecture tuning can improve leakage yield by 73%



Leakage Yields for more Classes

	Homo-Vt					Hetero-Vt						Homo-Vt+G				
(N,K)	Vdd	Vt	Y	Mean	SD	Vdd	CVt	IVt	Y	Mean	SD	Vdd	Vt	Y	Mean	SD
	(V)	(V)	(%)	(W)	(%)	(V)	(V)	(V)	(%)	(W)	(%)	(V)	(V)	(%)	(W)	(%)
(6,4)	0.9	0.3	97	0.07	48	0.9	0.3	0.35	99	0.06	46	0.9	0.3	99	0.04	48
(8,4)	0.9	0.3	97	0.08	48	0.9	0.3	0.35	99	0.06	46	0.9	0.3	99	0.04	48
(10,4)	0.9	0.3	96	0.08	48	0.9	0.3	0.35	98	0.06	46	0.9	0.3	99	0.04	48
(12,4)	0.9	0.3	89	0.11	49	0.9	0.3	0.35	96	0.08	45	0.9	0.3	99	0.05	48
(6,5)	0.9	0.3	96	0.08	49	0.9	0.3	0.35	98	0.06	46	0.9	0.3	99	0.05	48
(8,5)	0.9	0.3	88	0.11	49	0.9	0.3	0.35	95	0.08	46	0.9	0.3	98	0.05	48
(10,5)	0.9	0.3	87	0.11	49	0.9	0.3	0.35	95	0.08	46	0.9	0.3	98	0.05	48
(6,6)	0.9	0.3	78	0.15	49	0.9	0.3	0.35	86	0.11	46	0.9	0.3	92	0.08	48
(8,6)	0.9	0.3	78	0.15	49	0.9	0.3	0.35	85	0.12	46	0.9	0.3	91	0.08	48
(6,7)	0.9	0.3	72	0.17	49	0.9	0.3	0.35	77	0.14	47	0.9	0.3	83	0.11	48
Avg	0.9	0.3	88	0.11	49	0.9	0.3	0.35	93	0.08	46	0.9	0.3	96	0.06	48

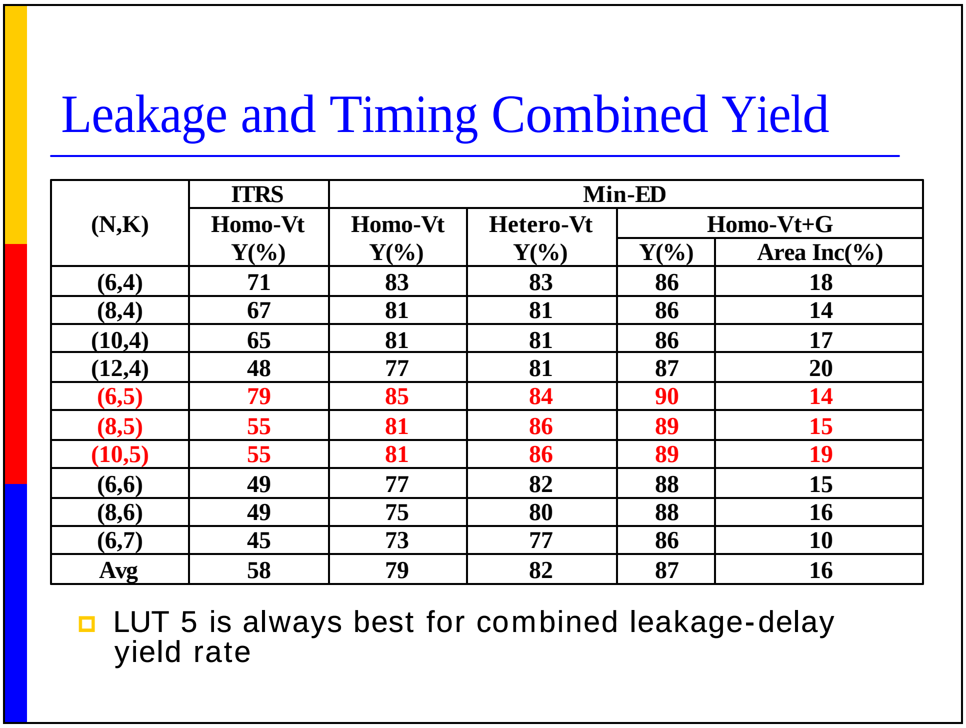
- Power gate improves yield more than hetero-Vt
- LUT 4 is always best for leakage yield rate (as well as area and leakage energy)



Timing Yield for Hetero-Vt+G

	Y (1.1X) (%)	Y (1.1X) (%)	Mean (ns)
(6,4)	69	86	39.9
(8,4)	70	86	40.7
(10,4)	69	86	41.5
(12,4)	71	88	38.3
(6,5)	75	91	36.4
(8,5)	74	90	34.6
(10,5)	74	90	34.7
(6,6)	77	93	30.8
(8,6)	78	94	29.9
(6,7)	79	95	27.7
Avg	75	90	35.4

- LUT 7 is the best for timing yield rate (and performance)
  - Same for other classes



Leakage and Timing Combined Yield

(N,K)	ITRS	Min-ED			
	Homo-Vt	Homo-Vt	Hetero-Vt	Homo-Vt+G	
	Y(%)	Y(%)	Y(%)	Y(%)	Area Inc(%)
(6,4)	71	83	83	86	18
(8,4)	67	81	81	86	14
(10,4)	65	81	81	86	17
(12,4)	48	77	81	87	20
(6,5)	79	85	84	90	14
(8,5)	55	81	86	89	15
(10,5)	55	81	86	89	19
(6,6)	49	77	82	88	15
(8,6)	49	75	80	88	16
(6,7)	45	73	77	86	10
Avg	58	79	82	87	16

- LUT 5 is always best for combined leakage-delay yield rate

Conclusions

- Field programmability is a must for dual-Vdd to obtain power reduction without performance loss
- Field programmability can be achieved with little SRAM increase by programming Vdd path (rather than signal path)
- Simultaneous device and architecture tuning obtains largest gain
- FPGA architectures are NOT equal in terms of parametric yield
  - In addition to area, performance and power

References at http://eda.ee.ucla.edu

- P. Wong, L. Cheng, Y. Lin and L. He, ‘FPGA Device and Architecture Evaluation Considering Process Variation ”, ICCAD 05
- L. Cheng, P. Wong, F. Li, Y. Lin and L. He, ‘Device and Architecture Co-Optimization for FPGA Power Reduction ”, DAC 05
- Y. Lin and L. He, ‘Leakage efficient chip-level dual-vdd assignment with time slack allocation for FPGA power reduction ”, DAC 05
- Y. Lin, F. Li and L. He, ‘Power Modeling and Architecture Evaluation for FPGA with Novel Circuits for Vdd Programmability ”, FPGA 05
- Y. Lin, F. Li and L. He, ‘Routing Track Duplication with Fine-Grained Power-Gating for FPGA Interconnect Power Reduction ”, ASPDAC 05
- F. Li, Y. Lin and L. He, ‘FPGA Power Reduction Using Configurable Dual-Vdd ”, ICCAD 04
- F. Li, Y. Lin and L. He, ‘FPGA Power Reduction Using Configurable Dual-Vdd ”, DAC 04
- F. Li, Y. Lin and L. He, ‘Low Power FPGA Using Pre-defined Dual-Vdd/Dual-Vt Fabrics ”, FPGA 04
- F. Li et al, ‘Architecture Evaluation for Power-efficient FPGAs ”, FPGA 03
- Patents pending