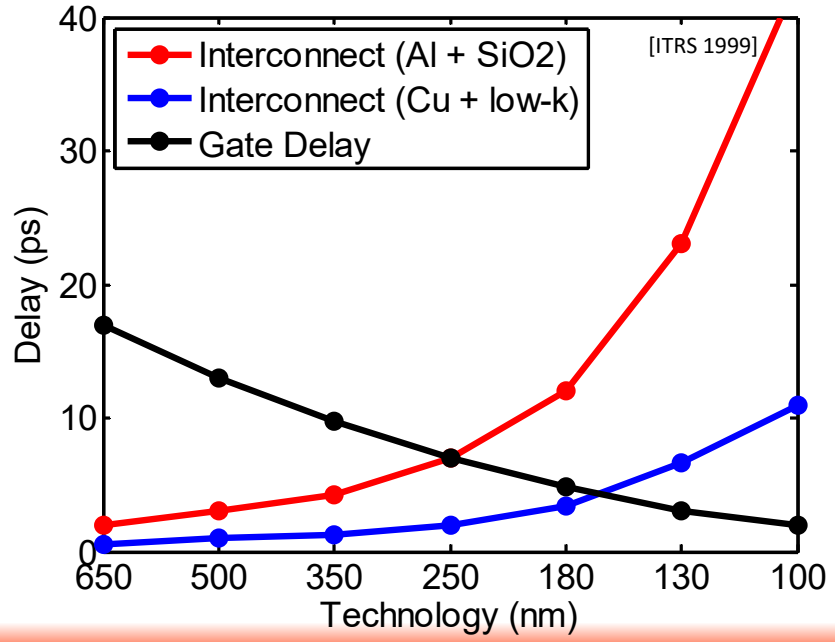
Micro-Fluidically Cooled 3D FPGAs

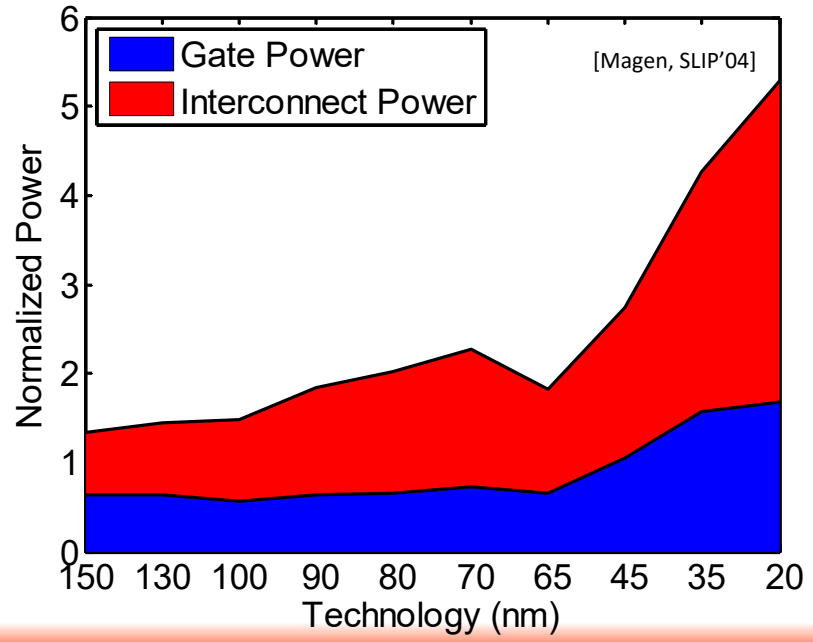
-- Physical Design Exploration

Zhiyuan Yang, Ankur Srivastava
Department of Electrical and Computer Engineering
University of Maryland, College Park

3D FGPAs

- **Dilemma of 2D FPGAs**
  - Large amount of interconnect fabrics
    - **Dominate** delay
    - **Dominate** power
  - Possible Solution: **3D Integration**

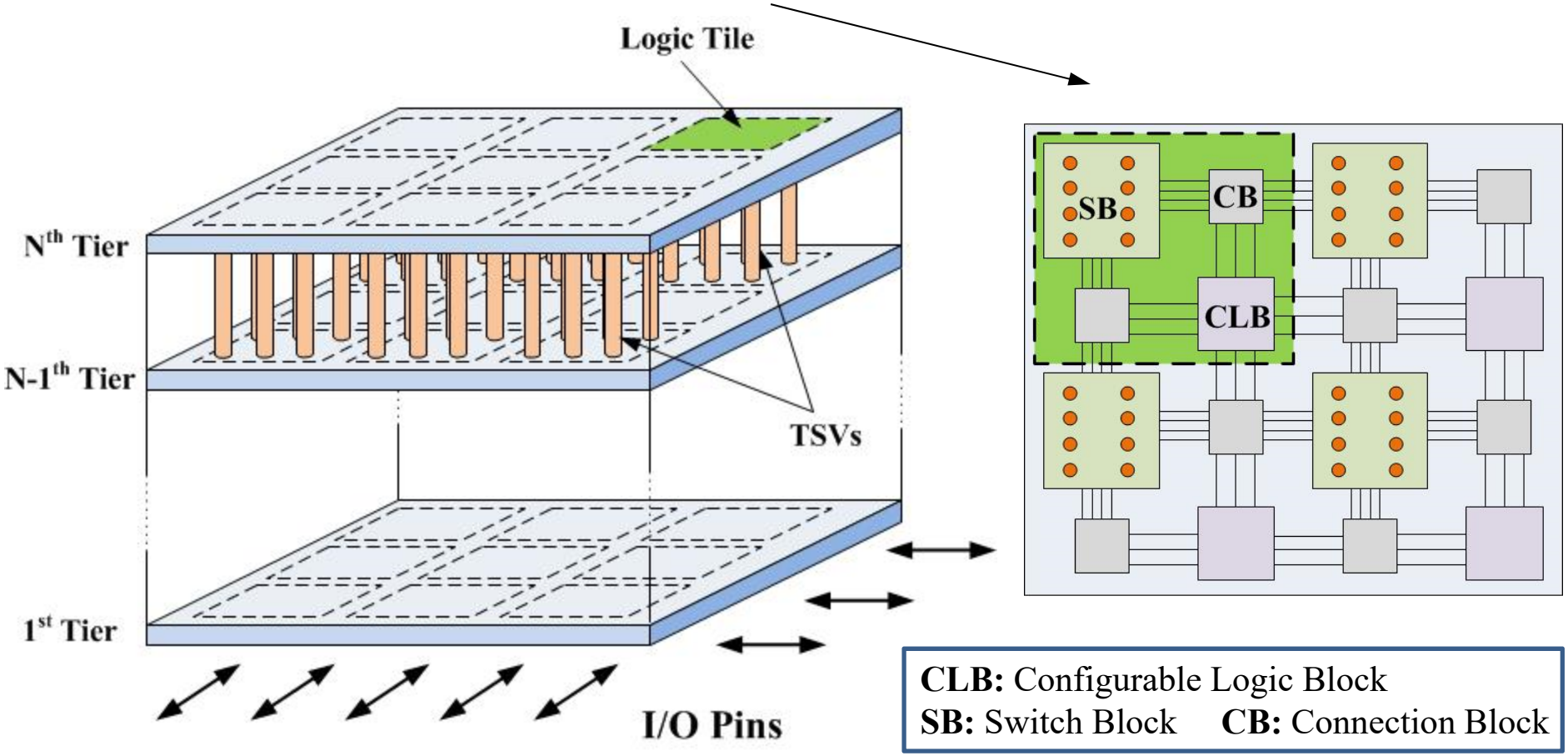




3D FGPAs

- **Stacking Multiple FPGA Tiers**
  - ❑ Physical Structure

3D Tile: the logic tile containing TSVs
2D Tile: the logic tile containing no TSVs

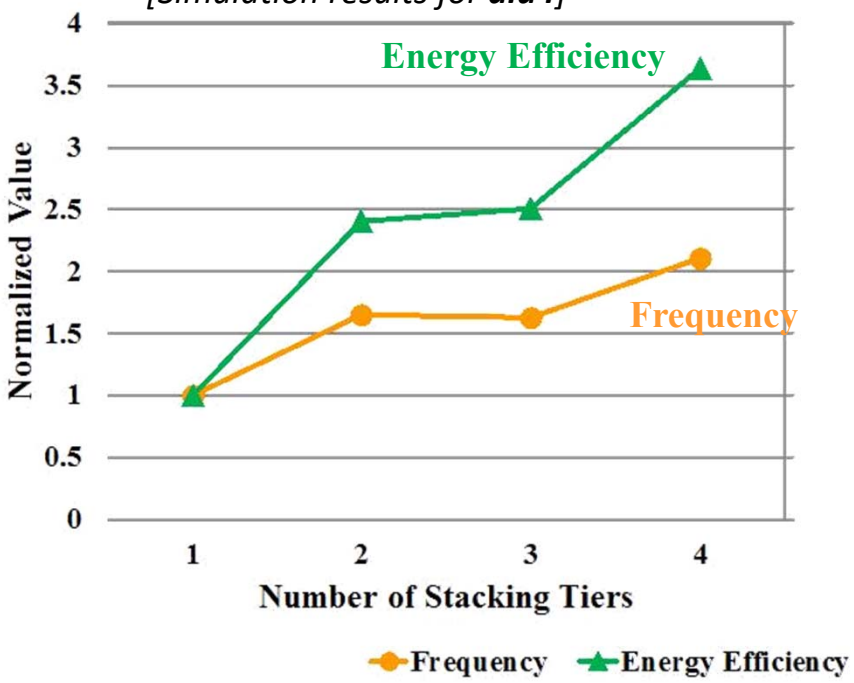


CLB: Configurable Logic Block
SB: Switch Block **CB:** Connection Block

3D FGPAs

- **Stacking Multiple FPGA Tiers**
  - ❑ Benefits
    - Reduce interconnect length
    - Improve performance and efficiency

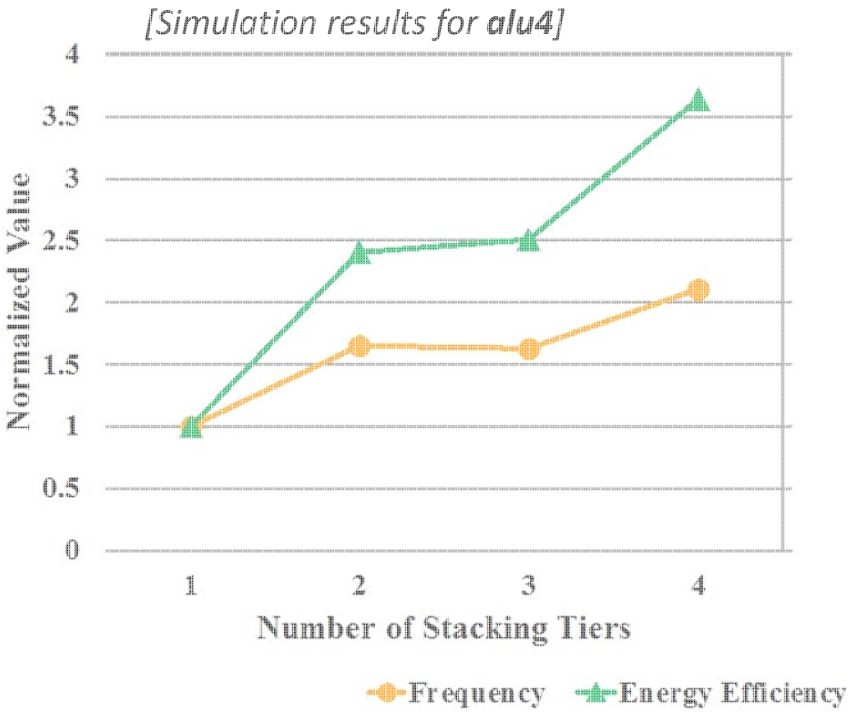
*[Simulation results for **alu4**]*



$$Energy\ Efficiency = \frac{Frequency^2}{Power}$$

3D FGPAs

- **Stacking Multiple FPGA Tiers**
  - ❑ Benefits
    - Reduce interconnect length
    - Improve performance and efficiency
  - ❑ Thermal Challenges



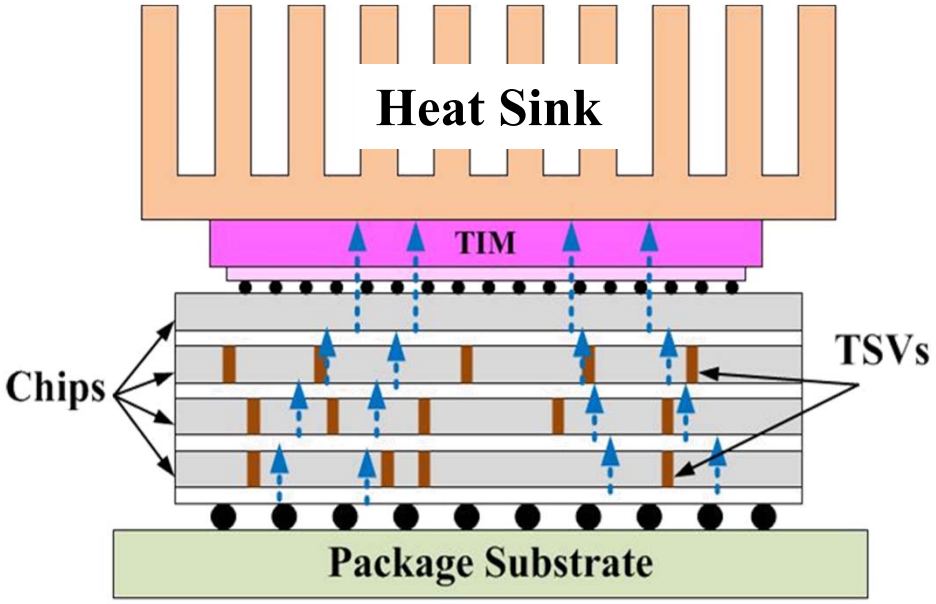
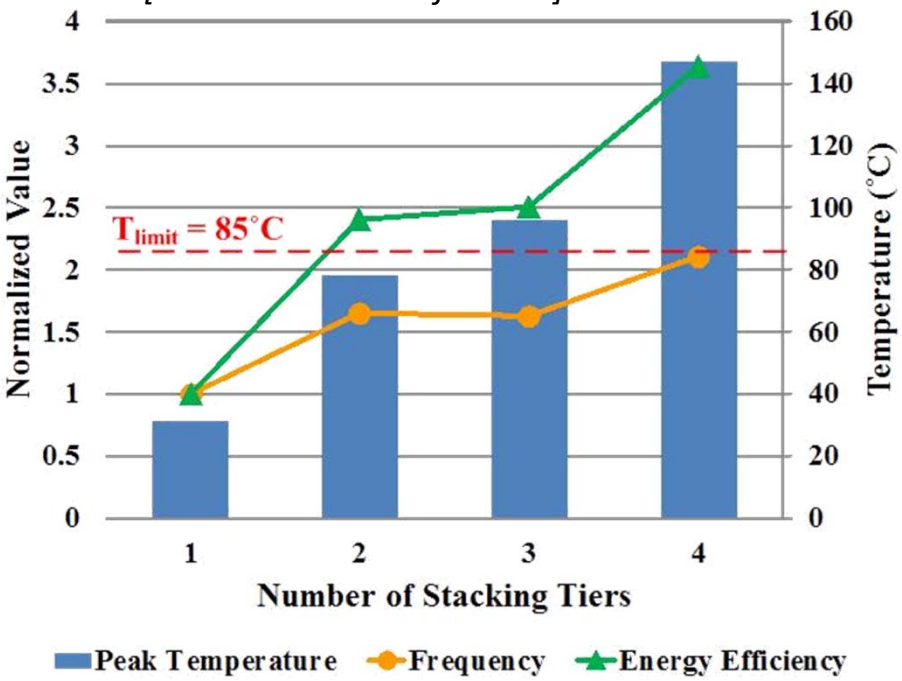


Illustration of Air Cooling

3D FGPAs

- **Stacking Multiple FPGA Tiers**
  - ❑ Benefits
    - Reduce interconnect length
    - Improve performance and efficiency
  - ❑ Thermal Challenges
    - Thermal violation (Peak Temperature > T_{limit})
    - Degrade performance and energy efficiency

*[Simulation results for **alu4**]*



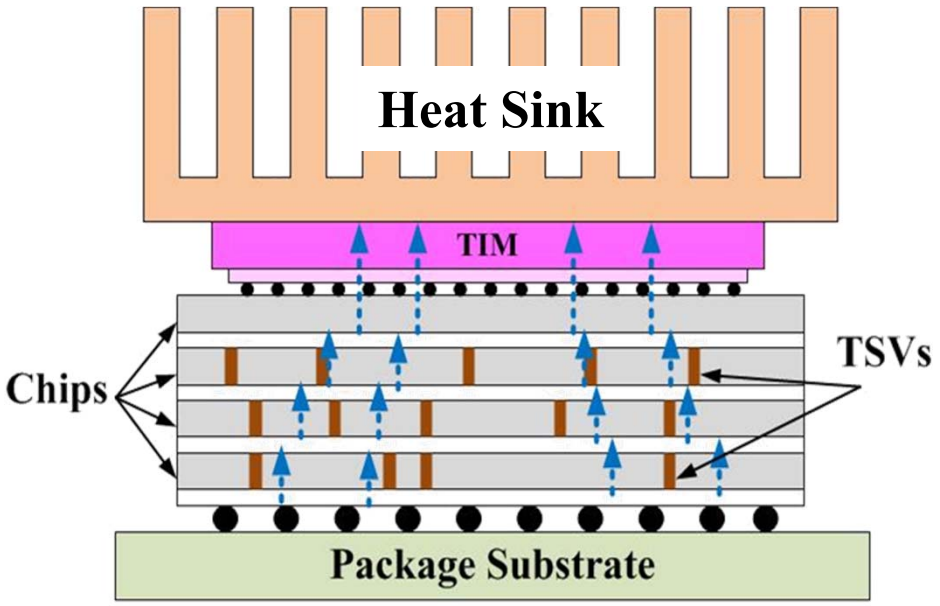
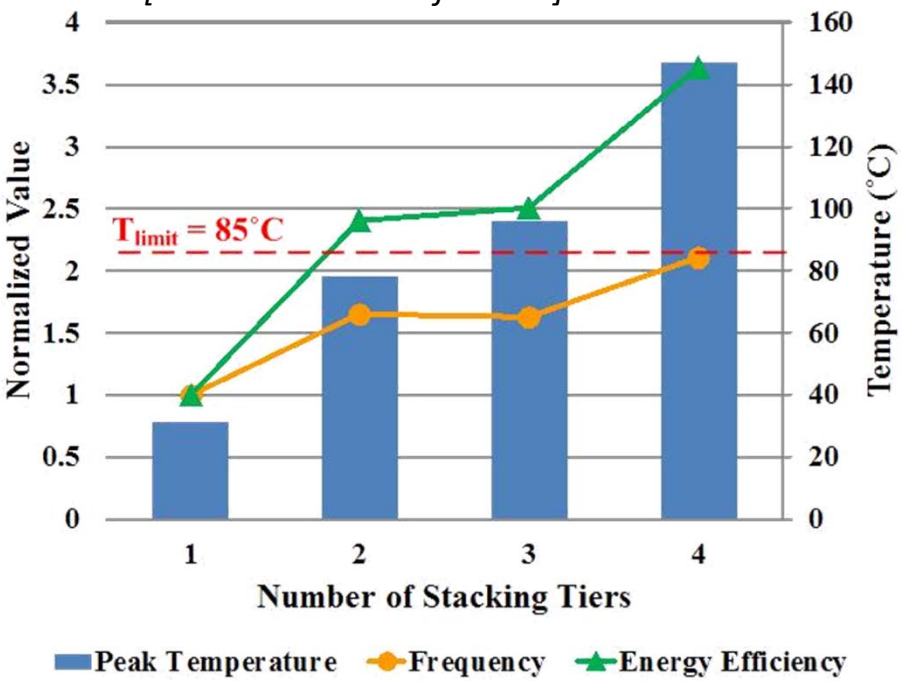


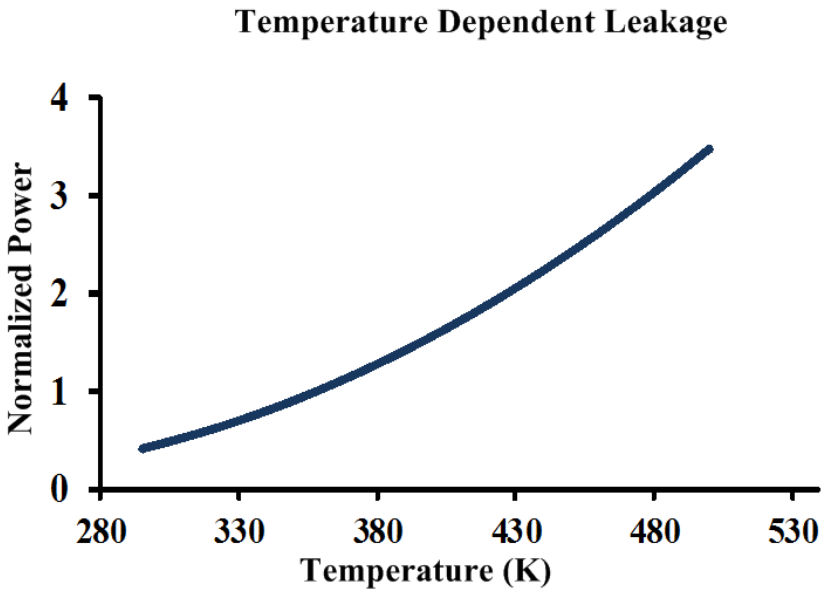
Illustration of Air Cooling

3D FGPAs

- **Stacking Multiple FPGA Tiers**
  - ❑ Benefits
    - Reduce interconnect length
    - Improve performance and efficiency
  - ❑ Thermal Challenges
    - Thermal violation (Peak Temperature > T_{limit})
    - Degrade performance and energy efficiency
    - **"Thermal Runaway"**

*[Simulation results for **alu4**]*

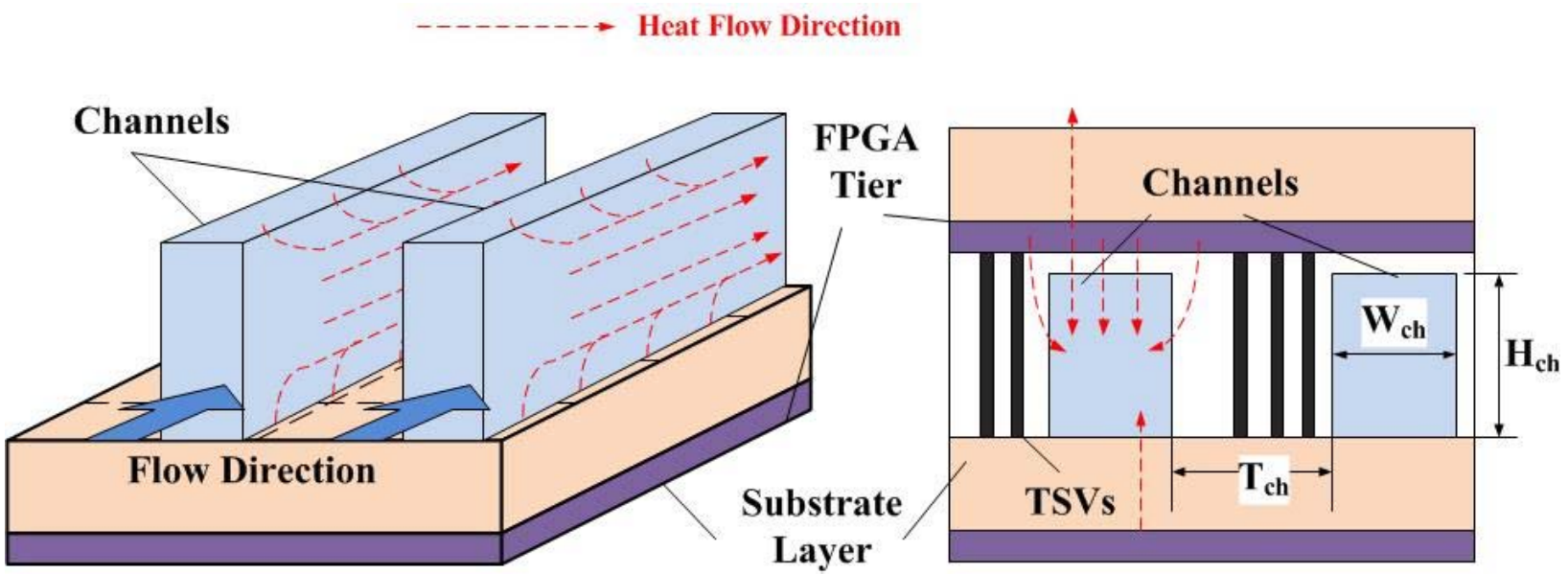




Micro-Fluidic (MF) Cooling

- **Structure and Benefits**

Illustration of Micro-channel-based Fluidic Cooling



Micro-Fluidic (MF) Cooling

- **Structure and Benefits**
  - Improvement: Shorten heat transfer path

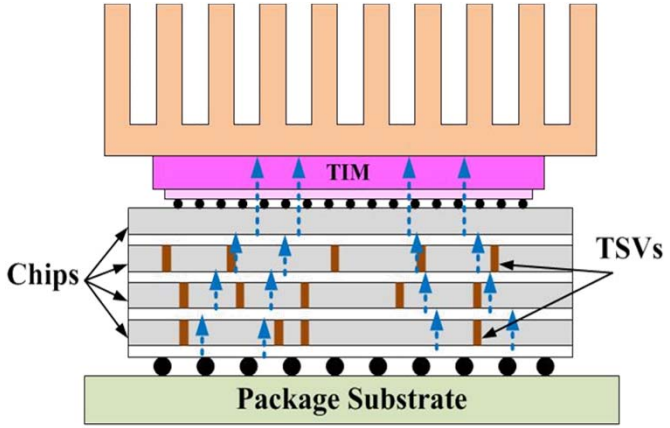
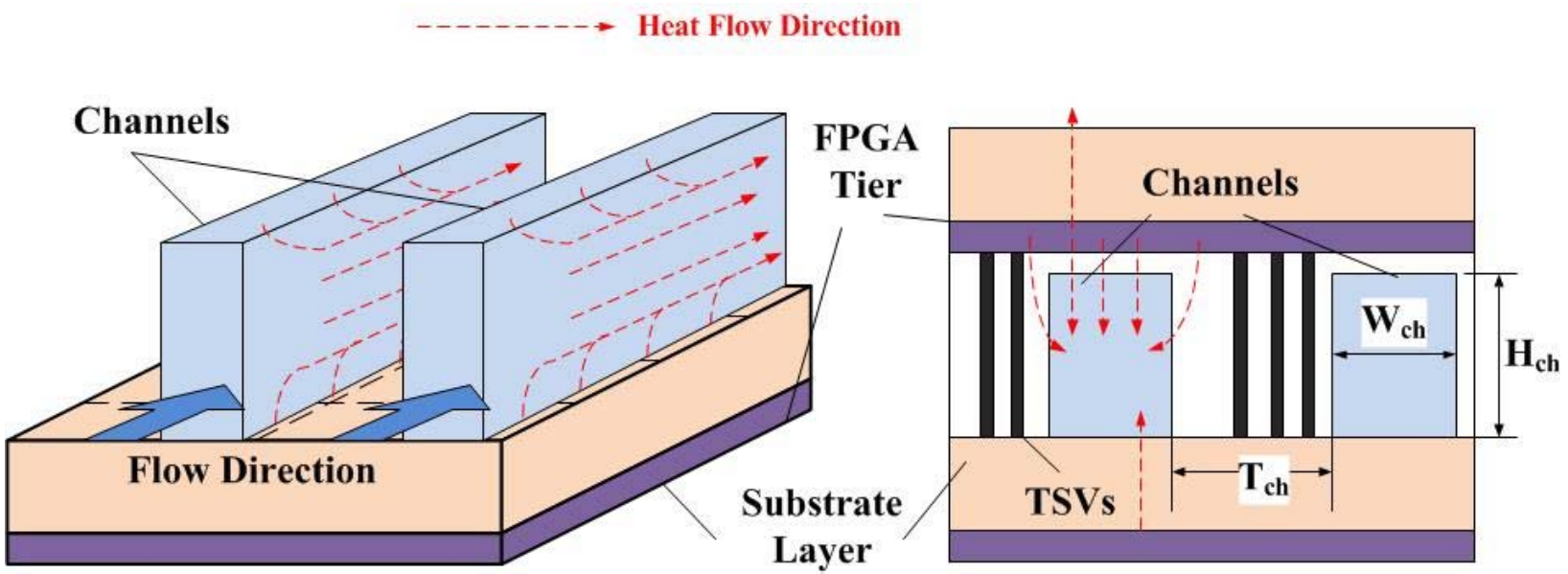


Illustration of Micro-channel-based Fluidic Cooling



Micro-Fluidic (MF) Cooling

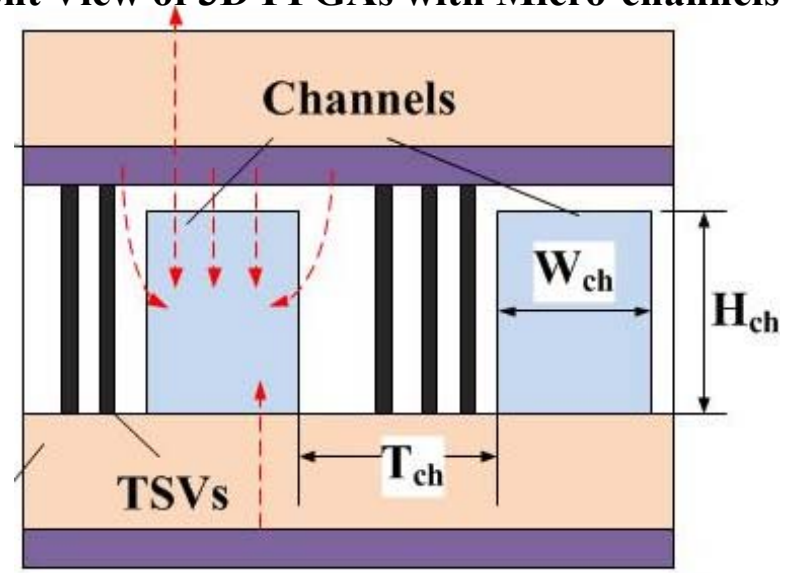
➢ **Design Challenges**

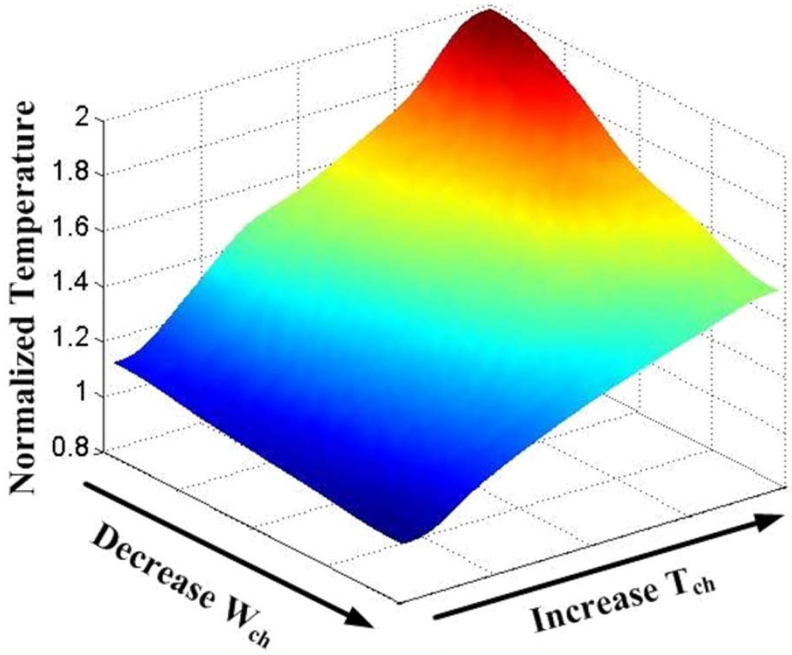
❑ Conflict between TSVs and Micro-channels

- 3D Bandwidth ⇧ T_{ch} ⇧ Cooling Performance ⇩

(assume the size and density of TSVs are fixed)

Front View of 3D FPGAs with Micro-channels



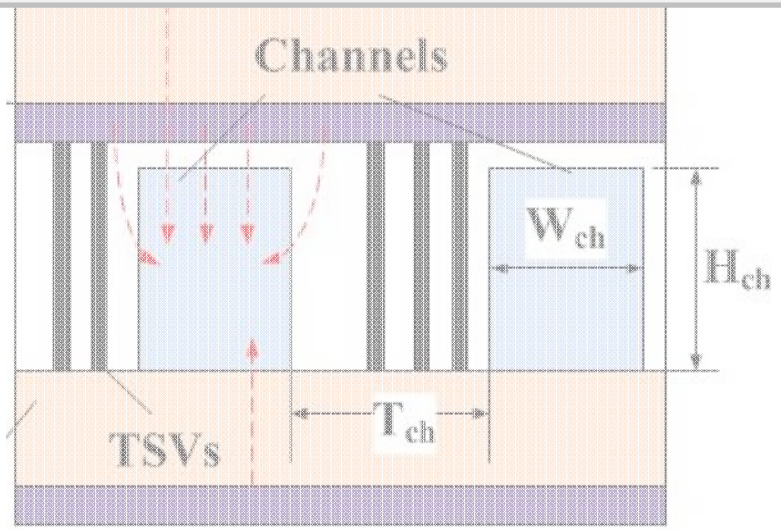


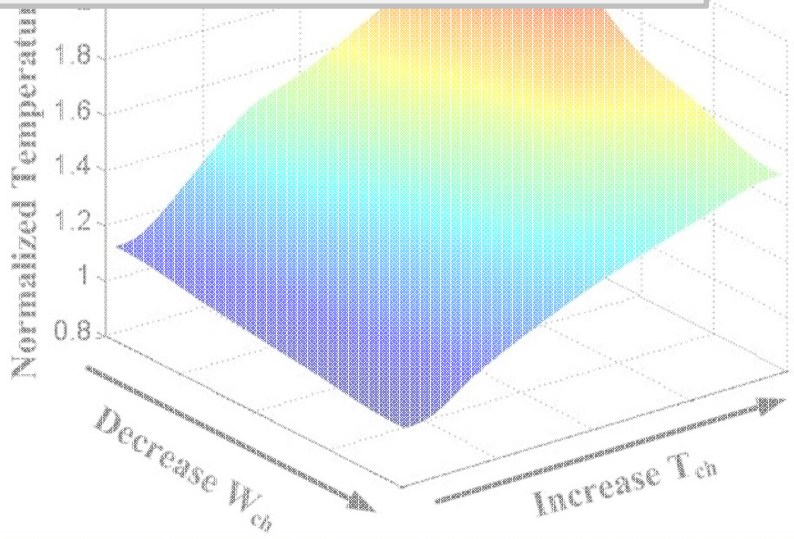
Micro-Fluidic (MF) Cooling

- **Design Challenges**
  - Conflict between TSVs and Micro-channels
    - 3D Bandwidth ⇧ T_{ch} ⇧ Cooling Performance ⇩

(assume the size and density of TSVs are fixed)

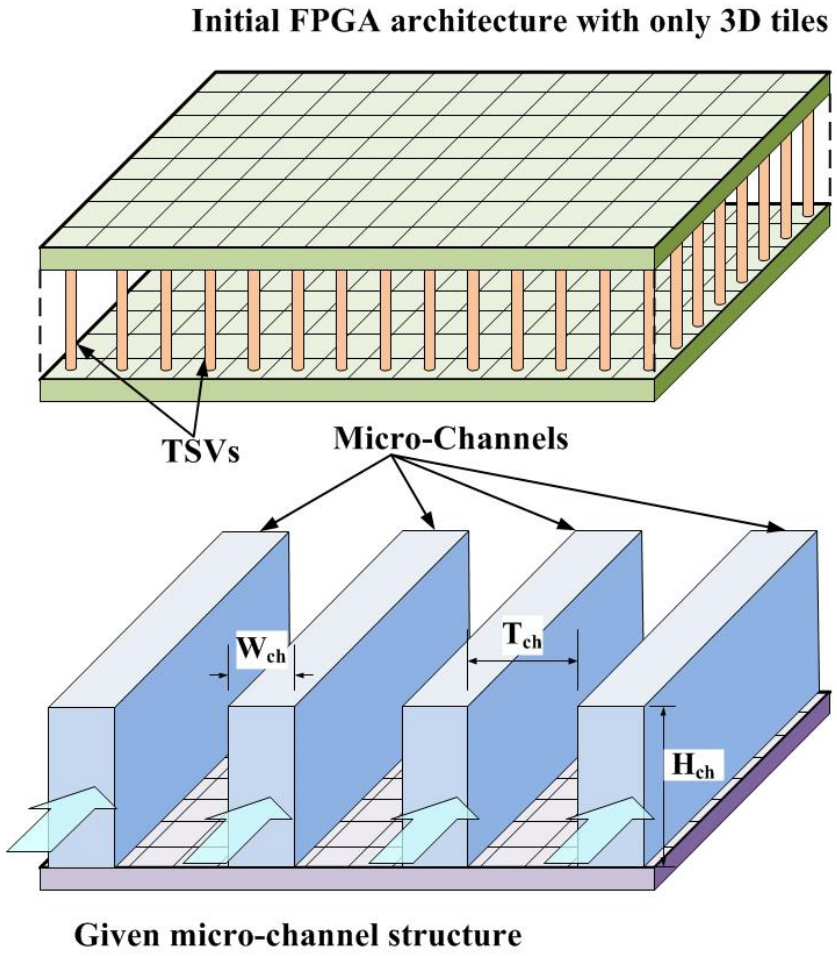
Objective:
Simultaneous exploration of the configuration of micro-channels and the design of 3D FPGAs





Design Exploration of MF Cooled 3D FPGAs

- **Match micro-channel configuration with 3D FPGA architecture**
  - Conflict between TSVs and Micro-channels



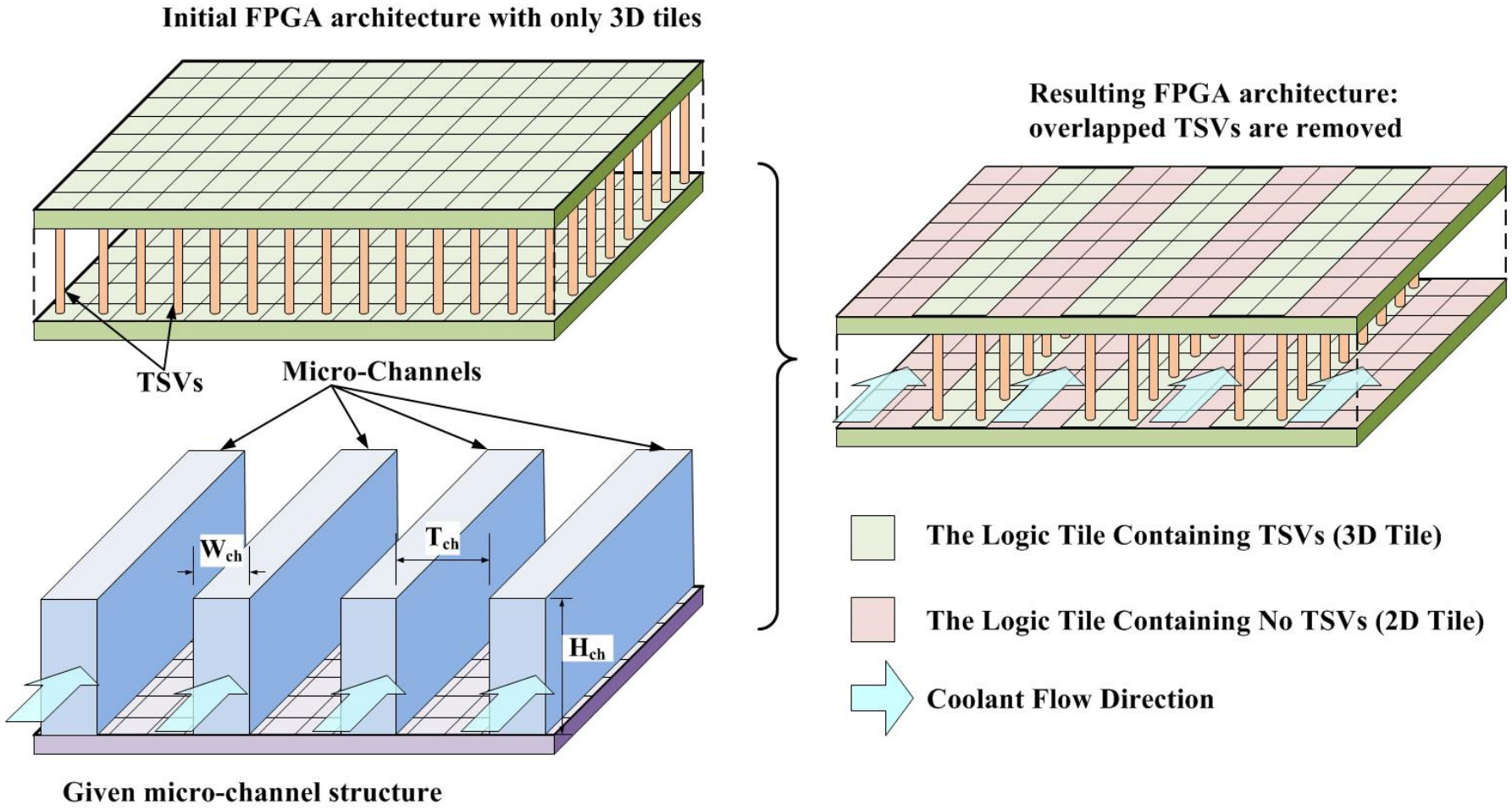
The Logic Tile Containing TSVs (3D Tile)

The Logic Tile Containing No TSVs (2D Tile)

Coolant Flow Direction

Design Exploration of MF Cooled 3D FPGAs

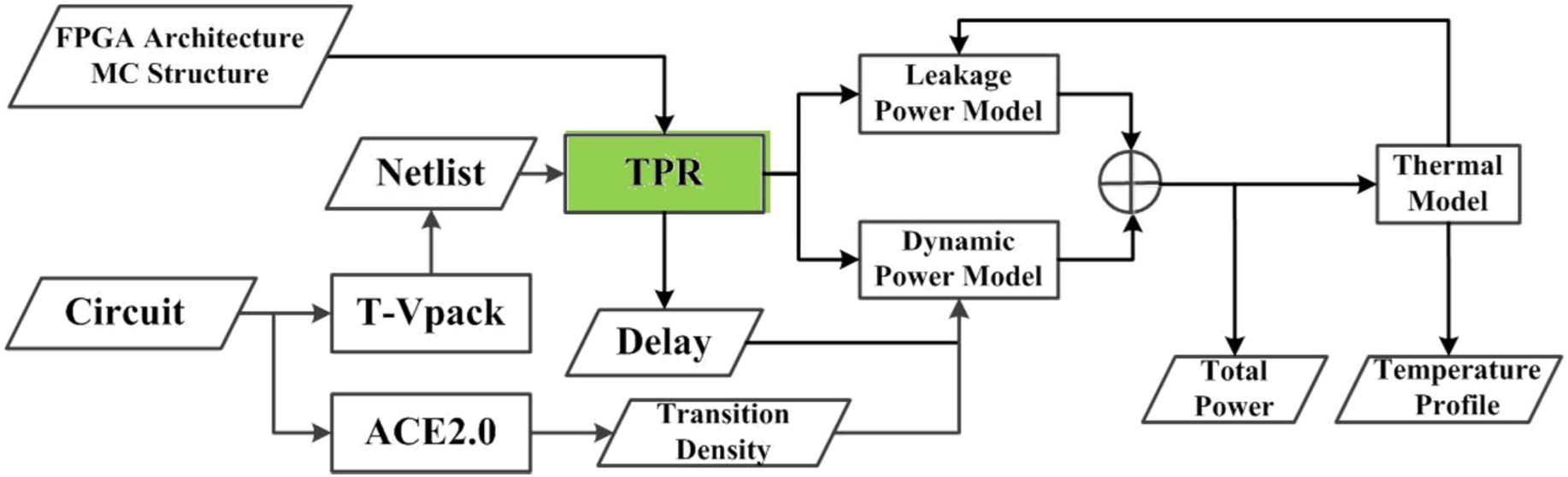
- **Match micro-channel configuration with 3D FPGA architecture**
  - Conflict between TSVs and Micro-channels



Design Exploration of MF Cooled 3D FPGAs

➢ Design Exploration Framework

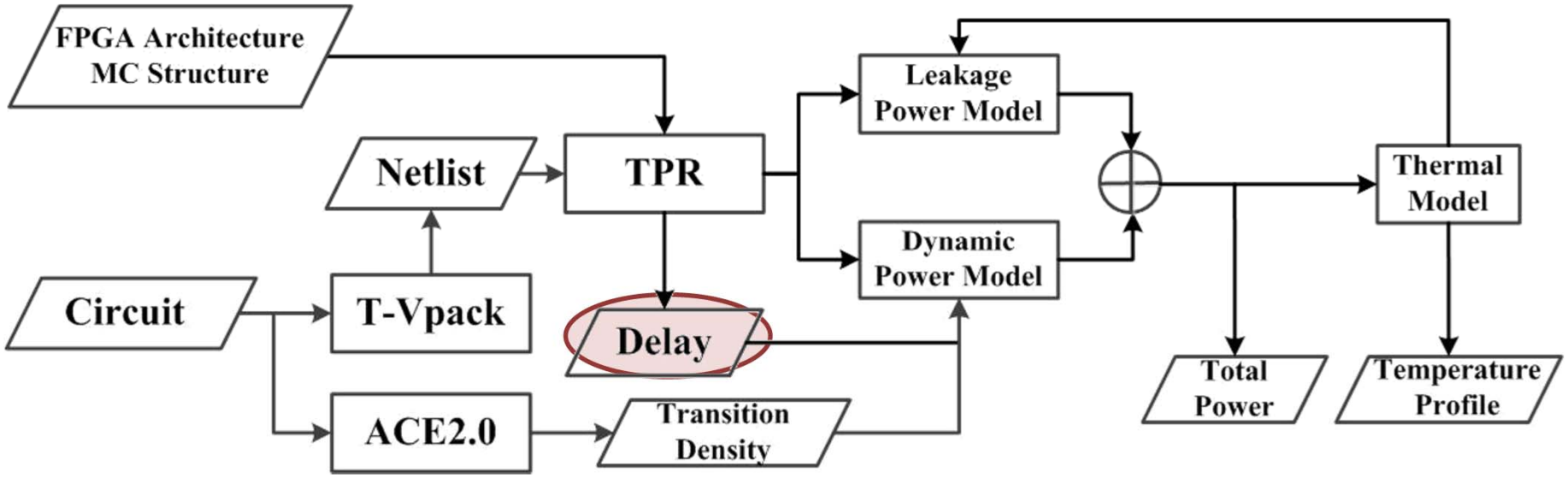
- The framework is based on TPR (modified to support multiple types of logic tiles)



Design Exploration of MF Cooled 3D FPGAs

➢ Design Exploration Framework

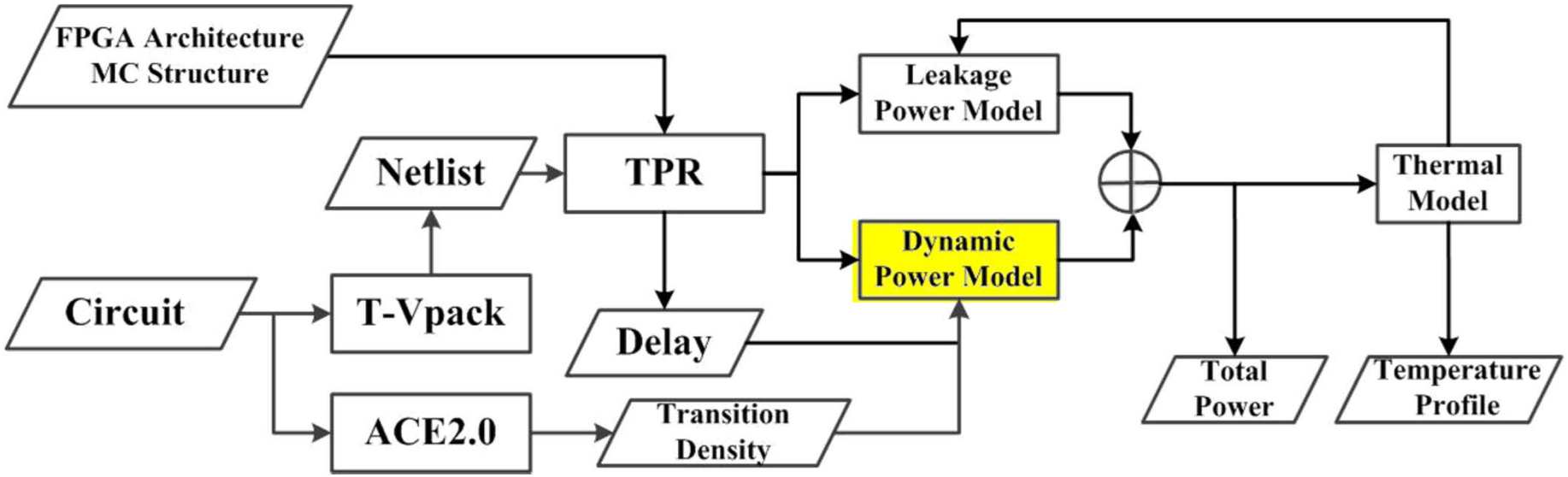
- The framework is based on TPR (modified to support multiple types of logic tiles)
- Delay determines the maximum operating frequency



Design Exploration of MF Cooled 3D FPGAs

➢ Design Exploration Framework

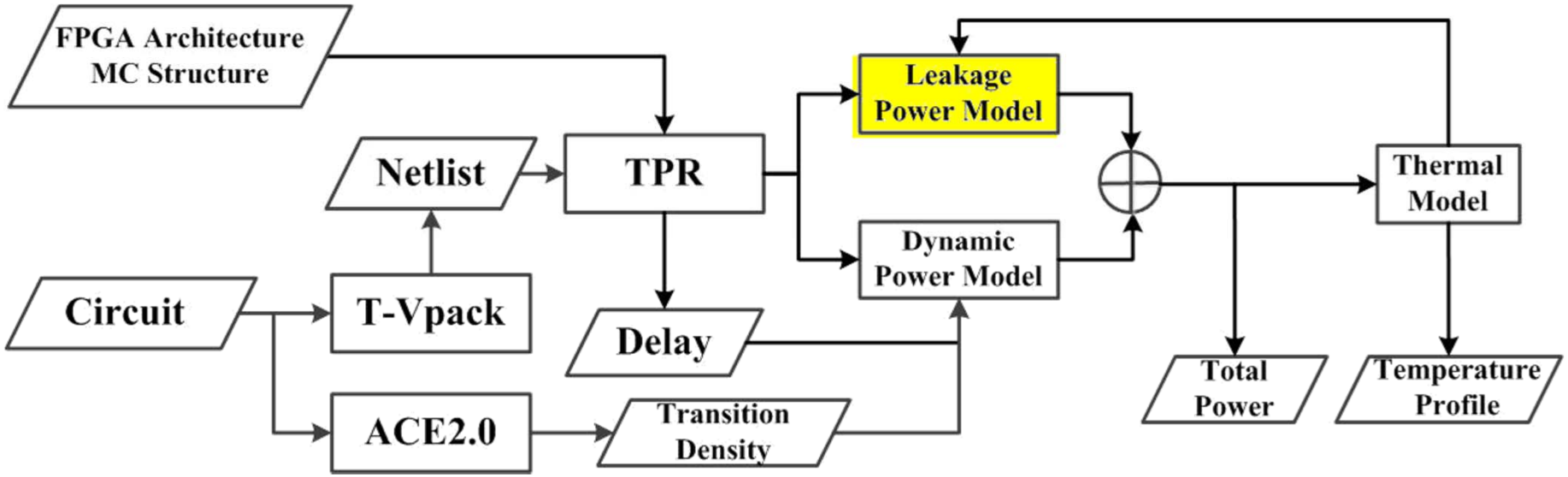
- The framework is based on TPR (modified to support multiple types of logic tiles)
- Delay determines the maximum operating frequency
- Dynamic power: $P = \alpha C V^2 f$



Design Exploration of MF Cooled 3D FPGAs

➢ Design Exploration Framework

- The framework is based on TPR (modified to support multiple types of logic tiles)
- Delay determines the maximum operating frequency
- Dynamic power: $P = \alpha C V^2 f$
- Leakage Power:



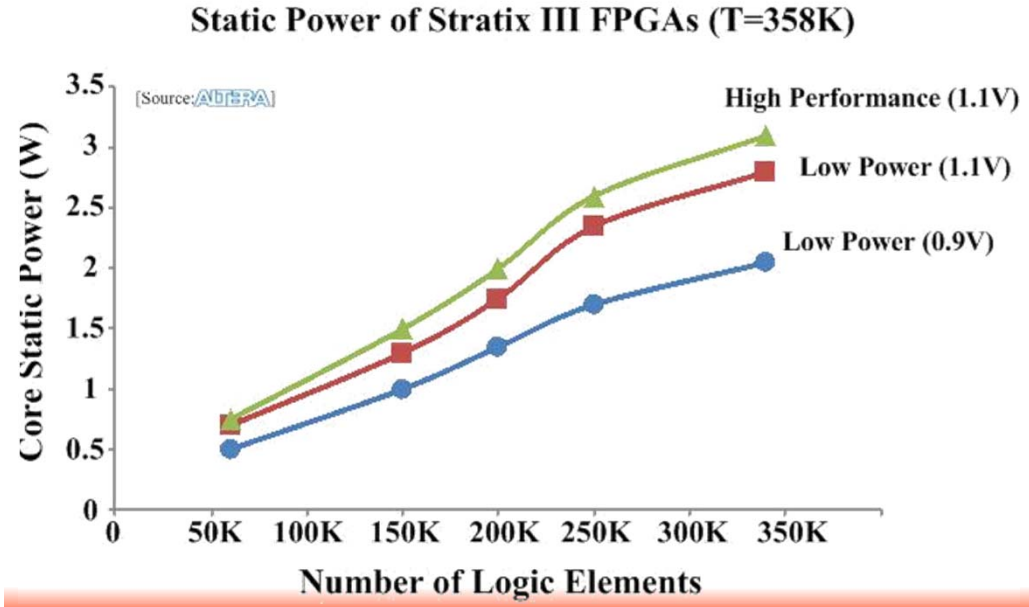
Design Exploration of MF Cooled 3D FPGAs

➢ Design Exploration Framework

- The framework is based on TPR (modified to support multiple types of logic tiles)
- Delay determines the maximum operating frequency
- Dynamic power: $P = \alpha C V^2 f$
- Leakage Power:

❑ Leakage Power

Step 1: Average leakage power at nominal temperature



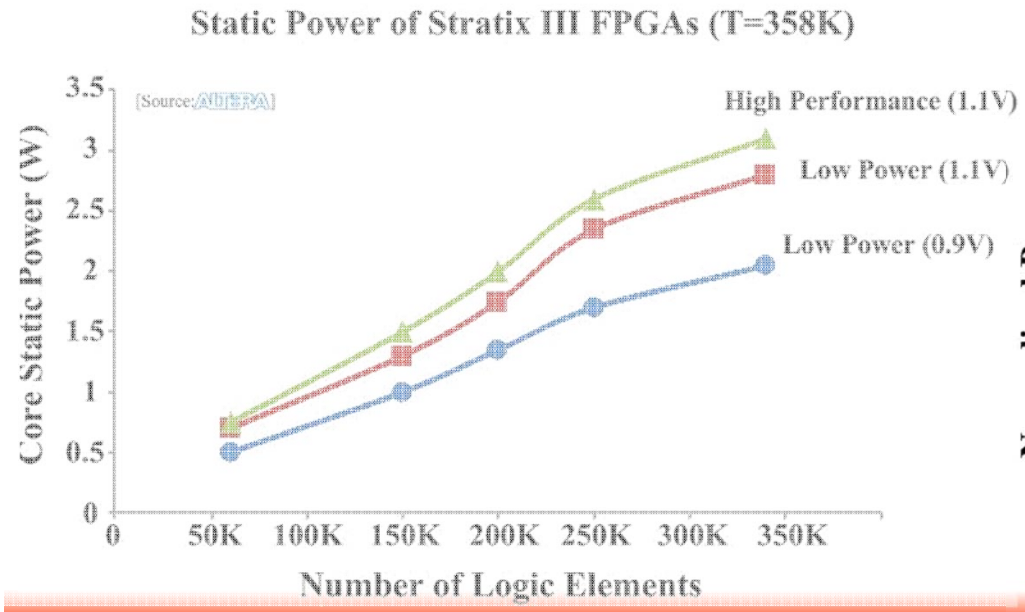
Design Exploration of MF Cooled 3D FPGAs

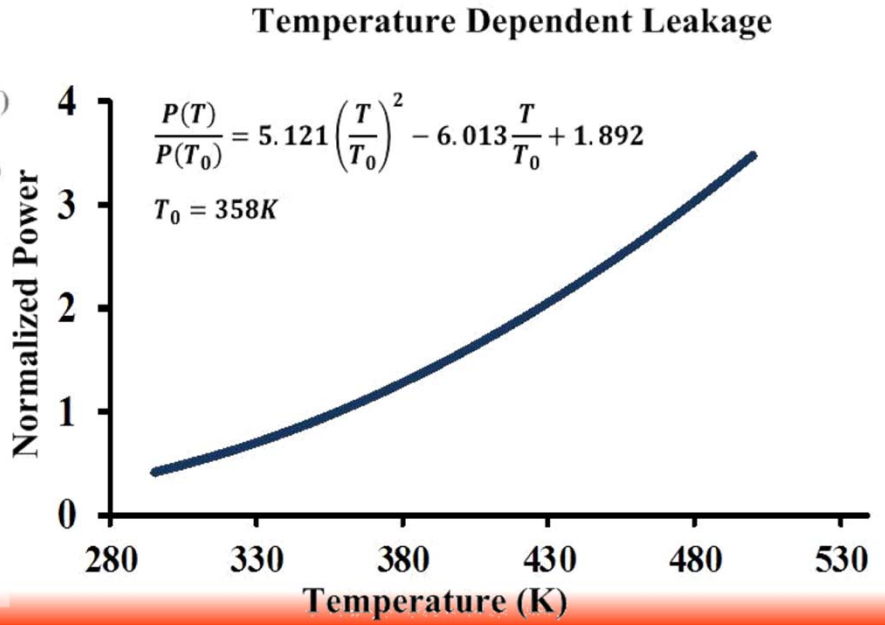
➢ Design Exploration Framework

- The framework is based on TPR (modified to support multiple types of logic tiles)
- Delay determines the maximum operating frequency
- Dynamic power: $P = \alpha C V^2 f$
- Leakage Power:

❑ Leakage Power

Step 2: Leakage power look-up-table for the whole range of temperature

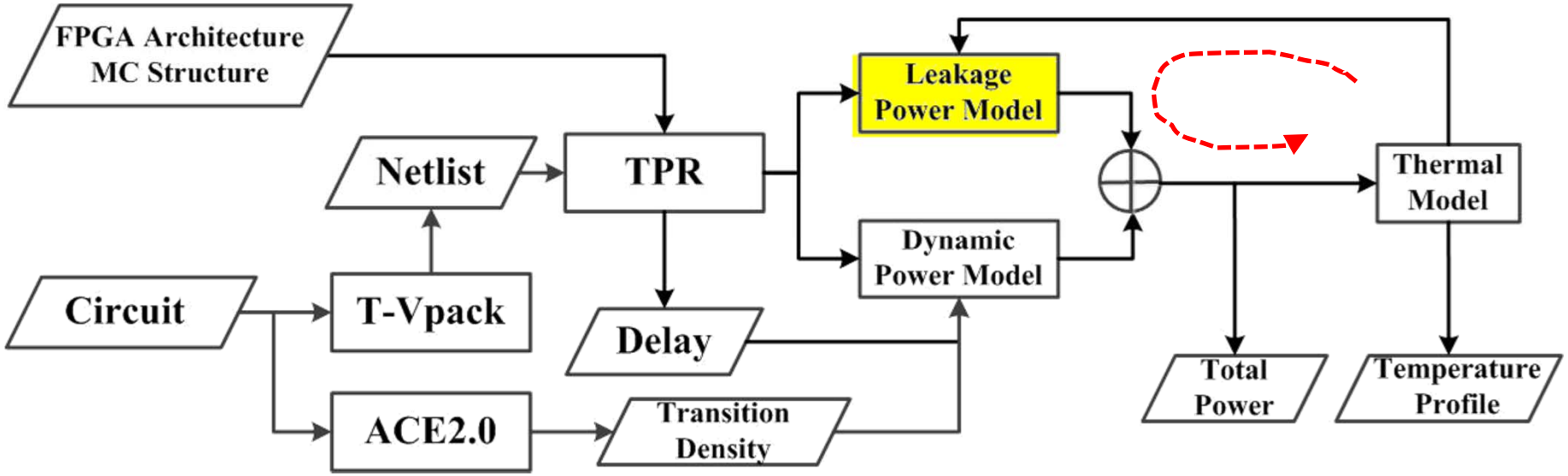




Design Exploration of MF Cooled 3D FPGAs

➢ Design Exploration Framework

- The framework is based on TPR (modified to support multiple types of logic tiles)
- Delay determines the maximum operating frequency
- Dynamic power: $P = \alpha C V^2 f$
- Leakage Power: **Thermo-Power Loop**

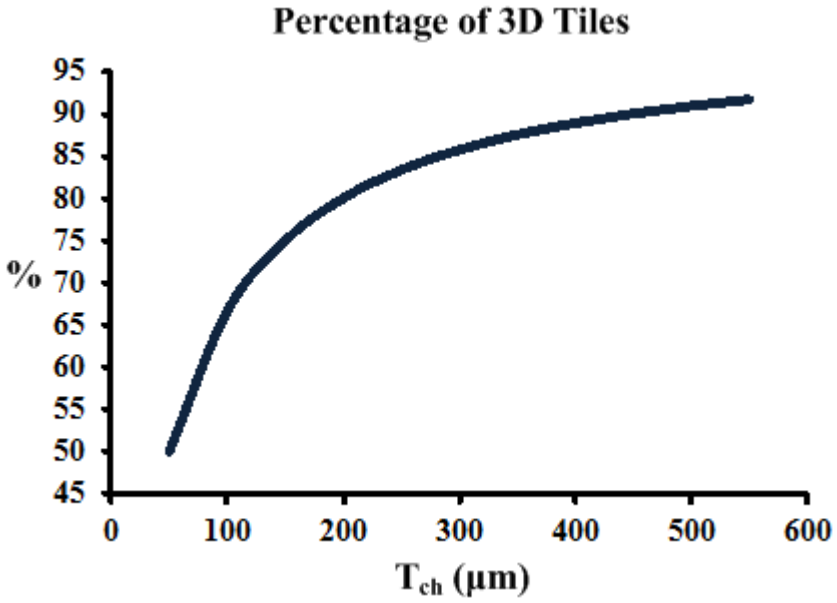


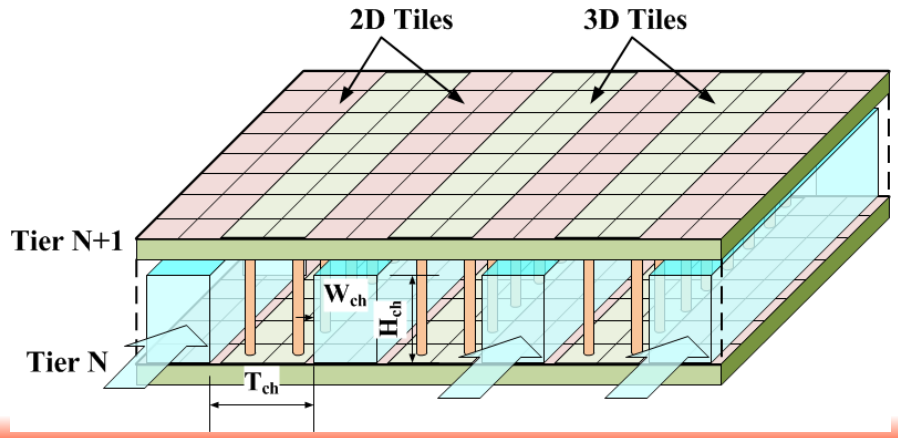
Experiment and Results

➢ Experiment Setup

Variables	Value
Layer # of 3D FPGA	2 ~ 4
T_{ch}	50 μm ~ 550 μm

Parameters	Value
Fluid Velocity	1 $m\ s^{-1}$
W_{ch}	50 μm
H_{ch}	100 μm
Ambient Temperature	298 K
TSV Diameter	10 μm
TSV Height	150 μm
TSV # per 3D Tile	4
Horizontal Routing-Channel Width	30



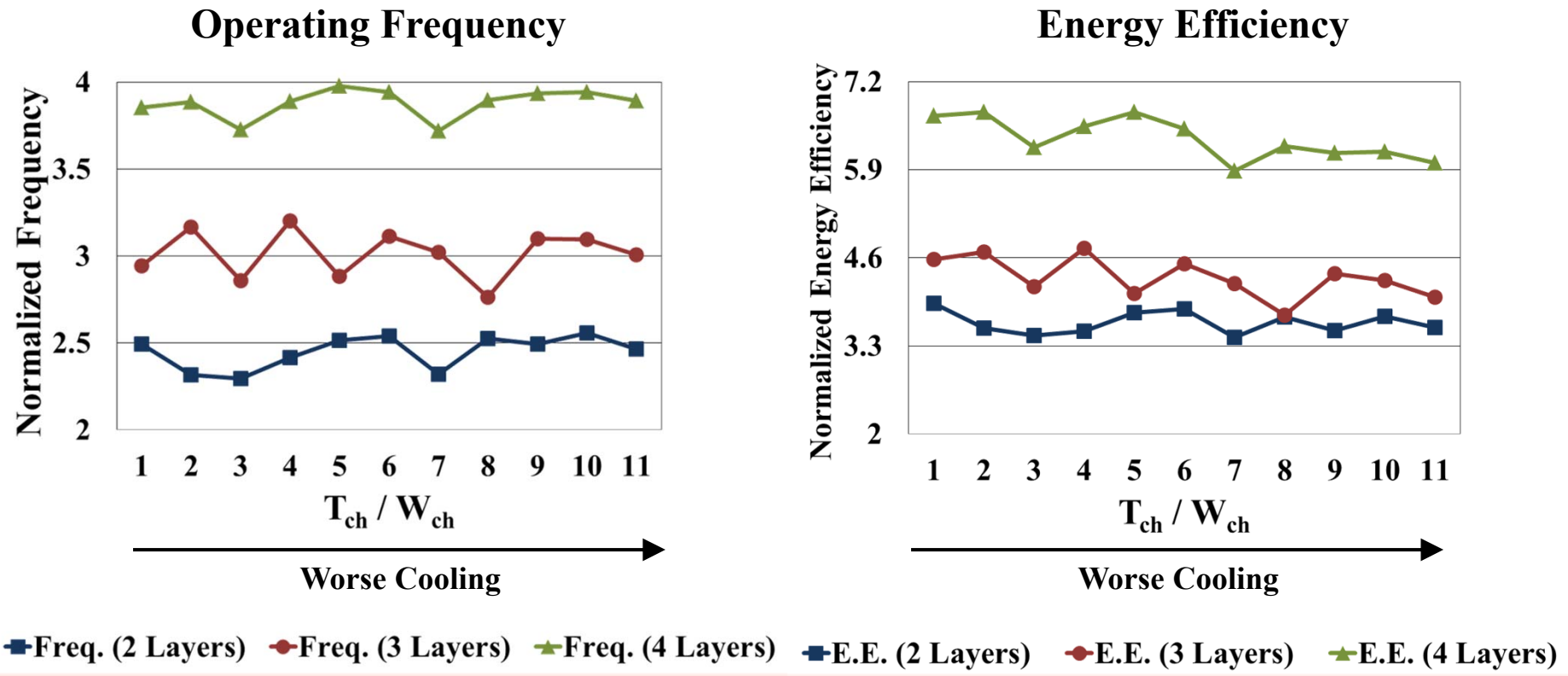


Experiment and Results

➢ **Results 1**

- More layers, better frequency and energy efficiency
- Nonmonotonic relationship between frequency and T_{ch}/W_{ch}

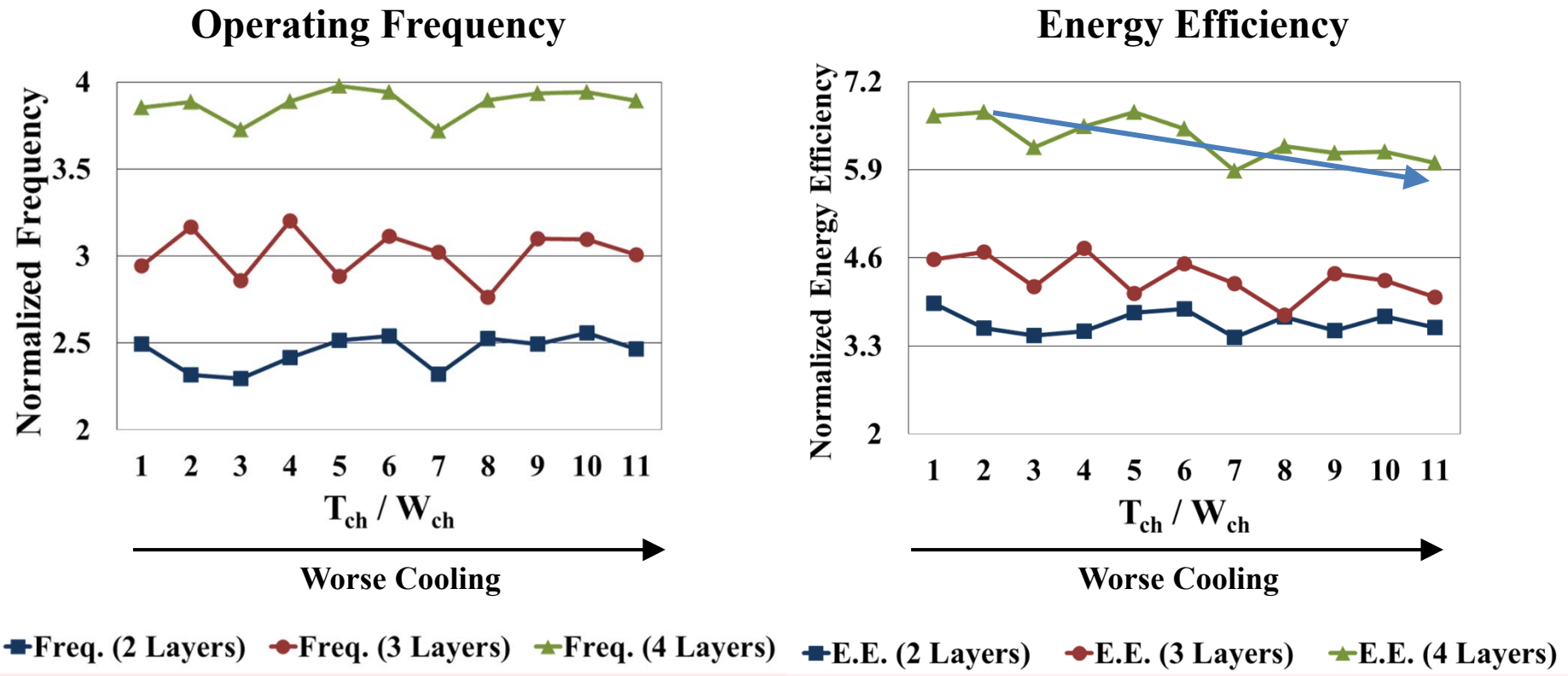
$$Energy\ Efficiency = \frac{Frequency^2}{Power}$$



Experiment and Results

➢ **Results 1**

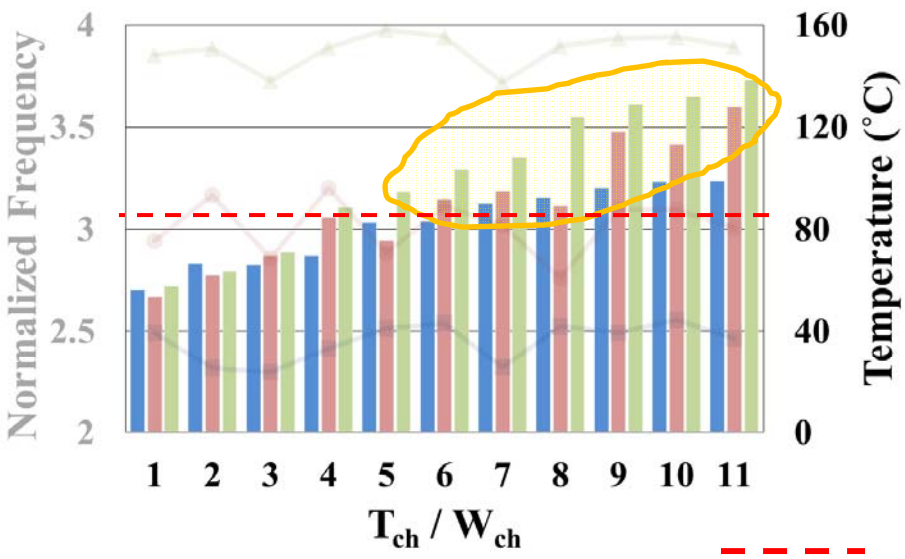
- More layers, better frequency and energy efficiency
- Nonmonotonic relationship between frequency and T_{ch}/W_{ch}
- Reduction trend of energy efficiency with the increase of T_{ch}/W_{ch}



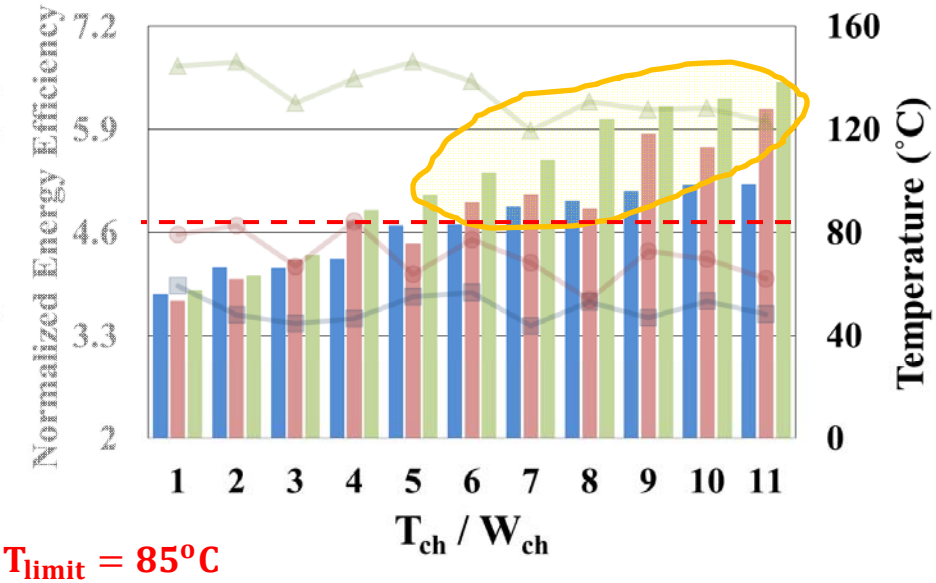
Experiment and Results

- **Results 1**
  - More layers, better frequency and energy efficiency
  - Nonmonotonic relationship between frequency and T_{ch}/W_{ch}
  - Reduction of energy efficiency with the increase of T_{ch}/W_{ch}
  - Thermal violation when T_{ch}/W_{ch} is large

Operating Frequency



Energy Efficiency

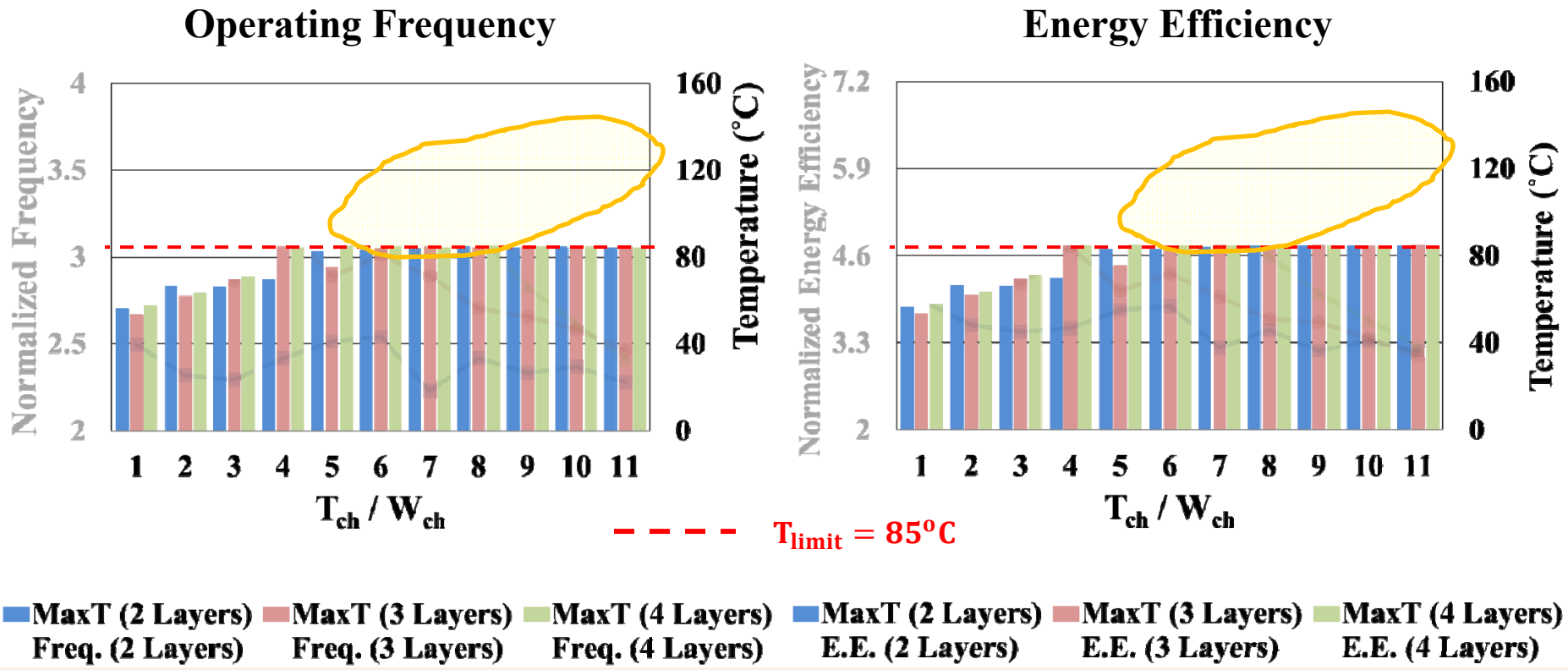


$T_{limit} = 85°C$

MaxT (2 Layers) MaxT (3 Layers) MaxT (4 Layers) MaxT (2 Layers) MaxT (3 Layers) MaxT (4 Layers)
Freq. (2 Layers) Freq. (3 Layers) Freq. (4 Layers) E.E. (2 Layers) E.E. (3 Layers) E.E. (4 Layers)

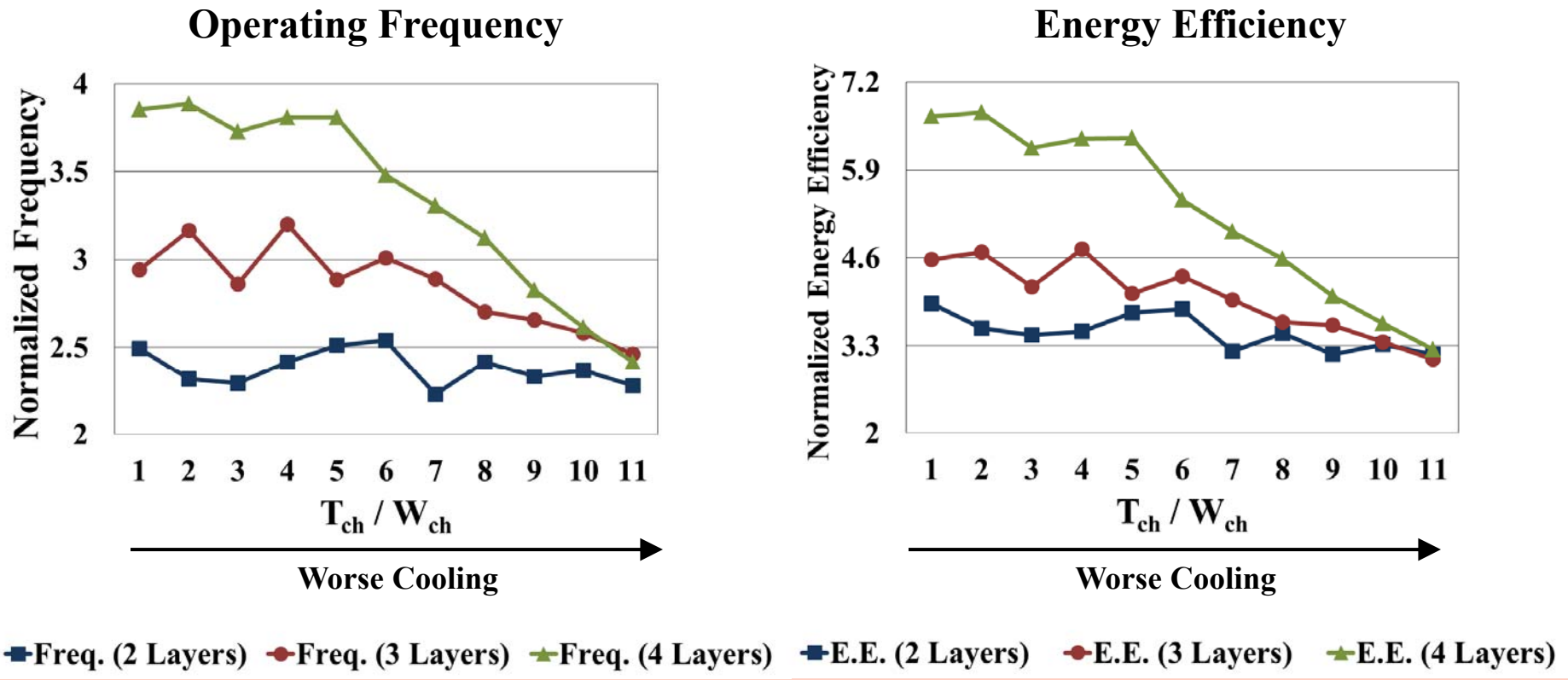
Experiment and Results

- **Results 2**
  - Fix thermal problem by reducing operating frequency



Experiment and Results

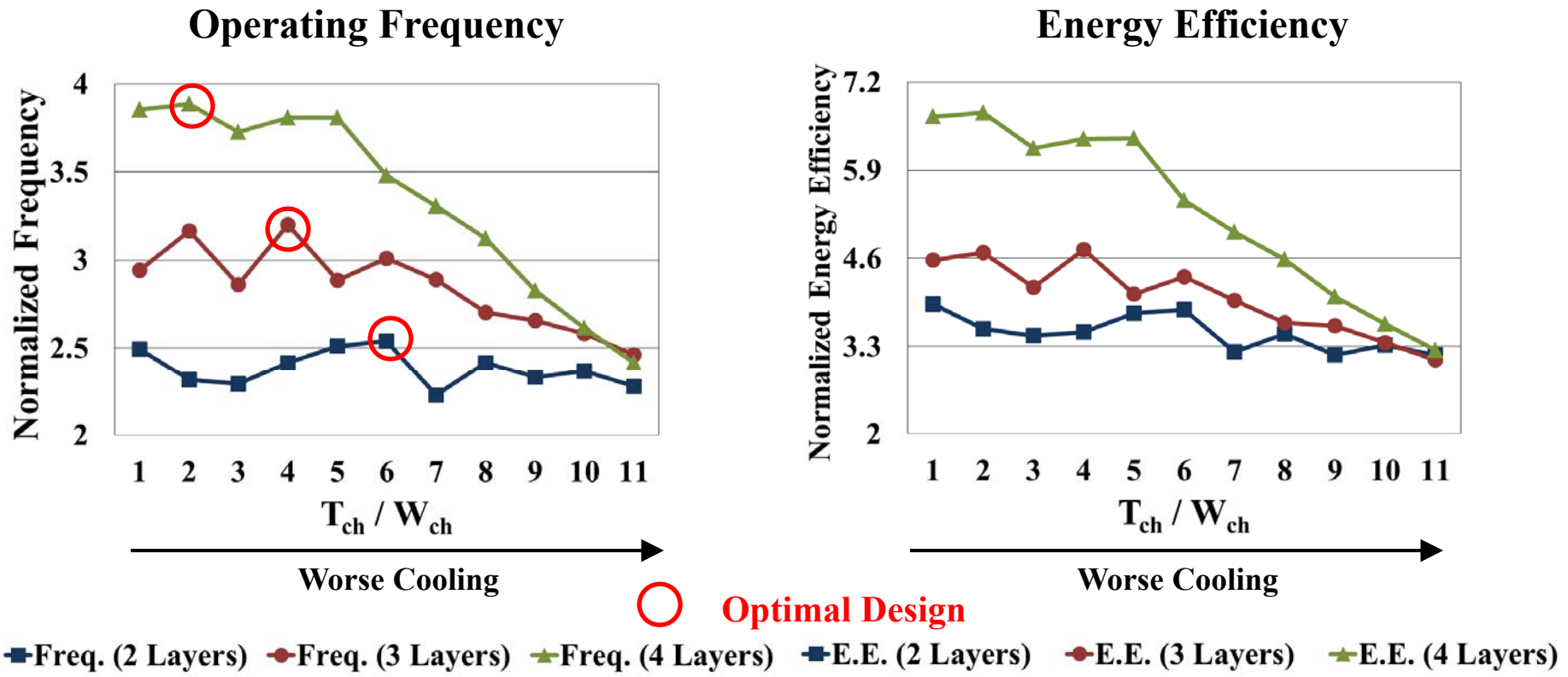
- **Results 2**
  - Performance and Energy Efficiency degrade when T_{ch}/W_{ch} is large



Experiment and Results

➢ **Results 2**

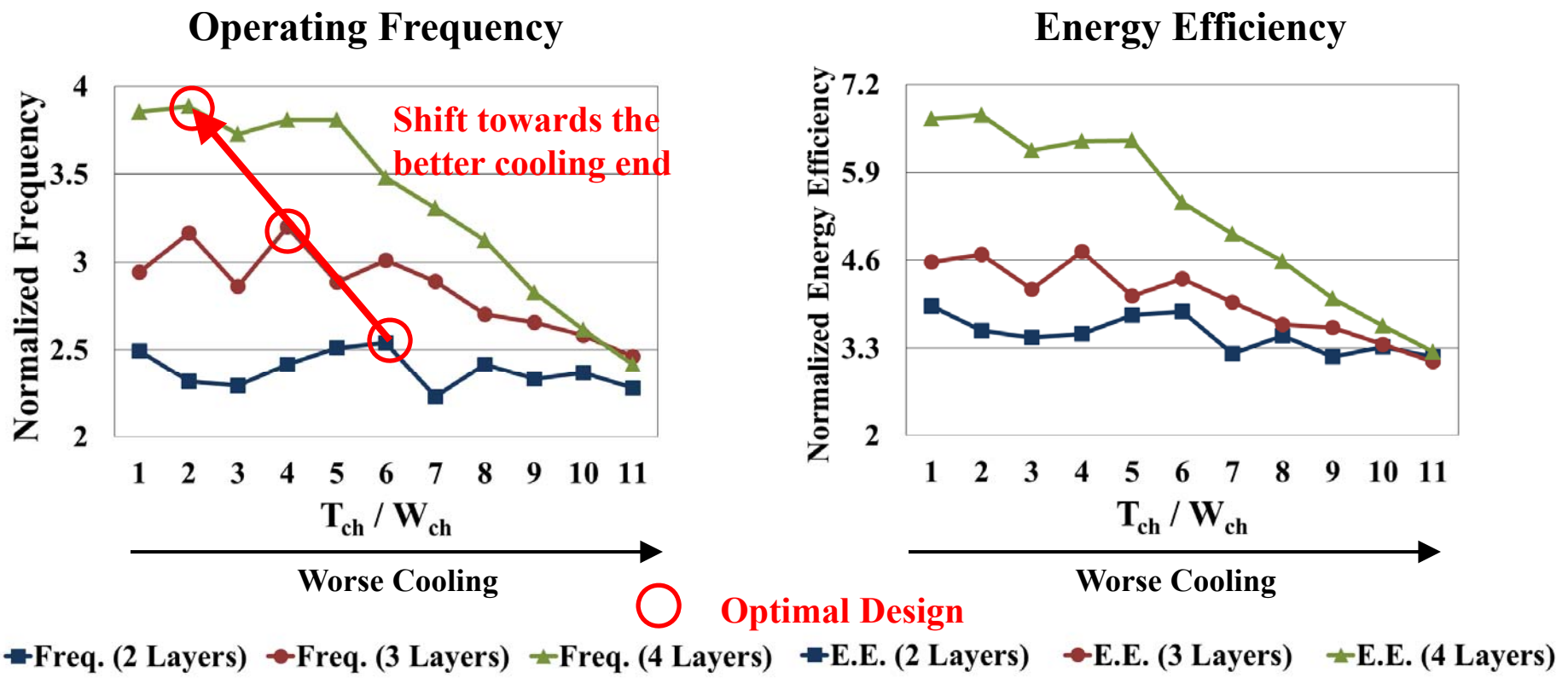
- Performance and Energy Efficiency degrade when T_{ch}/W_{ch} is large
- Neither the highest 3D bandwidth or the best cooling design is optimal
- Optimal T_{ch}/W_{ch} varies for different number of layers



Experiment and Results

➢ **Results 2**

- Performance and Energy Efficiency degrade when T_{ch}/W_{ch} is large
- Neither the highest 3D bandwidth or the best cooling design is optimal
- Optimal T_{ch}/W_{ch} varies for different number of layers



Conclusion

- 3D FPGAs enjoy several improvements while suffering thermal problems. Micro-channel cooling is a potential solution to the thermal problems.
- Micro-channel density impacts the cooling efficiency and performance of 3D FPGAs.
- We propose a framework to explore the design of 3D FPGAs with micro-channel cooling.
- Neither the design to achieve highest 3D bandwidth or highest cooling performance is optimal. Optimal MC structure differs for different number of layers.

Acknowledgement

- NSF Grant CCF1302375

- DARPA ICECOOL

Question

Thank You!