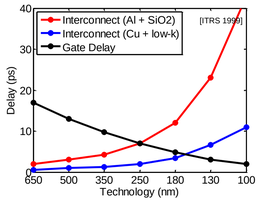
3D IC Architectural-Physical Co-Design

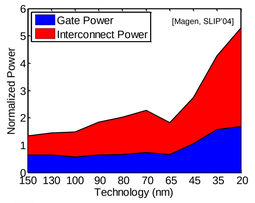
Caleb Serafy, Tiantao Lu and Zhiyuan Yang
Advisor: Ankur Srivastava

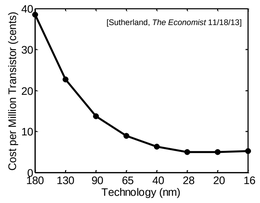
THIRTY YEARS OF SYSTEMS RESEARCH EXCELLENCE

INSTITUTE FOR SYSTEMS RESEARCH
A. JAMES CLARK SCHOOL OF ENGINEERING

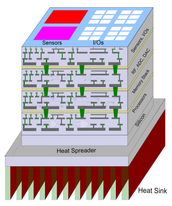
Background and Motivation



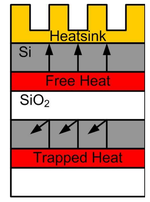




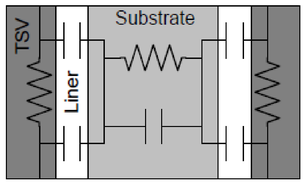
- As transistor shrinks in 2D-IC:
  - Interconnect delay/power becomes dominating factor
  - Cost scaling flattens
- 3D-IC expands planar circuits into 3D space
  - Shorter wirelengths are faster and consume less power
  - Transistor stacking increases density without scaling
- Increased power density creates thermal challenges
- TSV Challenges: Coupling, Stress, Electromigration



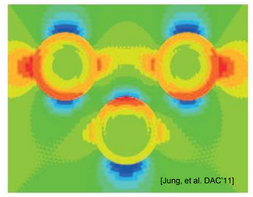
3D-IC



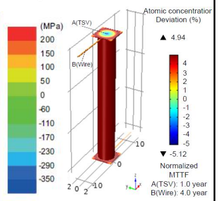
Trapped Heat Effect



TSV-TSV Coupling

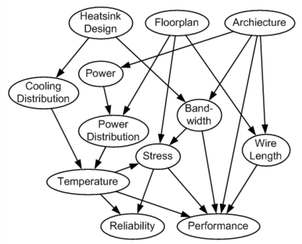


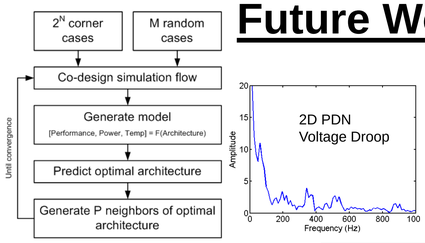
Thermo-Mechanical Stress

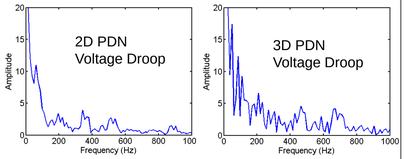


Electromigration

Future Work







Co-design Graph Model
- Multiple design variables interact through a cause and effect graph
- Modeling of all interactions (edges) can be overly complex without adding significant accuracy
- Assign a weight to each edge to represent the strength of the interaction
- Trade off model complexity and accuracy by changing parameter W
  - Only model edges with weight > W

3D CPU Design Space Exploration
- Past work has exhaustively simulated architectural design space
  - This computationally limits the scope of the design space
- Design space exploration (DSE) algorithm required to find optimal design without exhaustive simulation
- Requires model of metrics vs design
- Increase accuracy of model as exploration occurs

3D IC Power Delivery Network (PDN)
- Past work has shown 3D CPUs with MF cooling can increase performance at the expense of increased transistor power
- This can lead to increased frequency and amplitude of VDD noise and IR drop
- Vertical power routing changes PDN design requirements
- PDN constraints could limit potential of 3D CPUs
- PDN constraints and optimization must be integrated into DSE framework

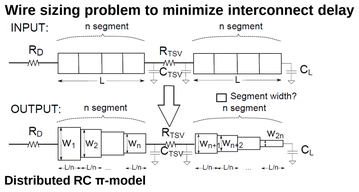
3D FPGA with Micro-Fluidic Cooling
- Past work has shown the peak temperature of 3D FPGA increases with the number of layers
- This necessitates the use of more aggressive cooling such as micro-fluidic cooling
- Heat sink structure impacts the arrangement of 3D switch boxes thereby influencing the placement and routing performance
- Different from ASIC, routing of FPGA determines connection boxes which has great impact on power distribution thereby affecting the temperature
- Micro-fluidic cooling should be integrated in the placement and routing algorithm of 3D FPGA

3D IC Reliability Study
- Past work has focused on TSV-induced failure in 3D-IC, and has shown the co-design scheme can prolong the interconnect lifetime without performance overhead.
- Gate's failure (HCI, TDDB, NBTI) is also crucial for the system
- Call for co-design at different levels of abstraction (device, circuit, architectural level)

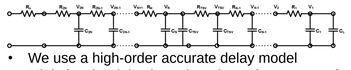
Physical Level Design of 3D-ICs

Designing for Interconnect Performance

Wire sizing problem to minimize interconnect delay



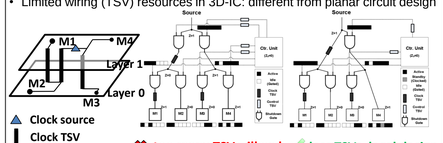
Distributed RC π-model



- We use a high-order accurate delay model
- Minimize the delay based on dynamic programming

Designing for Low-Power

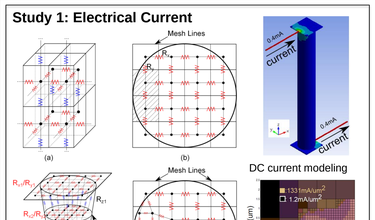
- Clock tree power accounts for up to 70% of the circuit total dynamic power
- Shutdown gates periodically turns off idle sequential logic thus saves power
- These gates need special control logic (wiring overhead)
- Limited wiring (TSV) resources in 3D-IC: different from planar circuit design



We have developed algorithms to :
- Generate the entire 3D clock tree
- Decide the location of the shutdown gates
- Place clock TSVs and control TSVs within given whitespace

Designing for Reliability

Study 1: Electrical Current

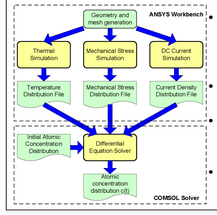


DC current modeling

Adaptive resistive mesh algorithm

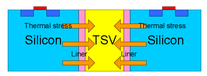
Non-uniform mesh

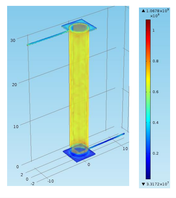
Study 3: Atom Migration



- Atom migration refers to the movement of metal atoms in interconnect
- Result in open/short circuit
- Decided by electrical current, stress, and thermal condition
- Multi-physics time-dependent simulation

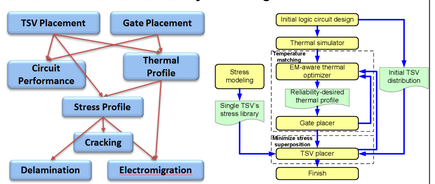
Study 2: Thermal Mechanical Stress





- When cooled, Cu contracts more than Si, forms residual stress
- Stress leads to mechanical failures such as crack and delamination
- Finite-element-method (FEM) thermal-mechanical simulation

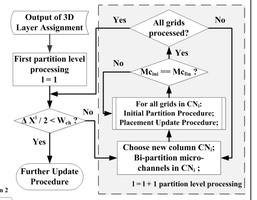
Electrical-Thermal-Reliability Co-Design

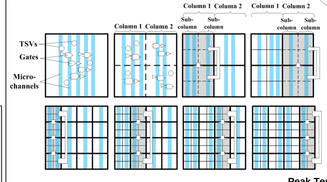


- We propose a unified framework to optimize both 3D-IC's performance and reliability
- Novel reliability models (including electromigration and material fracture) to avoid slow FEM simulations
- Results show we can design more reliable circuit without scarifying performance

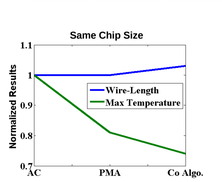
Co-Design with Micro-Fluidic Cooling

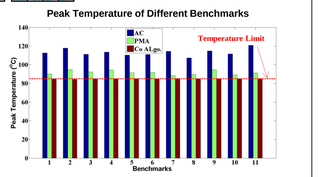
Motivations for Co-Design
- Conflict between TSVs and micro-channel/micropin-fin structure
- Interaction between gates and micro-channels
- Temperature of coolant fluid increases along flow direction: potentially large spatial temperature variation
- Extra power introduced by micro-fluidic cooling

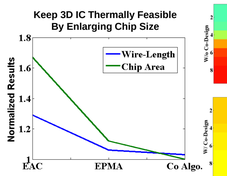


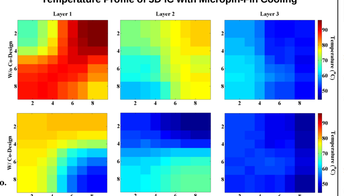


Co-Design Framework
- 2-stage algorithm: (1) Layer-assignment of gates; (2) Hierarchically partitioning based placement
- Objective: provide sufficient cooling for 3D IC with more compact layout



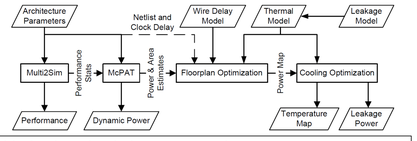




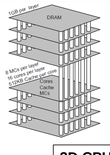


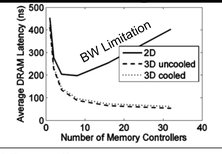
Results an Conclusion
- Air cooing scheme (AC) and the scheme allocate micro-channels after gate and TSV placement (PMA) fail to sufficiently cool 3D IC for compact layout
- Co-design can keep the chip temperature below the limit while maintaining compact layout
- Enlarging chip size (EAC and EPMA) can provide sufficient cooling with larger overhead in chip area and wire-length than co-design

Architectural-Physical Co-design of 3D CPUs

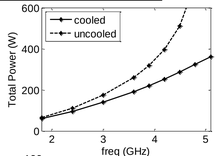


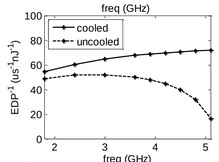
Multi-physics Simulation Framework
- Simulate performance, power and temperature as a function of 3D CPU architecture and software workload
- Built in floorplan and micro-fluidic heatsink optimization loops

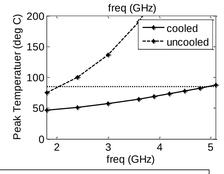




3D CPU Design Space
- Number of CPU cores: stack logic layers
- Number of memory controllers: increase on-chip DRAM bandwidth
- Core Frequency: MF cooling improves feasibility and efficiency of frequency scaling

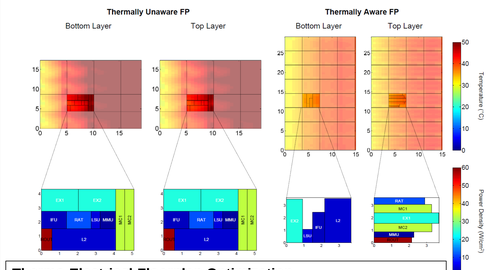




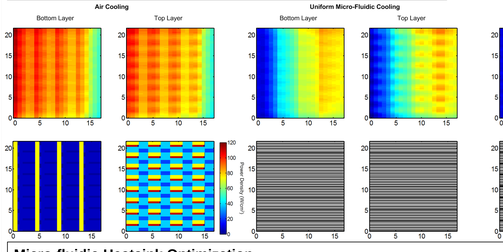


Objective
- Show potential for thermo-electrical co-design at physical and architectural level

Contributions
- Unified multi-physics simulation framework with integrated physical optimization routines

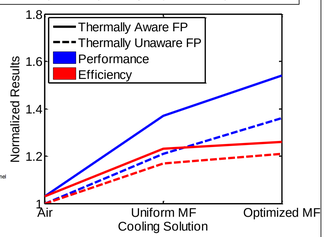


Thermo-Electrical Floorplan Optimization
- Rearranging the power density in horizontal and vertical direction can reduce peak temperatures
- Tradeoff between temperature and timing
  - Spreading blocks reduces hotspots but increases wirelength
- Changing chip aspect ratio can affect channel length and count



Micro-fluidic Heatsink Optimization
- Micro-fluidic heatsink drastically reduces temperature as compared to traditional air cooling
- There is a tradeoff between number of channels and fluid velocity (assuming a constant pressure drop)
- The number and placement of channels can be optimized to yield further thermal improvements

Results and Conclusions
- MF cooling unlocks full potential of 3D CPU
- FP and Heatsink optimizations are necessary to maximize feasible performance and efficiency in 3D CPU
- Significant interactions between heatsink, architecture and floorplan
- Unified co-design approach is necessary in early design space analysis



Optimal Performance/Efficiency
- The optimal performance or efficiency that is thermally and timing feasible across 3D CPU design space
- Both optimizations show significant improvements

Funded by NSF, ISR and DARPA