Minimizing Thermal Gradient and Pumping Power in 3D IC Liquid Cooling Network Design

Gengjie Chen, Jian Kuang, Zhiliang Zeng, Hang Zhang,
Evangeline F. Y. Young, Bei Yu

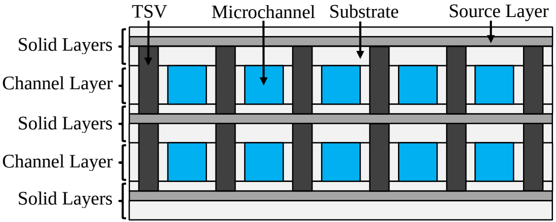
Department of Computer Science & Engineering
The Chinese University of Hong Kong

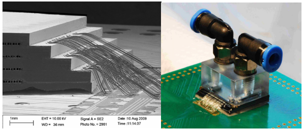
June 21, 2017

香港中文大學
The Chinese University of Hong Kong

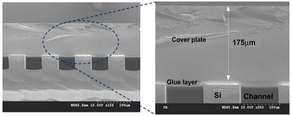
Why 3D IC Liquid Cooling?

- **Power** is the number one problem in chip design
- **3D IC** is promising for increasing computer performance
- But 3D IC **worsens** power problem by
  - higher heat dissipation density
  - larger thermal resistance from junction to ambient
- Microchannel-based liquid cooling is proposed as a solution

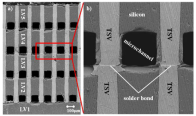




[Brunschwiler+, 3DIC'09]



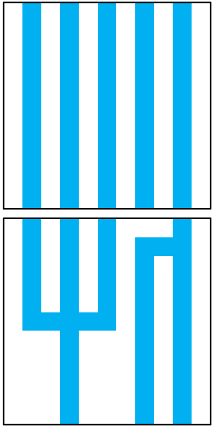
[Dang+, TAP'10]



[Madhour+, ICEPT'12]

Challenges for 3D IC Liquid Cooling

- Hot downstream and cool upstream $\Longrightarrow$ **large thermal gradient** $\Longrightarrow$ reliability and timing issues
- limited channel diameter $\Longrightarrow$ **high pumping requirement** $\Longrightarrow$ overhead to whole system
- Limitation of previous work
  - No considering thermal gradient
  - Assuming unidirectional straight channels
  - Assuming unrealistic constant-temperature heat source

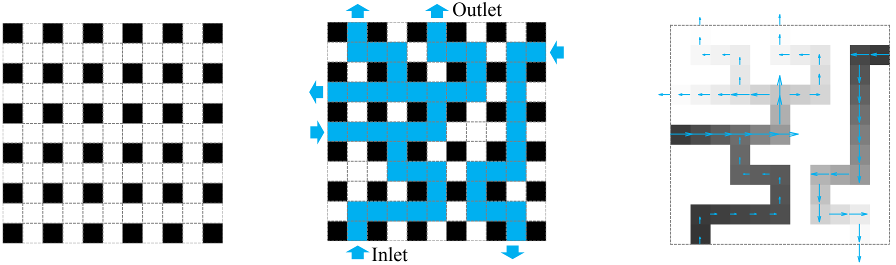


Thermal Modeling Background

- Most existing models assume unidirectional straight channels
- 4-register model (4RM) in 3D-ICE [Sridhar+, TOC'14]
  - Accurate
  - Has been extended for flexible topology
  - **Slow**
- We construct a fast 2-register model (2RM) for cooling network

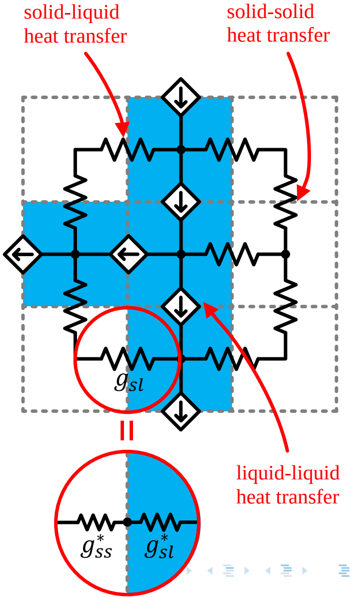
Thermal Modeling Basics

- Divide channel layer into **basic cells** with a 2D grid
- Solve local pressure and flow rate from a **linear system**



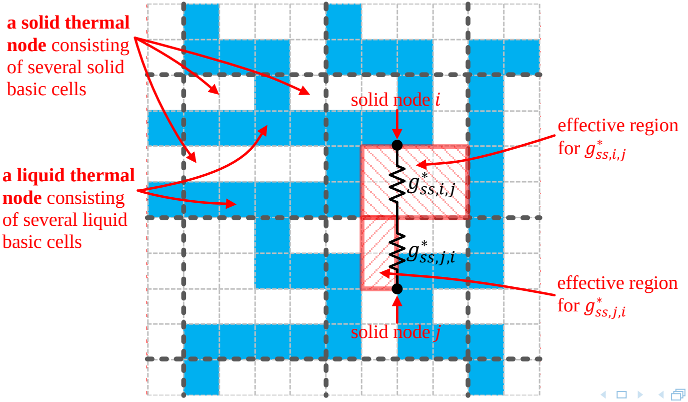
4RM Model

- **Thermal cell** = basic cell
- Solve temperature from a **linear system** considering three kinds of heat transfer
  - Solid-solid
  - Solid-liquid
  - Liquid-liquid



Faster 2RM Model

- No conforming channel geometry $\Longrightarrow$ **larger** and fewer thermal cells $\Longrightarrow$ **speed-up**
- In solid layers, $m \times m$ basic cells $=$ a thermal node
- In channel layers, $m \times m$ basic cells $=$ a solid thermal node $+$ a liquid one



Problem Formulations

Decision variables

- **Cooling network topology** N
- **System pressure drop** P_{sys}

Metrics

- **Pumping power** $W_{pump} = \frac{P_{sys} \cdot Q_{sys}}{\eta}$
  - Q_{sys}: system flow rate; η: efficiency term
- **Thermal gradient** $\Delta T = \max_i(\Delta T_i)$
  - ΔT_i: range of node temperatures in i-th source layer
- **Peak temperature** T_{max}

Problem Formulations

- **Problem 1: Pumping Power Minimization**

$$\begin{aligned} \text{min} \quad & W_{pump}, \\ \text{s.t.} \quad & P_{sys} \in \mathbb{R}^{+},\ \boldsymbol{N} \in \mathcal{N},\ T_{max} \leq T^{*}_{max},\ \Delta T \leq \Delta T^{*}. \end{aligned} \tag{1}$$

($\mathcal{N}$: all legal cooling networks)

- **Problem 2: Thermal Gradient Minimization**

$$\begin{aligned} \text{min} \quad & \Delta T, \\ \text{s.t.} \quad & P_{sys} \in \mathbb{R}^{+},\ \boldsymbol{N} \in \mathcal{N},\ T_{max} \leq T^{*}_{max},\ W_{pump} \leq W^{*}_{pump}. \end{aligned} \tag{2}$$

- Design rules from ICCAD 2015 Contest

Pumping Power Minimization – Flow

Input: $\boldsymbol{N}_{init}$, ΔT^*, T^*_{max}, stack description and floorplan files.
Output: $\boldsymbol{N}$, P_{sys}.

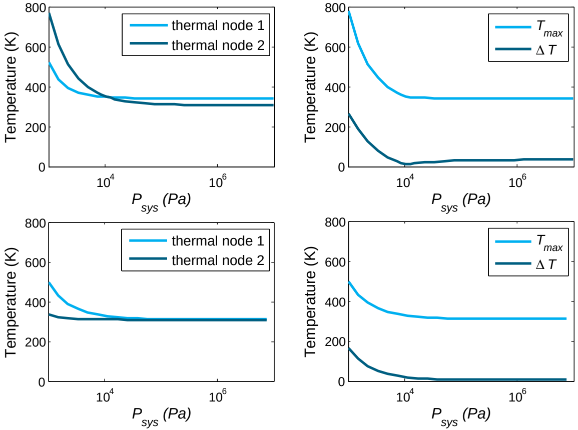
$\boldsymbol{N} \leftarrow \boldsymbol{N}_{init}$;
while #iteration is within the limit **do**
  Obtain neighboring network solution $\boldsymbol{N}'$;
  $W'_{pump} \leftarrow$ EVALUATENETWORK ($\boldsymbol{N}'$, ΔT^*, T^*_{max});
  $\boldsymbol{N} \leftarrow \boldsymbol{N}'$ or not according to SA mechanism;
  if W'_{pump} converges **then** return $\boldsymbol{N}$ and P_{sys};
end while

The problem is divided into two levels:

- **Inner**: P_{sys} is varied to minimize W_{pump} for a specific $\boldsymbol{N}$, which evaluates $\boldsymbol{N}$
- **Outer**: simulated annealing (SA) searches for a good $\boldsymbol{N}$

Pumping Power Minimization – Temperature vs. Pressure

- As P_{sys} increases, T_{max} decreases and finally becomes approximately constant
- $\Delta T = f(P_{sys})$ is either uni-modal or monotonically decreasing



Pumping Power Minimization – Network Evaluation

- Replace W_{pump} by P_{sys}, as W_{pump} vs. P_{sys} is monotonic for a specific $\boldsymbol{N}$
- Ignore T_{max} first, as it is easier to handle
  - Step 1: solve the problem without constraint T^*_{max}
  - Step 2: check T_{max} and find optimal solution by binary search

function EvaluateNetwork($\boldsymbol{N}$, ΔT^*, T^*_{max})
    Minimize W_{pump} s.t. $\Delta T \leq \Delta T^*$;
    if $\Delta T > \Delta T^*$ **then**
        return $+\infty$;
    else if $T_{max} > T^*_{max}$ **then**
        Minimize W_{pump} s.t. $T_{max} \leq T^*_{max}$;
        if $\Delta T > \Delta T^*$ or $T_{max} > T^*_{max}$ **then**
            return $+\infty$;
        else
            return W_{pump};
        end if
    else
        return W_{pump};
    end if
end function

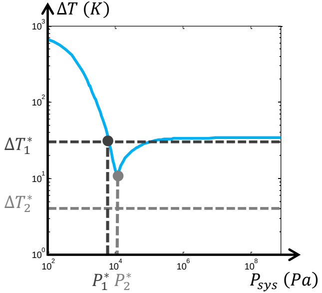
Pumping Power Minimization – Network Evaluation

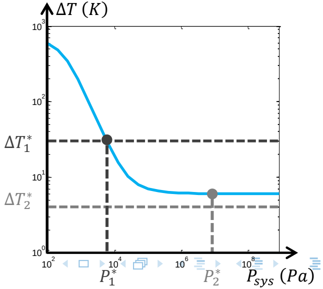
In step 1, by further substituting $\Delta T = f(P_{sys})$, Problem 1 becomes single-variable:

$$\begin{aligned} \text{min} \quad & P_{sys}, \\ \text{s.t.} \quad & P_{sys} \in \mathbb{R}^+,\ f(P_{sys}) \leq \Delta T^*. \end{aligned} \tag{3}$$

Solve (3) by searching (with three probing points):

- If a feasible P_{sys} exists, return optimal P_{sys}
- Otherwise, return the P_{sys} for minimum f (show the nonexistence of feasible P_{sys})

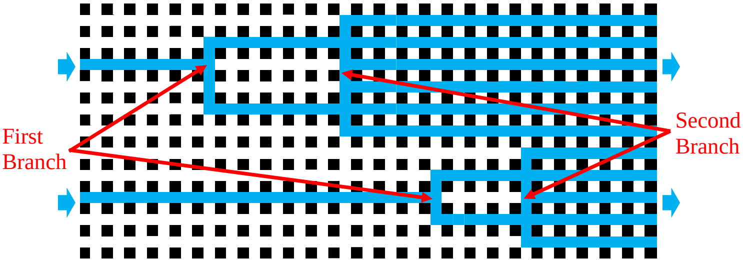




Pumping Power Minimization – Tree-like Cooling Network

Hierarchical tree-like structure is simple and can balance cooling:

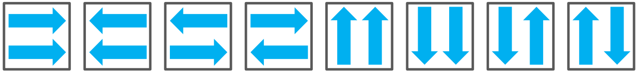
- Between upstream and downstream
- Among different trees



Pumping Power Minimization – Network Topology Optimization

Stage #	Step Size	Objective Function	Simulator	Runtime for an Iteration
1	10	ΔT	2RM	short
2	10	W'_{pump}	2RM	medium
3	2	W'_{pump}	2RM	medium
4	2	W'_{pump}	4RM	long

- In stage 1, ΔT under a **fixed** P_{sys} is used as cost function to accelerate
- Eight types of global flow directions are attempted



Thermal Gradient Minimization – Network Evaluation

Problem for a specific $\boldsymbol{N}$ can be similarly solved:

- Its simplified form becomes:

$$\begin{aligned} \text{min} \quad & f(P_{sys}), \\ \text{s.t.} \quad & P_{sys} \in \mathbb{R}^+, \; P_{sys} \leq P^*_{sys}, \end{aligned} \tag{4}$$

- Solving (4) is simpler:
  - If P^*_{sys} locates on falling side of f, it is optimal already
  - Otherwise, adopt golden section search

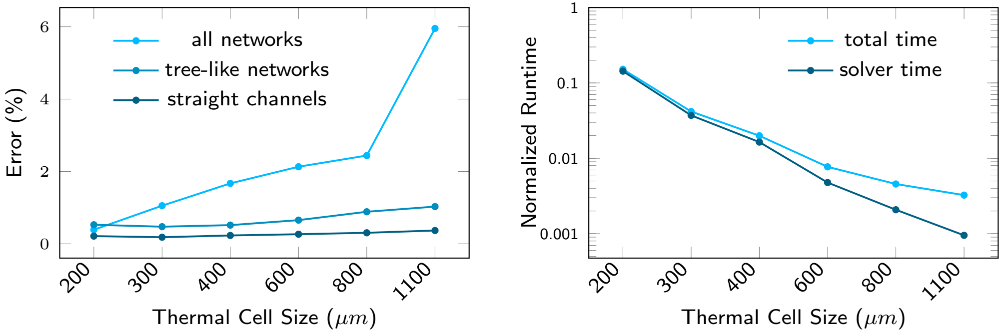
Thermal Gradient Minimization – Network Topology Optimization

Stage #	Step Size	Objective Function	Simulator	Runtime for an Iteration
1	10	$\Delta T'$	2RM	short
2	10	$\Delta T'$	4RM	medium
3	2	$\Delta T'$	4RM	medium

Minimizing W_{pump} under a fixed P_{sys} is unrelated to temperature and meaningless, but minimizing ΔT under a fixed P_{sys} is safe $\implies$ **speed-up**

- Some iterations are evaluated by one simulation under a fixed P_{sys}
- The original stage 1 is no longer needed

Experimental Results – Faster 2RM Model



- 5 benchmarks, 40 network samples, 6 thermal cell sizes and 13 pressures
- Tree-like networks, $400\mu m$ thermal cells: 0.52% errors (compared to 4RM), runtime reduced from **3.37s to 0.07s**

Experimental Results – Pumping Power Minimization

	Case #	1	2	3	4	5
Baseline	P_{sys} (kPa)	12.98	6.23	7.85	9.71	N/A
	T_{max} (K)	322	314	321	314	N/A
	ΔT (K)	15.0	10.0	15.0	10.0	N/A
	$\boldsymbol{W_{pump}}$ $\boldsymbol{(mW)}$	**10.41**	**6.91**	**8.34**	**11.65**	**N/A**
Manual (1st place in ICCAD Contest)	P_{sys} (kPa)	8.86	5.54	6.98	9.45	40.1
	T_{max} (K)	357	336	328	336	338
	ΔT (K)	15.0	10.0	15.0	10.0	10.0
	$\boldsymbol{W_{pump}}$ $\boldsymbol{(mW)}$	**1.72**	**1.51**	**3.36**	**2.96**	**113.96**
Ours	P_{sys} (kPa)	8.72	5.13	5.81	8.27	40.10
	P_{system} (kPa)	358	336	337	335	338
	ΔT (K)	15.00	10.0	15.0	10.00	10.00
	$\boldsymbol{W_{pump}}$ $\boldsymbol{(mW)}$	**1.66**	**1.37**	**1.90**	**2.68**	**113.96**

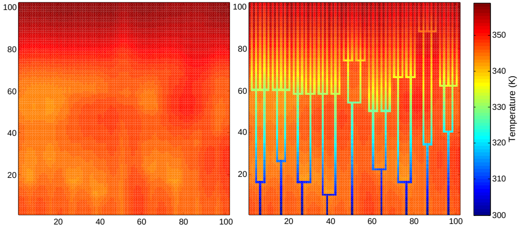
- **79.61%** better than baseline (unidirectional straight channels)
- **16.35%** better than 1st place in ICCAD 2015 Contest

Experimental Results – Thermal Gradient Minimization

	Case #	1	2	3	4	5
Baseline	P_{sys} (kPa)	26.08	14.43	17.82	26.51	45.81
	T_{max} (K)	316	309	316	308	338
	W_{pump} (mW)	42.0	37.0	43.0	43.4	148.2
	$\boldsymbol{\Delta T}$ $\boldsymbol{(K)}$	**8.75**	**5.42**	**11.42**	**4.76**	**26.48**
Ours	P_{sys} (kPa)	16.51	8.96	11.46	13.80	40.06
	T_{max} (K)	338	319	327	321	338
	W_{pump} (mW)	5.67	5.66	6.56	4.16	113.80
	$\boldsymbol{\Delta T}$ $\boldsymbol{(K)}$	**5.54**	**3.81**	**7.12**	**3.87**	**9.64**

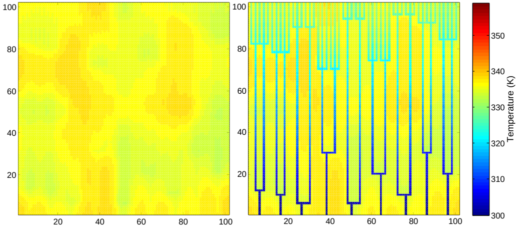
- Constraint W^*_{pump} on W_{pump} is set to 0.1% of die power
- **37.27%** better than baseline

Experimental Results – Example Temperature Maps



bottom source layer channel layer

(a) Puming power minimization



bottom source layer channel layer

(b) Thermal gradient minimization