State of the Art Cooling Technologies
Ways Toward Zero Emission Datacenters

Thomas Brunschewiler, Werner Escher, Javier Goicochea Ingmar Meijer, Bruno Michel, Monika Müller, Jeff Ong, Stephan Paredes, Patrick Ruch

Advanced Thermal Packaging, IBM Research - Zurich
Outline

Introduction
- The holistic view

Liquid cooling
- Ultra Thin High Efficiency Heat Sinks
- Direct Jet Impingement

Thermal Interface Materials
- Hierarchical nested channels for improved thermal contact and reduced assembly times

Being ready for future technologies – 3D stacking
- Why 3D?
- Improved thermal contact between stacked chips – The Thermal Underfill
- Interlayer cooling between 3D stacked chips

Improved Heat Transfer on the Nanoscale
- Phonon engineering using molecular dynamics
Outline

Introduction
- The holistic view

Liquid cooling
- Ultra Thin High Efficiency Heat Sinks
- Direct Jet Impingement

Thermal Interface Materials
- Hierarchical nested channels for improved thermal contact and reduced assembly times

Being ready for future technologies – 3D stacking
- Why 3D?
- Improved thermal contact between stacked chips – The Thermal Underfill
- Interlayer cooling between 3D stacked chips

Improved Heat Transfer on the Nanoscale
- Phonon engineering using molecular dynamics
The zero emission data center concept

Brunschwiler et al., IBM J. Res. & Dev. 53 (2009) 11
The Holistic View - From the Transistor to the Globe
Outline

Introduction
- The holistic view

Liquid cooling
- Manifold Micro-channel Heat Sinks
- Direct Jet Impingement

Thermal Interface Materials
- Hierarchical nested channels for improved thermal contact and reduced assembly times

Being ready for future technologies – 3D stacking
- Why 3D?
- Improved thermal contact between stacked chips – The Thermal Underfill
- Interlayer cooling between 3D stacked chips

Improved Heat Transfer on the Nanoscale
- Phonon engineering using molecular dynamics
Ultra Thin High Efficiency Heat Sinks

- Motivation: Find the best coolant and the best structure for ultra-compact heatsinks (thickness < 2 mm)

- Nanofluid thermal properties explained by effective medium theory which means they cannot ‘magically’ improve heat transfer
  
  ➞ Water provides the best combination of material properties

- Flat heatsinks reduce the board pitch of future systems from >30 mm (1U) to 3 mm (1/10 U)

- Optimum design provides a total thermal resistance of 0.09 cm²K/W @ V =1.3 l/min, Δp = 0.22 bar
  
  ➞ maximum power density > 700 W/cm² for ΔT = 65 K

- Increasing inlet temperature to 70ºC (190°F) enhances the heat sink efficiency >40%
**Liquid Cooled Micro-channel Heat Sinks**

- **Thermal Performance:**
  - Total thermal resistance:
    \[
    R''_{\text{tot}} = R''_{\text{end}} + R''_{\text{conv,eff}} + R''_{\text{bulk}}
    \]

    \[
    R''_{\text{end}} = \frac{h_{\text{base}}}{k_s} \quad R''_{\text{conv,eff}} = \frac{w_{\text{ch}} + w_w}{w_{\text{ch}} \alpha_{\text{conv}} + 2h_{\text{ch}} \alpha_{\text{conv}} \gamma_w}, \quad \alpha_{\text{conv}} = \frac{Nu \lambda_f}{d_h}
    \]

- **Hydrodynamic Performance:**
  - Frictional pressure loss:
    \[
    \Delta p = \frac{2 f_{\text{app}} \rho_f \bar{u}^2 L_{\text{ch}}}{d_h} \sim \frac{\mu_f \sqrt{E_{\text{ch}}}}{d_h^4}
    \]
What is the most efficient cooler structure?

Bifurcating vs. Massive Multi-branching

Parallel Channel network results in: >2x Q for P = const. => >2 COP

Massive multi-branching is more efficient than just bifurcating

Manifold Micro-Channel Heat sink

Optimization by Analytical Model

Design Optimum at: \[
\frac{R_{\text{bulk}}}{w_{\text{HT,ch}}} = -\frac{R_{\text{cnv}}}{w_{\text{HT,ch}}}
\]

Optimization by Analytical Model

Design Optimum at: \[
\frac{R_{\text{bulk}}}{w_{HT,ch}} = -\frac{R_{\text{cvn}}}{w_{HT,ch}} \quad \& \quad \frac{R_{\text{bulk, opt}}}{K} = -\frac{R_{\text{cvn, opt}}}{K}
\]

Experimental Investigation – Testvehicles

Hydrodynamic and thermal performance

Cooling of up to > 700 W/cm²

Direct Liquid jet impingement

Arrayed jets, distributed return

Biological vascular systems are optimized for the mass transport at low pressure

Cooling of up to 350 W/cm²

SEM cross-section of two-level jet plate with diameter of 35µm

Direct Liquid Jet-Impingement Cooling with Micron-Sized Nozzle Array and Distributed Return Architecture, T. Brunschwiler et al., ITERM 2006
Silicon Heat Sink - Packaging

Outline

Introduction
- The holistic view

Liquid cooling
- Ultra Thin High Efficiency Heat Sinks
- Direct Jet Impingement

Thermal Interface Materials
- Hierarchical nested channels for improved thermal contact and reduced assembly times

Being ready for future technologies – 3D stacking
- Why 3D?
- Improved thermal contact between stacked chips – The Thermal Underfill
- Interlayer cooling between 3D stacked chips

Improved Heat Transfer on the Nanoscale
- Phonon enginiering using molecular dynamics
Thermal interface background...

- Interfaces are large portion of total resistance

- Particle filled materials have cost benefit compared to solder or indium
  - Easier processing, no metallization, flexibility for many applications

- Conductivity increase with higher particle loading
  - Viscosity and shear strength also increase
  - If bondline thickness increases - **No Gain!**

- Assembly loads cannot be too high
  - C4 crushing, chip cracking
  - Substrates bend trapping thick TIM

- Hierarchical Nested Channel (HNC) creates thinner bondlines with higher conductivity materials using low assembly forces...

→ TIM up to 50% of system thermal resistance !!!
Filled pastes

**Requirement**
- High fill factor → up to 80%
- Small particles resulting in small gap
- Low voiding due to paste pumping

**Composition:**
- Filler: Al2O3, ZnOx
- Paste: silicone oil, epoxy oil

**Rheometric measurement**
pastes are Bingham fluids

- **ATC 3.8 filler particles**
  - diameter distribution:
    - average 5μm
    - max 40 μm

- **ATC 3.8 after cycling**
  - voiding due to paste pumping
Fluidic analysis of particle stacking

The Paste Cross: lines between corners are regions of stacked particles

Bondline cross section: Particles along gradient do not move with the matrix and start piling up...

- Paste cross develops along flow bifurcation line (between corners)
  - Longer flow distance (higher pressure drop) to chip corner
  - Cross line: high particle concentration, large particles and stacks of particles

- Stacking must be optimized for best bondline performance
  - Uniformity of interface material, assembly loads, reduced failures...
First experiments in particle stacking...

- Surface channels create lower pressure drop and disrupt flow bifurcation
- Pressure drop between diagonal and vertical/horizontal channels must be equal for optimized control of particle stacking

- Channel designed for uniform pressure drop across chip
Optimization of HNC for thermal interface

- **Design tradeoff between:**
  - Reduced bondline thickness with smaller HNC cell sizes (→ more channels)
  - Increasing resistance of HNC with smaller HNC cell size (→ fewer channels)

- **Optimum can be modeled and seen experimentally**
Results with IBM material and HNC

- 40% reduction over Flat surface: $8 \text{ Cmm}^2/\text{W}$, 25um bondlines
- Measurement from center sensor directly below HNC channel
- Rough cooler surface, ±3 um planarity accuracy

Test Specifications: 5 bar assembly load, 20mm chip, 50W/cm², error bars = average deviation of 8 tests
HNC Benefits Increase with Larger Areas

- **Large areas are increasingly difficult to control bondline thickness**
  - Large assembly forces (scale with $L^2$)
  - High forces cause components to bend, trapping TIM material in thick gaps
- **HNC reduces the bondline thickness by up 90% with low cost commercial materials**

![Pressure limited bondline benefit with HTCI](image)

(2 bar assembly load, Wacker P12 TIM)
Outline

Introduction
- The holistic view

Liquid cooling
- Ultra Thin High Efficiency Heat Sinks
- Direct Jet Impingement

Thermal Interface Materials
- Hierarchical nested channels for improved thermal contact and reduced assembly times

Being ready for future technologies – 3D stacking
- Why 3D?
- Improved thermal contact between stacked chips – The Thermal Underfill
- Interlayer cooling between 3D stacked chips

Improved Heat Transfer on the Nanoscale
- Phonon engineering using molecular dynamics
From Frequency Scaling to Multi-Core Architectures

**Frequency and power cap**

**Intel CPU Trends**
(sources: Intel, Wikipedia, K. Oluotun)

- Dual-Core Itanium 2
- Pentium 4
- Pentium
- 386

**Multi-core examples**

- P6, 2 cores
- BlueGene/L, 2 cores
- Cell-BE, 8 cores
- Z6, 4 cores

**Parallel computing:**
→ demanding software development
Multi-Core Architecture: Core – Cache Bandwidth Bottleneck

... at constant off-chip bandwidth

- Cores share constant off-chip bandwidth
- Core proportional system performance demands cubic cache size scaling
- Total chip area increase
  - signal delay in wires
  - lithographic limit reached (~6cm²)
3D Stacking Types

3D Stacking

Low bandwidth:
→ high parasitics
→ low interconnect density

Vertical Integration

High bandwidth:
→ high complexity
→ high power densities

Monolithic Integration
Fabrication Sequence

Vertical interconnects
Through-Silicon-Vias

Transistor fabrication

Wafer thinning

Bonding: soldering

Chip stack
3D chip stacks and heat dissipation

- 3D CS cavity height (H): 60 µm → 20 µm
- Chip stacks constrain heat dissipation
- Bottleneck: Underfill material. Initially designed to transfer mechanical stress from solder balls
  → enhance heat dissipation

Underfill materials

- Epoxy containing dispersed dielectric particles that match CTE of solder balls (i.e. SiO₂ + Al₂O₃, SiC, AlN, BN, etc.)
- Stress reduction in solder balls and BEOL (i.e. at the periphery)
  - Stress reduction of 10x increases 10-100x lifetime
- Dispensed using capillary forces (Laplace-Young equation)

Thermal performance

- Low filling fractions ~ 60 wt% or 30 vol%
- No thermal percolation between particles
- Viscosity limited → increases exponentially with particle loading
- Typically K_{eff} < 1 W/m-K ≈ 0.6 - 0.8 W/m-K
  (worse than most thermal interface materials)

- Ideally: High filling fraction and thermal percolation
Heat conduction modes in UF

A: Series
B: Parallel
- $K(\text{epoxy}) = 0.23$ W/m-K
- $K(\text{SiO}_2) = 1.50$ W/m-K
- $K(\text{Al}_2\text{O}_3) = 22.0$ W/m-K
- Particle distance > 0.5 µm

What is the impact of particle arrangement on the UF thermal performance?
Possible desired configurations:

- $d/H = 1.00$
- $d/H = 0.52$
- $d/H = 0.60$
- $d/H = 0.50$
Sequential Underfilling

- Method

1) Convective particle stacking
   - Syringe with stirrer
   - Particles
   - Carrier fluid
   - Module with chip stack

2) Evaporation
3) Capillary filling

Percolating thermal UF
Capillary thermal UF

4) Cure
Sequential Underfilling

Method

1) Convective particle stacking
   - Syringe with stirrer
   - Pressure supply
   - Particles: SiO$_2$: 10, 38, 45, 53 ± 1 µm
     - Al$_2$O$_3$: 36 ± 8 µm
   - Concentration < 1 vol%
   - Magnetic steering

2) Evaporation
   - Hot plate (65 °C) → 1-2 min

3) Capillary filling

4) Cure
   - 3 hours @ 65 °C

Particle filling:
- Applied pressure < 10$^5$ Pa
- 10 ml syringe: water or isopropanol
- Particles:
  - SiO$_2$: 10, 38, 45, 53 ± 1 µm
  - Al$_2$O$_3$: 36 ± 8 µm
- Concentration < 1 vol%
- Magnetic steering

Evaporation:
- Hot plate (65 °C) → 1-2 min

Epoxy filling:
- EPO-TEK 301-2
- Time: 115 – 350 s

Epoxy curing:
- 3 hours @ 65 °C
Particle arrangement: Experimental results

- Filling fraction of particles confined in cavities

- Filling fractions follow theoretical predictions
- Lower fractions reached due to random arrangement and size variations
- For small d/H values RCP fraction becomes the upper bound reachable

Filling fraction:

- FCC close-packing
- Random close-packing
- Confined space close-packing
- Optical microscope, H=62µm
- Optical microscope, H=58µm
- CT-reconstruction, H=62µm
- Carman-Kozeny correlation, H=62µm

CT scan
Particle arrangement: Percolation

- Cross-section: $\text{SiO}_2 \rightarrow d = 10$ and 38 $\mu$m

- Particle percolation: $\text{SiO}_2 \rightarrow d = 38 \mu$m, $H = 58 \mu$m
**Particle filling**

- Time to fill a 10 mm cavity vs particle diameter

![Graph showing time to fill a 10 mm cavity vs particle diameter](image)

- Initial concentration 1/8 vol% and \( f = 60\% \)
- Carman-Kozeny assumed valid for all diameters
- Filling velocity controlled by the resulting hydraulic diameter
- Time increases as particle diameter becomes smaller (i.e. \( t \sim 1/d^2 \))
**Summary and outlook**

- Sequential method for the formulation of percolating thermal underfill:
  - Convective filling
  - Fluid carrier evaporation
  - Epoxy refilling
  - Curing

- Filter element promotes high filling fractions and thermal percolation. Key ingredients for achieving high thermal conductivity values

- 36 µm Al₂O₃ particles → $K_{eff} = 1.3 \pm 0.1 \text{ W/m-K} \ (60 \% > \text{commercial TUF})$ in a 62 µm height Si-Si cavity

- Decoupling the formulation steps allows the characterization of the resulting particle filling and associated thermal performance. Provides great flexibility on particle, carrier fluid and epoxy selection
Limits of Traditional Back-Side Heat Removal

1-dimensional heat flux model with thermal gradient budget of 65K

Back-side cooling potential and limit

Guidelines from a thermal perspective:
- MPU as close as possible to the cold plate
  - Lower peak temperature \(\rightarrow\) high heat flux is conducted through minimum number of layers
  - Memory can handle 15K higher junction temperatures
- Non-identical hot spot locations

Unacceptable:
- Two identical MPU’s with overlapping hot spot
- More then two MPU layers

Heat removal limit constrains electrical design
Scalable Heat Removal by 3D Interlayer Cooling

- 3D integration will require interlayer cooling for stacked logic chips.
- Bonding scheme to isolate electrical interconnects from coolant.
- Heat removal scales with the number of dies.

- **Cool between logical layers with optimal vias**
  - Best performance with 200 μm pin fins
  - Through-silicon via height limit, typically 150μm
  - Microchannel, pin fins staggered/in line, drop shape

- **Interlayer cooling of 3D stacked chips**
  - Remove 180 W/cm² per layer or
  - Remove 7.2 KW from 10 layers with 4 cm²

**Interconnect compatible heat transfer structures**
- Microchannel
- Pin fin inline / staggered
Electro-Thermal Co-Design

- Chip design
- Heat Transfer Building Blocks
- Heat Transfer Structure Design

**Efficient heat removal**
- Heat transfer structure
- Modulation of heat transfer structure

**Increase in local hot spot flow rate**
- Fluid focusing

**Feedback**
Experimental Validation: Pyramid Chip Stack

**Thermal Demonstrator:**
- Three active tiers, cooled with four cavities
- Polyimide bonding → represents wiring levels
- Multi-scale modeling accuracy validated (+/-10%)

**Realistic Product Style Stack:**
- Aligned hot-spot heat flux of 250W/cm² possible
Nano-Surface: “Lotus-Effect”

Super-hydrophobicity:
- Micron / nano-sized topography with wax
- Only in contact with 0.1% of the surface
  → reduction of pressure drop

Water drops on a Lotus leaf

Biology -- surface morphology -- Engineered
Outline

Introduction
- The holistic view

Liquid cooling
- Ultra Thin High Efficiency Heat Sinks
- Direct Jet Impingement

Thermal Interface Materials
- Hierarchical nested channels for improved thermal contact and reduced assembly times

Being ready for future technologies – 3D stacking
- Why 3D?
- Improved thermal contact between stacked chips – The Thermal Underfill
- Interlayer cooling between 3D stacked chips

Improved Heat Transfer on the Nanoscale
- Phonon engineering using molecular dynamics
Thermal transport in electronics

▪ Heat path in microprocessors

![Diagram showing heat path in microprocessors]

- Strained Si, Ge, SiGe
- Gate
- Buried oxide
- Silicon substrate
- TIM1
- TIM2
- Heat sink
- Ambient
- Electron
- Hot spot
- Optical modes
- Acoustic modes
- Buried oxide
- Silicon substrate
- TIM1
- Cap
- TIM2
- Heat sink
- Ambient

Transistor level

Interface

Heat path
On-chip thermal challenges

UTB FD-SOI (2012) + SiGe, Ge, III-V, FinFETs (2014) & Nanowires (2024)

- Smaller dissipation areas → UTB FD-SOI, Nanowires, etc.
- Lower dimensionality (≈ thermal conductivity ~5-10x lower)
- III-V Semiconductors (≈ thermal conductivity ~1-9x lower)
  
  Si: 140, Ge: 61, lnSb: 17, GaAs: 46, InAs: 27
  
  lnP: 77, GaN: 130, GaP: 110 [W/mK], etc.
- Multiple interfaces and large presence of oxide layers (e.g. SOI)
- Heat transport described in terms of phonons → Fourier law not valid!

Could electrical improvements deal with additional heat dissipation constraints?
Modeling heat conduction in microprocessors

Sub-continuum heat transfer in semiconductors

Steady state:

\[ \frac{\partial T}{\partial x} = 0 \quad \text{and} \quad \frac{\partial T}{\partial y} = 0 \]

Knudsen number (kn) = \( \Lambda / d \)

Transient:

\[ \frac{\partial T}{\partial y} = 0 \]

• Creation of hotspots due to phonon confinement

Temperature

Knudsen number (kn) = 0.01

Knudsen number (kn) = 1.00

Knudsen number (kn) = 100

Characteristic dimension is reduced

Diffusive (conduction)

Ballistic (radiation)

Thermal Conductivity

\[ k_{\text{eff}} \]

\[ k_{\text{bulk}} \]
Multiscale heat conduction

Sub-continuum heat transfer in semiconductors

- Failure of heat conduction equation
- Heat transport is described in terms of phonons
- Heat dissipation is a function of atomic forces

Temporal and spatial range of applicability of the currently thermal transport models used to predict heat conduction

\[ \lambda \]
heat carrier wavelength

\[ \Lambda \]
heat carrier mean free path

\[ l_r \]
relaxation length

\[ \tau_r \]
collision time

\[ \tau_c \]
heat carrier relaxation time

\[ t >> \tau_r \]
time scale greater than collision time

\[ t > \tau_r \]
time scale greater than relaxation time

\[ \Lambda > l_r \]
spatial characteristic dimension

\[ L >> l_r \]
Spatial characteristic dimension
Heat dissipation at solid-solid interfaces

- Interactions at Si + Metal (Al, Ag or Au) interfaces*

  Electrons: \[ k_e \frac{d^2T_e}{dx^2} - G(T_e - T_{ph}) = 0 \]

  Phonons: \[ k_{ph} \frac{d^2T_{ph}}{dx^2} + G(T_e - T_{ph}) = 0 \]

* To be presented at Semi-Therm 27, 2011
Heat dissipation at solid-fluid interfaces

- Thermal rectification at water/functionalized silica interfaces*

Silica-SAMs-Water

* Hu, Goicochea, et al. 2010, APL 95, 151903
### Summary

- **Thinking global about energy usage**
  - Crucial to allow exascale computers
  - Total cost of ownership perspective
  - Demand and supply of sensible heat: Thermal energy re-use
  - Cooling chips with “hot” water to obtain recyclable heat (65°C / 149 F) for remote heating
  - Joint project with ETH: Aquasar

- **Key components enabling efficiency and energy re-use (short term)**
  - Improved thermal conductivity, reduced bondline
  - Improved heat transfer with micro and nanotechnology
  - Minimized exergy losses with water and hotspot cooling

- **Future interlayer cooling of 3D stacked chips (medium term)**
  - Computer efficiency is dominated by communication
  - Collapse one rack of computers to one cubic centimeter and improve efficiency by more than 10x

- **Phonon Transport Engineering (long term)**
  - Improved nanoscale heat transfer from nano to macro
  - Chemical surface functionalization
  - Geometrical patterning for phonon resonance matching
  - Long-term research: Modeling and experiment needed