Intel Corporation

Meteor Lake and Arrow Lake
Intel Next-Gen 3D Client Architecture Platform with Foveros

Wilfred Gomes, Slade Morgan, Boyd Phelps, Tim Wilson, Erik Hallnor
Intel’s Mission
We create world-changing technology that improves the life of every person on the planet.

**PC Era**
1B Internet Connected Devices

**Mobile + Cloud Era**
10B Cloud Connected Devices

**Pervasive Intelligence Era**
100B Edge Connected Devices

**Empower Everyone**

- 1B Internet Connected Devices
- 10B Cloud Connected Devices
- 100B Edge Connected Devices

Intel’s Mission:
We create world-changing technology that improves the life of every person on the planet.
Experience Driven Client

- Experience First
- Purposeful Performance
- Dynamic
- Scale

Pervasive Intelligence Era
100B Edge Connected Devices

Experience Driven
Compute
Experience Driven

- Experience First
- Purposeful Performance
- Dynamic
- Scale

Implications for Client

- **Performance**
  - Performance, Perf/Watt
- **Flexibility**
  - Mix & Match Blocks and Functions
- **Innovation Pace**
  - Time-to-Market
- **Next Exponential**
  - The next generation of devices
<table>
<thead>
<tr>
<th>Monolithic</th>
<th>Disaggregated</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Highest</strong></td>
<td><strong>Limited</strong></td>
</tr>
<tr>
<td>Very Limited</td>
<td>Faster (release per new function)</td>
</tr>
<tr>
<td>Slow (per SOC basis)</td>
<td>Higher</td>
</tr>
<tr>
<td>Low</td>
<td><strong>Scale</strong></td>
</tr>
<tr>
<td><strong>Performance</strong></td>
<td><strong>Innovation Pace</strong></td>
</tr>
<tr>
<td><strong>Flexibility</strong></td>
<td><strong>Faster (release per new function)</strong></td>
</tr>
<tr>
<td><strong>Low</strong></td>
<td><strong>Higher</strong></td>
</tr>
</tbody>
</table>

Intel Confidential
Can we get **monolithic** performance with **disaggregated** architecture benefits?
## Disaggregation Journey

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Packaging</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Haswell</strong></td>
<td><strong>2D</strong></td>
<td>Intel 22nm</td>
</tr>
<tr>
<td>Ultra Thin &amp; Light</td>
<td>MCP</td>
<td>GloFo 14nm</td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td>Intel 14nm</td>
</tr>
<tr>
<td><strong>Kaby Lake G</strong></td>
<td><strong>2.5D + 2D</strong></td>
<td>Intel 22FETL</td>
</tr>
<tr>
<td>Ultra Thin &amp; Perf Graphics</td>
<td>EMIB + MCP</td>
<td>Intel N7</td>
</tr>
<tr>
<td>2017</td>
<td></td>
<td>TSMC N7</td>
</tr>
<tr>
<td><strong>Lakefield</strong></td>
<td><strong>3D</strong></td>
<td>Intel 10nm</td>
</tr>
<tr>
<td>Ultra Thin &amp; Light</td>
<td>50µm Foveros</td>
<td>TSMC N5</td>
</tr>
<tr>
<td>2019</td>
<td></td>
<td>Intel 7</td>
</tr>
<tr>
<td><strong>Ponte Vecchio</strong></td>
<td><strong>2.5D + 3D</strong></td>
<td></td>
</tr>
<tr>
<td>High Density &amp; Performance</td>
<td>EMIB + 36µm Foveros</td>
<td></td>
</tr>
<tr>
<td>2022</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Transistor Diversity Opportunity

- SOC Optimized
- Graphics Optimized
- Memory Optimized
- Mobile CPU Optimized
- Desktop CPU Optimized
- I/O Optimized
Advanced Packaging

**Foveros Direct**
- Bump pitch: < 10 microns
- Bump density: > 10,000/mm²
- Power: < 0.05 pJ/bit

**Foveros Omni**
- Bump pitch: 25 µm
- Bump density: 1,600/mm²
- Power: < 0.15 pJ/bit

**Foveros**
- Bump pitch: 50-25 µm
- Bump density: >400-1600/mm²
- Power: < 0.15 pJ/bit

**EMIB**
- Bump pitch: 55-45 µm
- Bump density: 330-772/mm²
- Power: 0.15 pJ/bit

**Standard Package**
- Bump pitch: 100 µm
- Bump density: 100/mm²
- Power: 1.7 pJ/bit
Advanced Packaging

Foveros Direct
- Bump pitch: <10 microns
- Bump density: >10,000/mm²
- Power: <0.05 pJ/bit

Foveros Omni
- Bump pitch: 25 µm
- Bump density: 1,600/mm²
- Power: <0.15 pJ/bit

Foveros
- Bump pitch: 50-25 µm
- Bump density: >400-1600/mm²
- Power: 0.15 pJ/bit

High Yield & High Volume Manufacturing for large number of tiles

Interconnect Density
- EMIB
- Die Center
- Die Edge

Power Efficiency

Cu Ni
Cu Ni
Cu

Power - 0.05-0.15 pJ/bit

Bump pitch - 25-55 µm
Bump density - 330-772/mm²

Intel Confidential
Advanced Packaging

Foveros Direct

Bump pitch = < 10 microns
Bump density = >10,000/mm²
Power = < 0.05 pJ/bit

Foveros Omni

Bump pitch = 25 µm
Bump density = 1,600/mm²
Power = <0.15 pJ/bit

Foveros

Bump pitch = 50-25 µm
Bump density = >400-1600/mm²
Power = 0.15 pJ/bit

EMIB

Bump pitch = 55-45 µm
Bump density = 330-772/mm²
Power = 0.50 pJ/bit

‘Mix and match’ tiles in base die complex

4x higher interconnect bump density vs EMIB
Advanced Packaging

**Foveros Direct**
- **Bump pitch**: < 10 microns
- **Bump density**: > 10,000/mm²
- **Power**: < 0.05 pJ/bit

**Foveros Omni**
- **Bump pitch**: 25 µm
- **Bump density**: 1,600/mm²
- **Power**: < 0.15 pJ/bit

**Foveros**
- **Bump pitch**: 50-25 µm
- **Bump density**: > 400-1600/mm²
- **Power**: 0.15 pJ/bit

**Bump pitch**
- 55-45 µm
- 33-772/mm²
- 0.50 pJ/bit

**16x higher interconnect bump density vs Foveros @ 36 µm pitch**

**Higher B/W at lower latency, power, & die area**
Architecture Evolution

Intel 386
1980
Design with Schematics
Sea of transistors/gates

Pentium
1990
Design with HDL
Sea of cells/blocks

Core Duo
2005
Design with IP
functional, units/IPs

Haswell
2013
Design with 2D Chiplets

Kaby Lake G
2017
Design with 2D Chiplet and 2.5D

Lakefield
2019
First Commercial 3D Logic on Logic Stacking

Ponte Vecchio
2022
2D+ 3D
Design Large Number of Tiles
## Disaggregation Journey

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Packaging</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU/GFX partitioning</td>
<td>Hybrid Architecture</td>
<td>47 Tiles</td>
</tr>
<tr>
<td>CPU/PCH partitioning</td>
<td>Compute/Memory/IO partitioning</td>
<td>Compute/Memory/IO partitioning</td>
</tr>
<tr>
<td>2.5D + 2D</td>
<td>3D</td>
<td>2.5D + 3D</td>
</tr>
<tr>
<td>EMIB + MCP</td>
<td>50µm Foveros</td>
<td>EMIB + 36µm Foveros</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Kaby Lake G</th>
<th>Lakefield</th>
<th>Ponte Vecchio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ultra Thin &amp; Perf Graphics</td>
<td>Ultra Thin &amp; Light</td>
<td>High Density &amp; Performance</td>
</tr>
<tr>
<td>2017</td>
<td>2019</td>
<td>2022</td>
</tr>
</tbody>
</table>

---

**Meteor Lake**

Next Step in our Disaggregation Journey
Architecture and Design tradeoffs

New Flexible Tiled Architecture
Scalable Architecture

Goals

- Flexibility with core arch, count, process
- Flexibility with Graphics cores
- IO modularity
- Process node flexibility
- Ability to scale graphics and compute
- Low power to discrete graphics performance
Compute Tile

Core Count Scalability

Core Generation Scalability

Node Scalability

Cache Scalability
Graphics Tile

Core Count Scalability

Node Scalability

Cache Scalability
SOC Tile

Scalable IP Blocks

- Low power IP
- SRAM
- High Voltage
- IO

[Diagram of scalable IP blocks with labels for IO, Low Power IP, Media, Memory Control, Imaging, Display, and XPU]
I/O Extender Tile

Scalable I/O Blocks

<table>
<thead>
<tr>
<th>Number of Lanes</th>
<th>Bandwidth</th>
<th>Protocol</th>
<th>Speed</th>
</tr>
</thead>
</table>

IO Control

IO Control
Meteor Lake

Construction

Base Tile
Meteor Lake

- **Known good tiles**
- **Back side metallization**
- **Base tile with large capacitance**
- **36 µm pitch Die2Die**
- **Metal layers for IO/power delivery and Die2Die routing**
Meteor Lake

- **Base Tile**
- **3D Capacitors**
- **Package Bump**
- **Die2Die**
  - Power delivery, package IO routing
- **Modularity with active silicon for memory and logic**
- **Redistribution layers with active silicon**
Meteor Lake

Colors Represent 3D Capacitors, Voltage Islands

- Graphics Tile
- SOC Tile
- CPU Tile
- IOE Tile
FDI - Foveros Die Interconnect

- Low Voltage CMOS interface
- High Bandwidth, Low Latency
- Synchronous and asynchronous signaling
- Low area overhead
- Operation @ 2 Ghz, 0.15 – 0.3 pJ/bit
# Meteor Lake

## Interconnect

<table>
<thead>
<tr>
<th>Link</th>
<th>Mainband width</th>
<th>Mainband Protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU - SoC</td>
<td>~2K</td>
<td>2x IDI</td>
</tr>
<tr>
<td>Graphics - SoC</td>
<td>~2K</td>
<td>2x iCXL</td>
</tr>
<tr>
<td>SoC - IOE</td>
<td>~1K</td>
<td>IOSF, 4x Display Port</td>
</tr>
</tbody>
</table>
Meteor Lake
Thermals

System, Software, Silicon Co-optimization

Floorplan
Materials
System
Process
Architecture

Previous Generation
Start
Innovations
Meteor Lake

Turbo Power Capability

<1X
Meteor Lake
Power Delivery

Capacitors on Base Tile
to implement power delivery solutions for mix/match

37
141
193
376
500

14 nm
10 nm
Intel 7
Intel 4
MTL Base Die

Total Capacitance FF/um²

Co-Optimization

System, Software, Silicon Co-optimization

Decoupling/Noise
Voltage Regulator
IP Mix and Match
Form Factor
Package Optimization

MTL Base Die
<table>
<thead>
<tr>
<th>Interface</th>
<th>OPIO (On Package IO)</th>
<th>FDI (Foveros Die Interconnect)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed</td>
<td>2-8GT/s</td>
<td>2 GT/s</td>
</tr>
<tr>
<td>IO/mm^2</td>
<td>1X (110 µm)</td>
<td>10X (36 µm)</td>
</tr>
<tr>
<td>Latency</td>
<td>10-20 ns</td>
<td>&lt; 10 ns</td>
</tr>
<tr>
<td>Power</td>
<td>1pJ/bit</td>
<td>0.2 - 0.3 pJ/bit</td>
</tr>
<tr>
<td>Number of Tiles</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

Client (Haswell) 2013

Meteor Lake 2023
Can we get **monolithic** performance with **disaggregated** architecture benefits?

- **Monolithic**
  - Highest
  - Very Limited
  - Slow (per SOC basis)
  - Low

- **Disaggregated**
  - Lower (tax on latency, power, B/W)
  - Limited
  - Faster (release per new function)
  - Higher
Monolithic vs. Disaggregated

- Monolithic (High Bandwidth interface)

<table>
<thead>
<tr>
<th>Disaggregated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Custom/Targeted Process</td>
</tr>
<tr>
<td>Optimized IP Density</td>
</tr>
<tr>
<td>IP Refresh Pace</td>
</tr>
<tr>
<td>Form Factor</td>
</tr>
</tbody>
</table>

- 2-3% Disaggregation Tax

- No Optimization

Organic Package  |  Foveros  | Bump Pitch (µm)
Bump Pitch (µm) vs. Performance

Monolithic High Bandwidth interface

Disaggregated
- Custom/Targeted Process
- Optimized IP Density
- IP Refresh Pace
- Form Factor

No Optimization

Performance

Organic Package

Foveros

Transistor Performance Uplift

Total Power (Normalized)

Frequency (GHz)

Intel 7

Intel 4 6VT

Intel 4 8VT

Meteor Lake Status

Meteor Lake

booted and
in the lab

*Graphics for illustrative purposes only and not to scale.
One Architecture – Multiple Performance Points

<10W
- Small Graphics Tile
- Efficient CPU Tile
- Reduced IO Tile

>100W
- Max Config GPU Tile
- Max Config CPU Tile
- Expanded IO Tile
Scalable Architecture across Multiple Generations

- **Meteor Lake**
  - Packaging: Foveros, 36 µm pitch

- **Arrow Lake**
  - Packaging: Foveros, 36 µm pitch

- **Lunar Lake & Beyond**
  - Packaging: Foveros, 25 µm pitch
“Experience First” Client drives New Era of **System level integration**

**Monolithic performance** with disaggregated benefits

Process, packaging and architecture together make this possible

Meteor Lake first tiled disaggregated architecture on Intel 4

Architecture is extremely flexible and scales across design points and **into future**
Future of Compute

Scalable Architecture, Construction and Packaging

Large range of Thermal and Performance Envelope

Flexible across process nodes, Monolithic benefits

3D Monolithic Multi-layer, Logic on Logic

Manufacturing at Scale for the next Billion Devices
Q&A