Challenges with Design and Test of 80-Core Chip

Tanay Karnik, Sriram Vangal, Tiju Jacob, Shailendra Jain, Saurabh Dighe, Jason Howard, Greg Ruhl, Yatin Hoskote, Dinesh Somasekhar, Howard Wilson, Vasantha Erraguntla, Nitin Borkar, Vivek De, Shekhar Borkar

Circuits Research Lab, Bangalore Design Lab
Intel® Corporation

November 13, 2008
What is Terascale?

Teraflops of performance operating on Terabytes of data

- Model-based Apps
  - Recognition
  - Mining
  - Synthesis
- Personal Media
  - Creation and Management
- Health
- Entertainment, learning, and virtual travel
- Financial Analytics

Performance
- TIPS
- GIPS
- MIPS
- KIPS

Dataset Size
- Kilobytes
- Megabytes
- Gigabytes
- Terabytes

Terascale
- 3D & Video
- Multi-Media
- Multi-core
- Single-core
- Text
From few to many cores...

Optimized for speed

Pentium® processor era chips optimized for raw speed on single threads, good latency, pipelined, out of order execution

Optimized for performance/watt

Today's chips use cores which balance single threaded and multi-threaded performance

5-10 years: 10s-100s of energy-efficient IA cores optimized for latency & throughput, interconnect network, some non-IA accelerators
Terascale platform

Scalable On-die Interconnect Fabric

Special Purpose Engines

Integrated IO devices

Integrated Memory Controllers

Off Die interconnect

High Bandwidth Memory

IO

Socket Inter-Connect
80-core research processor

Goals:

• Deliver Terascale performance
  – Single precision TFLOP at desktop power
  – Frequency 5GHz
  – Bi-section B/W order of Terabits/s
  – Link bandwidth in hundreds of GB/s

• Prototype two key technologies
  – High performance FP execution core
  – On-die interconnect fabric
  – 3D-stacked SRAM memory

• Develop scalable design methodology
  – Tiled design approach
  – Mesochronous clocking
  – Power-aware capability

<table>
<thead>
<tr>
<th>Technology</th>
<th>65nm, 1 poly, 8 metal (Cu)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transistors</td>
<td>100 Million (full-chip)</td>
</tr>
<tr>
<td></td>
<td>1.2 Million (tile)</td>
</tr>
<tr>
<td>Die Area</td>
<td>275mm² (full-chip)</td>
</tr>
<tr>
<td></td>
<td>3mm² (tile)</td>
</tr>
<tr>
<td>C4 bumps #</td>
<td>8390</td>
</tr>
</tbody>
</table>
Key ingredients

- **Special-purpose cores**
  - High performance Dual FPMACs

- **2D mesh interconnect**
  - High bandwidth low latency router
  - Phase-tolerant tile-to-tile communication

- **Mesochronous clocking**
  - Modular & scalable
  - Lower power

- **Workload-aware power management**
  - Sleep instructions and packets
  - Chip voltage & frequency control
  - Active clamp for SRAM data retention
Router architecture

- 5-port, 5-stage, 16-FLIT FIFO, two lane, 5GHz
- Shared crossbar architecture, two-stage arbitration
Mesochronous interface (MSINT)

- Circular FIFO, 4-deep
- Programmable strobe delay
- Low-latency setting
80-core processor power management

- Modular clocking, low power clock distribution
- Data enabling, wide buses, clock gating
- New instructions to sleep/wake
  • pipelined wake, regulated mem sleep
- Chip voltage & frequency control
  \( (0.7-1.3V, 0-5.8GHz) \)

**Dynamic sleep**

**STANDBY:**
- Memory retains data
- 50% less power/tile

**FULL SLEEP:**
- Memories fully off
- 80% less power/tile

**21 sleep regions per tile** (not all shown)

**Data Memory**
Sleeping: 57% less power

**Instruction Memory**
Sleeping: 56% less power

**Router**
Sleeping: 10% less power
(stays on to pass traffic)

**FP Engine 1**
Sleeping: 90% less power

**FP Engine 2**
Sleeping: 90% less power

**Industry leading energy-efficiency of 19.4 GFLOPS/Watt!**
Work in Progress: stacked memory prototype

- 256 KB SRAM per core
- 4X C4 bump density
- 3200 thru-silicon vias

80-tile processor with Cu bumps

Memory access to match the compute power
3D memory architecture

On-die Mesh Interconnect

Processor Tile

Memory Tile

Memory Bus

Signals and power from package, through memory, to the processor tile

<table>
<thead>
<tr>
<th>TSV Pitch</th>
<th>190μm</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRAM die size</td>
<td>275mm²</td>
</tr>
<tr>
<td>SRAM size</td>
<td>256KB per tile, 20MB total</td>
</tr>
<tr>
<td>SRAM Power</td>
<td>7W SRAM + 2.2W IO</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>12GB/sec/tile, ~1TB/sec total</td>
</tr>
</tbody>
</table>
Die versus wafer stacking

**Die Stacking**

Possible Application: Logic + Memory

- TSV Size: $\sim 50 \, \mu m$
- Thickness: $\sim 100 \, \mu m$
- Bonding Structure: $\sim$Bump Size
- Bonding Pitch: $\sim$Bump Pitch

**Wafer Stacking**

Possible Application: Logic + Logic

- TSV Size: $< \sim 5 \, \mu m$
- Thickness: $\sim 10 \, \mu m$
- Bonding Structure: $< \sim 5 \, \mu m$
- Bonding Pitch: $< \sim 8 \, \mu m$

Challenges

- Floorplanning
- TSV processing
  - keep out regions, redundancy, electrical, yield
  - Assembly
    - die size, wafer support system, dense Cu/solder plating, alignment
- Testing
  - sort, KGD, BIST structures
- Thermals
80-core verification challenges

• **Architectural verification**
  - task synchronization across multiple tiles

• **Mesochronous clocking**
  - ideal conditions during RTL simulations
  - variable latency makes validation harder

• **Clock gating**
  - enable generation and timing
    - e.g. enable must meet phase requirement
  - clock skew between gated and ungated clocks
  - distribution of gated and ungated clocks

• **Sleep verification**
  - sleep to non-sleep interfaces
  - multiple power grids due to sleep
  - equivalence verification
Design and verification flow

- Architecture Definition
- RTL Coding (Verilog)
- Functional Validation
- Schematic Entry
- Equivalent Verification
- Floor Planning
- Pre-layout Performance Verification
- Custom Layout
- Post-layout Performance Verification
- Layout Verification
- Dummy fill
- Decap Insertion
- Tape out
- Spice Simulation for Critical Signals
80-core functional validation

• **ModelSim™ for pre-silicon validation**
  – tests written in C
  – JTAG scan emulated through C library functions
  – output files used by dynamic logic verification
  – FPMAC validation
    – different paths of mantissa and exponent logic
    – single cycle accumulation algorithm
    – post normalization
  – program execution/flow control validation
  – simultaneous multiple instruction/data packets
    – execution resumes on arrival of the required data
  – simultaneous DMEM accesses

• **Sleep and clock gating validation**
  – one PE puts another PE into sleep
  – special packet for waking PE
80-core functional validation

- **Router**
  - reset Sequence (sleep/ awake)
  - routing
    - source directed routing protocol
    - lane/link arbitration – round robin algorithm
    - packets with bubbles, chain headers: zig-zag routing
  - flow Control
    - stream of packets: check “queue full” conditions, avoid deadlocks

- **Full chip**
  - tests covering different applications
  - unused PEs/Router ports clock gated/sleep

- **Total of ~200 tests covering different corner cases**
Formal equivalence verification

- **Formality™** for random logic (except memories)
  - standard cells are modeled in Verilog
  - transistor-level Verilog netlist of the standard cells
- **Transistor-level equivalence verification**
  - proprietary tool
  - memories, blocks instantiating memories
- **Challenges**
  - **Formality™** – cannot handle Memories
    - transistor-level netlist → gate level netlist
    - matching ‘Z’s : sleep logic
- **Schematic to RTL dynamic validation**
  - COSMOS-based switch-level simulation
    - validates initial conditions – false proof
Layout verification

• Hercules™ is used for LVS and DRC
• Tiled design approach to reduce top-level runtime
  – ~12 hours for chip level DRC/LVS run
• Sleep and virtual ground verification
  – dummy layer shorting them, during layout verification
  – runsets to check sleep transistor is evenly distributed
• Automated filler and decap cells insertion
  – hierarchical approach
  – fillers and decap cells as overlay cells

• Challenges
  – ~300mm², 100 Million transistor design
  – multiple ‘grounds’ due to NMOS-based sleep transistor
  – sleep transistor distribution
Performance verification

• FUB-level pre and post lay static TA with Pathmill™
  – assuming no sleep transistors
  – 8% performance penalty for sleep
• Dynamic simulations for critical signals

• Full chip
  – proprietary tools for tile-level/chip-level timing roll up
  – multiple iterations for the nets not meeting requirements
  – tile-based approach reduced timing closure at chip level
    – tile to tile communication only through mesochronous interface
    – eliminated global wiring to reduce top level verification
      – truly tile-based – zero global wires drawn at top level
Power-performance measurements

**Peak Performance**

- 80°C
  - (0.32 TFLOP) 1GHz
  - (1 TFLOP) 3.16GHz
  - (1.81 TFLOP) 5.67GHz

**Power Efficiency**

- 80°C, N=80
  - 1TFLOP @ 97W
  - 1.07V

**Measured Power**

- 80°C, N=80
  - 1.33TFLOP @ 230W
  - 1TFLOP @ 97W, 1.07V

**Leakage**

- 80°C, N=80
  - 394 GFLOPS

Charts source: Intel (lab measurements on research processor)
Estimated power breakdown

**Tile Power Profile**
- Clock dist.: 11%
- IMEM + DMEM: 21%
- 10-port RF: 4%
- Router + Links: 28%
- Dual FPMACs: 36%

**Communication Power**
- Crossbar: 15%
- Clocking: 33%
- Links: 17%
- MSINT: 6%
- Arbiter + Control: 7%
- Queues + Datapath: 22%

**4GHz, 1.2V, 110°C**
Key learnings

- **Teraflop performance within mainstream power**
  - 1.01TFLOP at 62 watts, peak 19.4 GOPS/watt
- **Tile-based methodology**
  - scalable chip size
- **Fine-grained power management**
  - hierarchical clock gating and sleep transistor techniques
  - 3X measured reduction in standby leakage power
- **Excellent SW performance with message-passing**
  - new instructions, larger memory, wider data ports

- **Key enabler for verification methodologies for future Terascale computing**
  - clock gating and sleep logic add verification complexity
  - tool accuracies at 4GHz+ operation were a concern
  - large design ~300mm\(^2\) with a small team
    - shared design and verification teams