5.7 A 48-Core IA-32 Message-Passing Processor with DVFS in 45nm CMOS

Jason Howard[1], Saurabh Dighe[1], Yatin Hoskote[1], Sriram Vangal[1], David Finan[1], Gregory Ruhl[1], David Jenkins[1], Howard Wilson[1], Nitin Borkar[1], Gerhard Schrom[1], Fabrice Pailet[1], Shailendra Jain[2], Tiju Jacob[2], Satish Yada[2], Sraven Marella[2], Praveen Salihundam[2], Vasantha Erraguntla[2], Michael Konow[3], Michael Riepen[3], Guido Droege[3], Joerg Lindemann[3], Matthias Gries[3], Thomas Apel[3], Kersten Henriss[3], Tor Lund-Larsen[3], Sebastian Steibl[3], Shekhar Borkar[1], Vivek De[1], Rob Van Der Wijngaart[4], Timothy Mattson[5]

[1]Intel, Hillsboro, OR
[2]Intel, Bangalore, India
[3]Intel, Braunschweig, Germany
[4]Intel, Santa Clara, CA
[5]Intel, DuPont, WA

Current developments in microprocessor design favor increased core counts over frequency scaling to improve processor performance and energy efficiency. Coupling this architectural trend with a message-passing protocol helps realize a data-center-on-a-die. The prototype chip (Figs. 5.7.1 and 5.7.7) described in this paper integrates 48 Pentium™ class IA-32 cores [1] on a 6×4 2D-mesh network of tiled core clusters with high-speed I/Os on the periphery. The chip contains 1.3B transistors. Each core has a private 256KB L2 cache (12MB total on-die) and is optimized to support a message-passing-programming model whereby cores communicate through shared memory. A 16KB message-passing buffer (MPB) is present in every tile, giving a total of 384KB on-die shared memory, for increased performance. Power is kept at a minimum by transmitting dynamic, fine-grained voltage-change commands over the network to an on-die voltage-regulator controller (VRC). Further power savings are achieved through active frequency scaling at the tile granularity. Memory accesses are distributed over four on-die DDR3 controllers for an aggregate peak memory bandwidth of 21GB/s at 4× burst. Additionally, an 8-byte bidirectional system interface (SIF) provides 6.4GB/s of I/O bandwidth. The die area is 567mm^2 and is implemented in 45nm high-κ metal-gate CMOS [2].

The design is organized in a 6×4 2D-array of tiles [3] to increase scalability. Each tile is a cluster of two enhanced IA-32 cores sharing a router for inter-tile communication. Cores operate in-order and are two-way superscalar. A 256 entry lookup table (LUT) extension of the 64-entry TLB translates 32-bit virtual addresses to 36-bit physical addresses. The separate L1 instruction and data caches are upsized to 16KB and support both write-through and write-back. Each L1 cache is reinforced by a unified 256KB 4-way write-back L2 cache. The L2 uses a 32-byte line size, matching the cache line size internal to the core, and has a 10-cycle hit latency. The L2 also uses in-line double-error-detection and single-error-correction for improved performance and several programmable sleep modes for power reduction. The L2 cache controller features a time-out-and-retry mechanism for increased system reliability.

Shared memory coherency is maintained through software protocols, such as MPI and OpenMP [4], in an effort to eliminate the communication and hardware overhead required for a memory coherent 2D-mesh. A new message-passing memory type (MPMT) is introduced as an architectural enhancement to optimize data sharing using these software procedures. A single bit in a core's TLB designates MPMT cache lines. The MPMT retains all the performance benefits of a conventional cache line, but distinguishes itself by addressing non-coherent shared memory. All MPMT cache lines are invalidated before reads/writes to the shared memory to prevent a core from working on stale data. A new instruction, MBINV, is added to the core to invalidate all MPMT cache entries in a single cycle. Subsequent reads/writes to invalidated cache lines cause cache misses, forcing the data to be read/written from the shared memory. Figure 5.7.2 illustrates the strict software-control protocol used to maintain cache coherency of the MPMT space. The MPB, used as a distributed on-die shared memory, further optimizes the design by decreasing the latency of memory access. Messages passed through the MPB see a 15× latency improvement over messages sent through DDR3.

The 5-port virtual cut-through router (Fig. 5.7.3) used to create the 2D-mesh network employs a credit-based flow-control protocol. Router ports are packet-switched, have 16-byte data links, and can operate at 2GHz at 1.1V. Each input port has five 24-entry queues, a route pre-computation unit, and a virtual-channel (VC) allocator. Route pre-computation for the outport of the next router is done on queued packets. An XY dimension ordered routing algorithm is strictly followed. Deadlock free routing is maintained by allocating 8 virtual channels (VCs) between 2 message classes on all outgoing packets. VC0 through VC5 are kept in a free pool, while VC6 and VC7 are reserved for request classes and response classes, respectively. Input port and output port arbitrations are done concurrently using a wrapped wave front arbiter. Crossbar switch allocation is done in a single clock cycle on a packet granularity. No-load router latency is 4 clock cycles, including link traversal. Individual routers offer 64GB/s interconnect bandwidth, enabling the total network to support 256GB/s of bisection bandwidth.

Fine-grain power management using voltage/frequency islands (VI/FI) combines the advantages of dynamic voltage and frequency scaling (DVFS) for improving energy efficiency. The die is divided into 8 VIs and 28 FIs (Fig. 5.7.4) to allow tiles and mesh to be independently modulated by software. The on-die VRC manages VI0 to VI6 by receiving commands over the network and interfacing to off-die voltage regulators for dynamic V_{CC} scaling from 1.3 to 0V in 6.25mV steps. VIs for idle cores can be set to 0.7V, a safe voltage for state retention, or completely collapsed to 0V, if retention is unnecessary. Voltage level isolation and translation circuitry allow VIs with active cores to continue execution with no impact from collapsed VIs and provide a clean interface across voltage domains. Individual tile frequencies, FI1 to FI24, can be scaled dynamically by software for further power savings. Tiles receiving frequency change commands react within 7 cycles. Specialized clock crossing FIFOs (CCFs), are used for deterministic synchronization between frequency domains.

Memory transactions are serviced by four DDR3 memory controllers [5] positioned at the periphery of the 2D-mesh network. The controllers feature support for DDR3-800, 1066 and 1333 speed grades and reach 75% bandwidth efficiency with rank and bank interleaving applying closed-page mode. By supporting dual rank and two DIMMs per channel, a system memory of 64GB is realized using 8GB DIMMs. I/O transactions are handled by an 8-byte interface. This interface is double-pumped and can sustain 6.4GB/s bandwidth at 200MHz. A system-level translation circuit converts the interface into the PCIe protocol and communicates over a ×4 PCIe link. A TAP controller with standard test and debug features is included.

All 48 IA-cores boot Linux simultaneously. A software library, similar to MPI, is written to take advantage of both the message-passing features and the power optimizations. When compiled using the library, HPLinpack (HPL) shows 8% average run time improvement and 16% better average power (Fig. 5.7.5). When operating under typical conditions, 1.14V and 1GHz, power consumption is measured to be 125W at 50°C. With DVFS, 0.7V and 125MHz, measured power is reduced by 80% to 25W at 50°C. Maximum tile and mesh frequencies versus V_{CC} are plotted in Fig. 5.7.6. A chip micrograph is shown in Fig. 5.7.7.

Acknowledgements:
The authors thank J. Rattner, J. Schutz, M. Haycock, G. Taylor, and J. Held for their leadership, encouragement, and support; and entire mask design team for chip layout.

References:
[1] J. Schutz, "A 3.3V 0.6um BiCMOS Superscalar Microprocessor", *ISSCC Dig. Tech. Papers*, pp. 202-203, Feb., 1994.
[2] K. Mistry, C. Allen, C. Auth, et al., "A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging", *IEDM Dig. Tech. Papers*, Dec. 2007.
[3] S. Vangal, J. Howard, G. Ruhl, et al., "An 80-Tile 1.28TFLOPS Network-on-Chip in 65nm CMOS", *ISSCC Dig. Tech. Papers*, pp. 98-99, Feb., 2007.
[4] L. Smith and M. Bull, "Development of hybrid mode MPI/OpenMP applications", *Scientific Programming*, Vol. 9, No 2-3, 83-98, 2001.
[5] R. Kumar, G. Hinton, "A Family of 45nm IA Processors", *ISSCC Dig. Tech. Papers*, pp. 58-59, Feb., 2009.

 978-1-4244-6034-2/10/$26.00

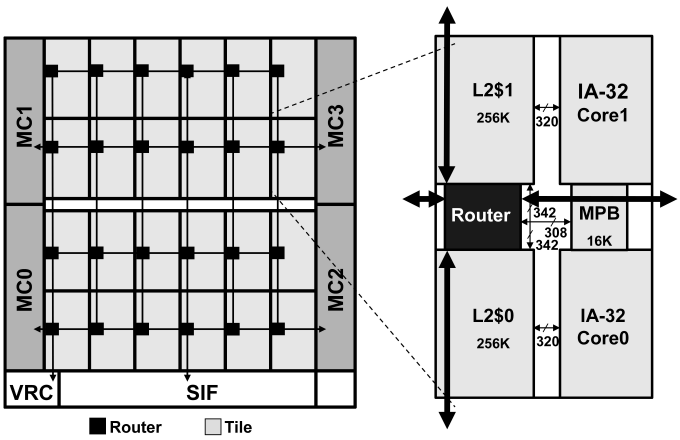


Figure 5.7.1: Block diagram and tile architecture.

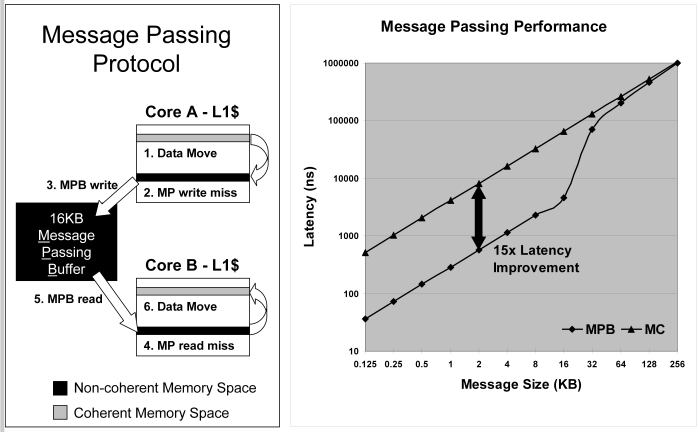
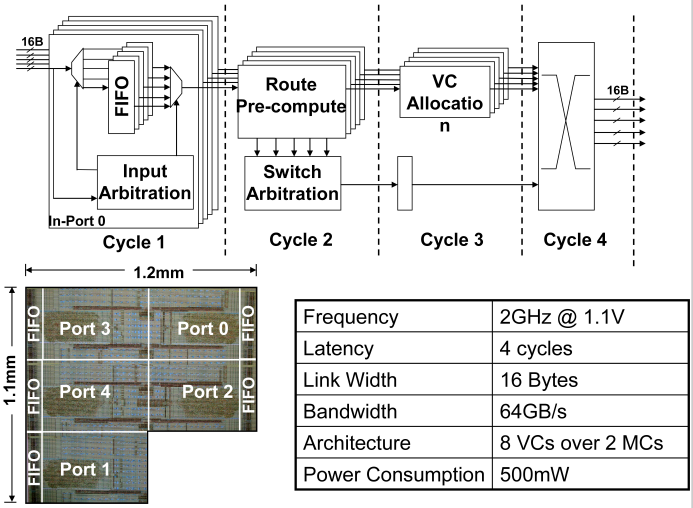


Figure 5.7.2: Message passing vs. DDR3-800 speed grade.



Frequency	2GHz @ 1.1V
Latency	4 cycles
Link Width	16 Bytes
Bandwidth	64GB/s
Architecture	8 VCs over 2 MCs
Power Consumption	500mW

Figure 5.7.3: Router architecture and latency.

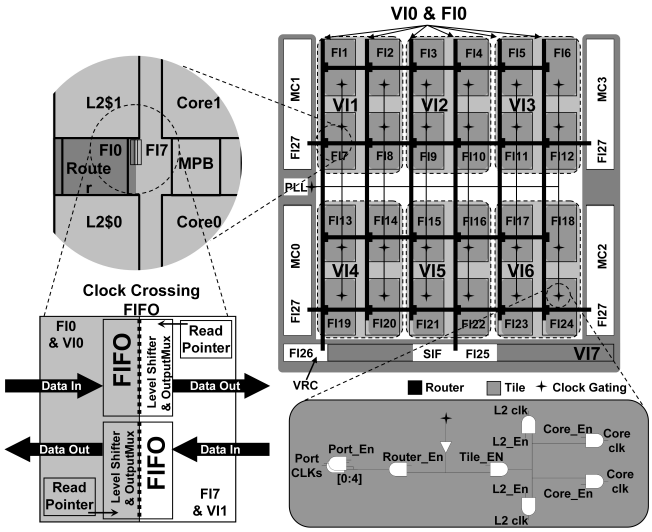


Figure 5.7.4: Voltage/Frequency Islands, Clock Crossing FIFOs and Clock Gating.

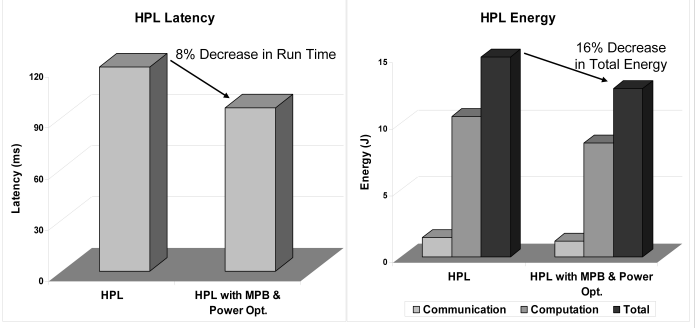


Figure 5.7.5: Measured HPL performance and energy benefits.

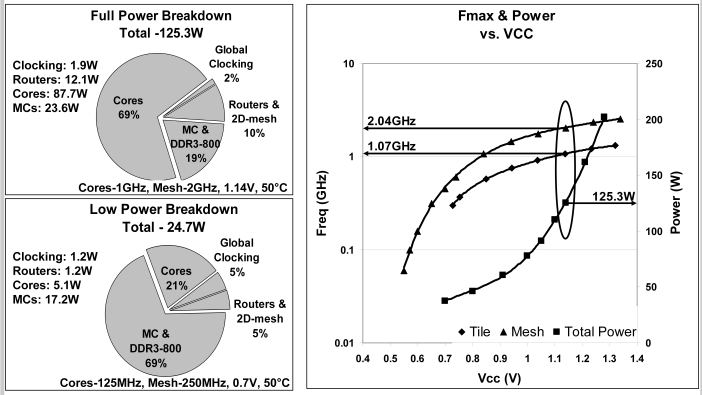
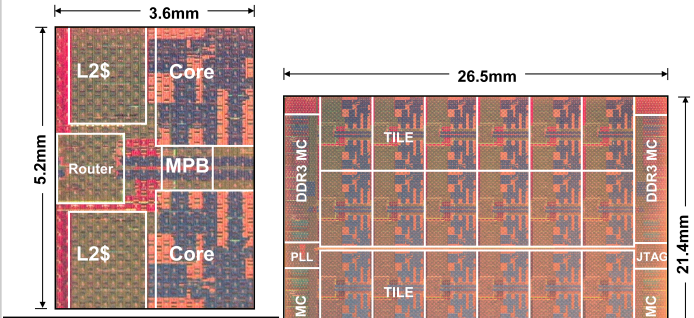


Figure 5.7.6: Measured full-chip power breakdowns, and maximum frequency (Fmax) and power versus VCC.



Technology	45nm Process
Interconnect	1 Poly, 9 Metal (Cu)
Transistors	Die: 1.3B, Tile: 48M
Tile Area	18.7mm^2
Die Area	567.1mm^2
Signals	970
Package	1567 pin LGA package

Figure 5.7.7: Full-Chip and tile micrograph and characteristics.

978-1-4244-6034-2/10/$26.00