Teraflops Prototype Processor with 80 Cores

Yatin Hoskote, Sriram Vangal, Saurabh Dighe, Nitin Borkar, Shekhar Borkar

Microprocessor Technology Labs,
Intel Corp.

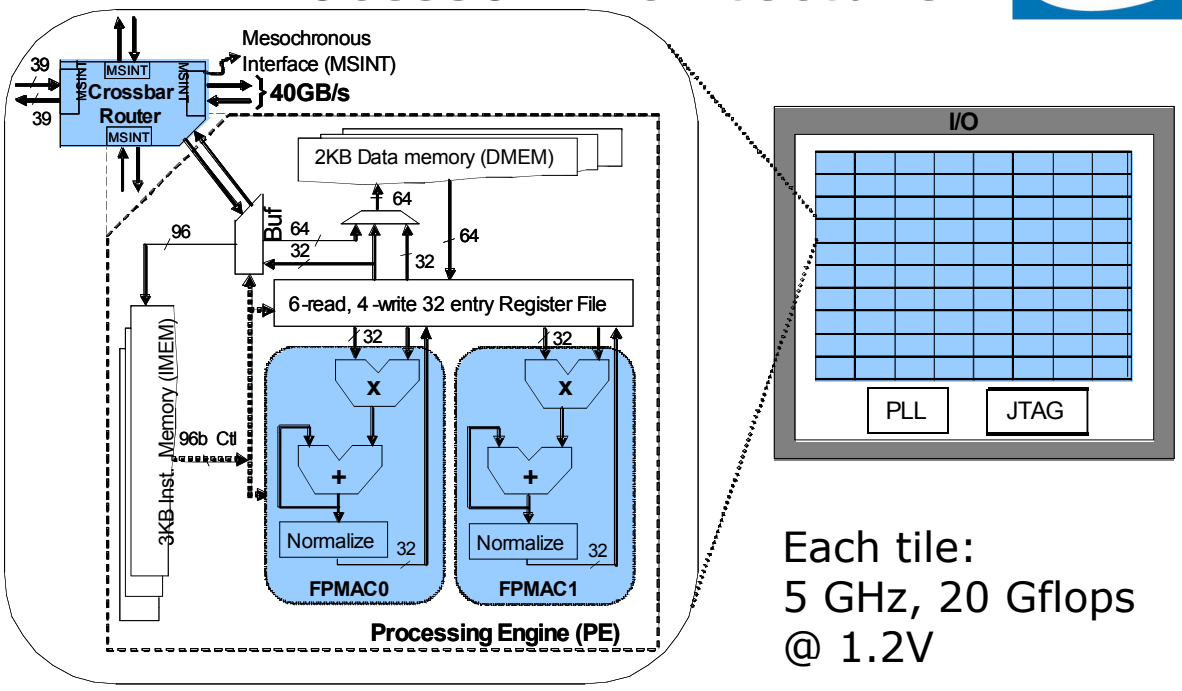
Agenda

- Goals
- Architecture overview
- Processing engine
- Interconnect fabric
- Power management
- Application mapping
- Experimental results

Goals

- Build Intel's first Network-on-Chip (NoC) research prototype
- Show teraflops performance under 100W
- Develop NoC design methodologies
  - Tiled design
  - Mesochronous (phase tolerant) clocking
  - Fine grain power management
- Prototype key NoC technologies
  - On-die 2D mesh fabric
  - 3D stacked memory

Processor Architecture



Each tile:
5 GHz, 20 Gflops
@ 1.2V

160 SP FP engines in 8x10 2D mesh

Instruction Set

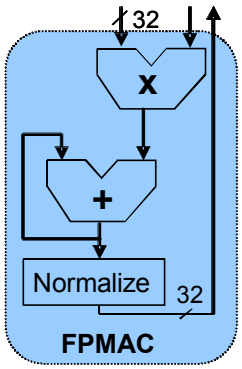
FPU (2)	LOAD/STORE	SND/RCV	PGM FLOW	SLEEP

- 96-bit instruction word, up to 8 operations/cycle

Instruction Type	Latency (cycles)
FPU	9
LOAD/STORE	2
SEND/RECEIVE	2
JUMP/BRANCH	1
STALL/WAIT FOR DATA	N/A
NAP/WAKE	1

SP FP Processing Engine

- Fast, single-cycle multiply-accumulate algorithm
- Key optimizations
  - Multiplier output is in carry-save format and uses 4-2 carry-save adders, removing expensive carry-propagate adders from the critical path
  - Accumulation is performed in base 32, converting expensive variable shifters in the accumulate loop to constant shifters
  - Costly normalization step is moved outside the accumulate loop



- Ref: S. Vangal, et al. IEEE JSSC, Oct. 2006

Accumulation in 15 Fanout-of-4 stages

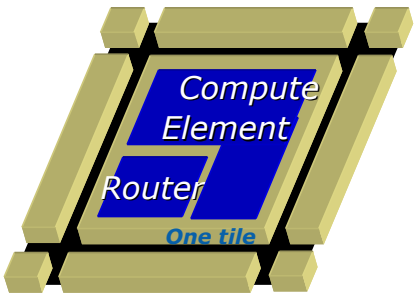
Sustained 2FLOPS per cycle

Agenda

- Goals
- Architecture overview
- Processing engine
- Interconnect fabric
- Power management
- Application mapping
- Experimental results

2D Mesh Interconnect

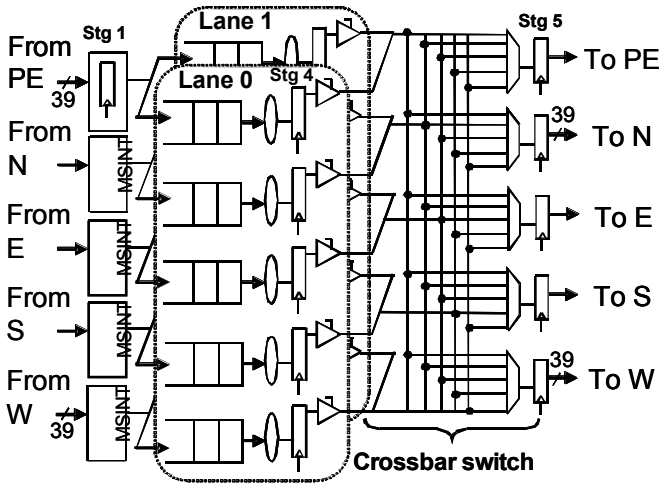
- 4 byte wide data links, 6 bits overhead
- Source directed, wormhole routing
- Two virtual lanes
- 5 port, fully non-blocking router
- 5GHz operation @1.2V
- 320GB/s bisection B/W
- 5 cycle fall-through latency
- On/off flow control

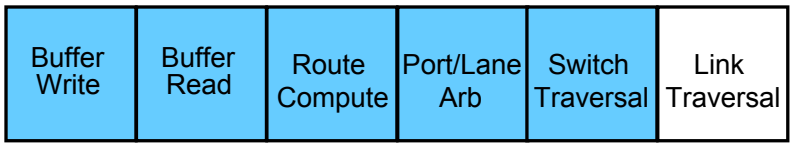


High bandwidth, low latency fabric

Router Architecture

- Five router pipe stages
- Input buffered
- Shared crossbar switch
- Distributed dual phase arbitration
- Ref: S. Vangal, et al., VLSI 2007





Router Pipe stages

Fabric Key Features

- Double pumped crossbar switch
  - Dual-edge triggered FFs on alternate data bits
  - Crossbar channel width reduced by 50% leading to 0.34mm2 router area
  - Ref: S. Vangal, et al. VLSI, 2005
- Mesochronous clocking
  - Phase-tolerant FIFO based synchronizers at interfaces between tiles
  - Enables low power, scalable global clock distribution
  - 2W global clock distribution power @ 1.2V 5GHz
  - Overhead of synchronizers: 6% of router power and 1-2 clock cycles in latency

Agenda

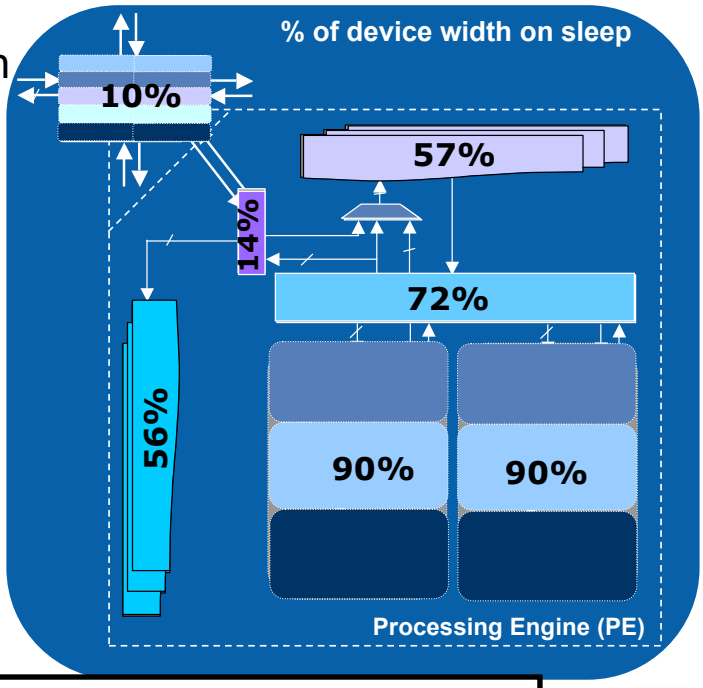
- Goals
- Architecture overview
- Processing engine
- Interconnect fabric
- Power management
- Application mapping
- Experimental results

Power Management Hooks

- Extensive use of sleep transistors
- Activation of sleep transistors done dynamically or statically
- Dynamic through special instructions or packets based on workload
  - NAP/WAKE: PE can put its FPMAC engines to sleep or wake them up
  - PESLEEP/PEWAKE: PE can put another PE to sleep or wake it up for processing tasks by sending sleep or wake packets
- Static through scan
  - Entire PE or individual router ports

Tile Sleep Regions

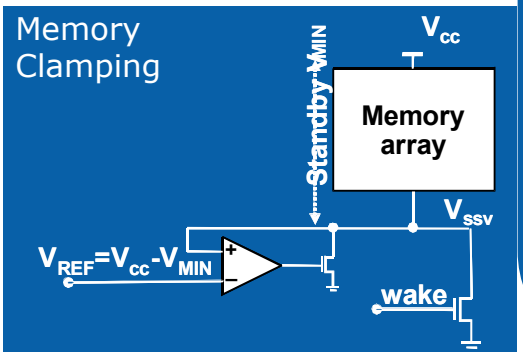
- 21 sleep regions with independent control
  - 74% of transistor device width
- Dynamic sleep
  - Individual FPMACs, cores or tiles
- Static sleep control
  - Scan chain
- Clock gating
  - Works with sleep

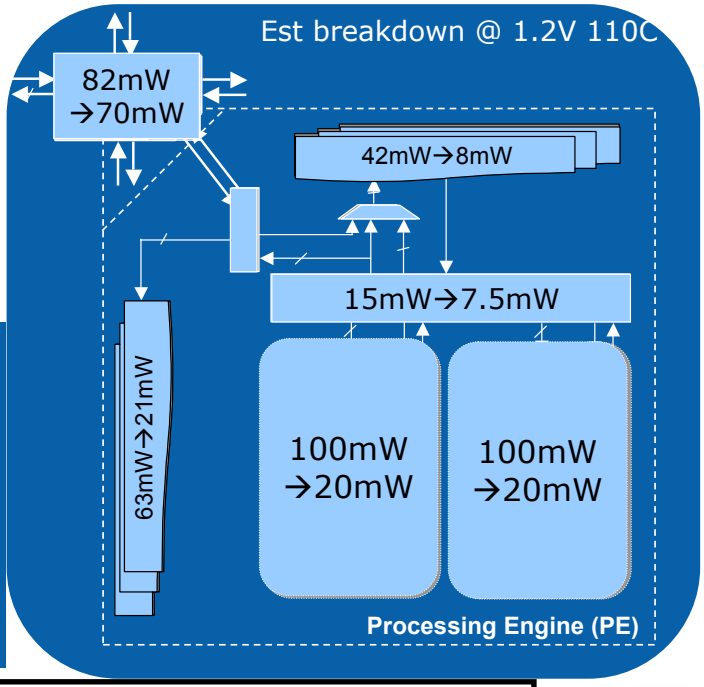


Fine grain power management

Leakage Savings

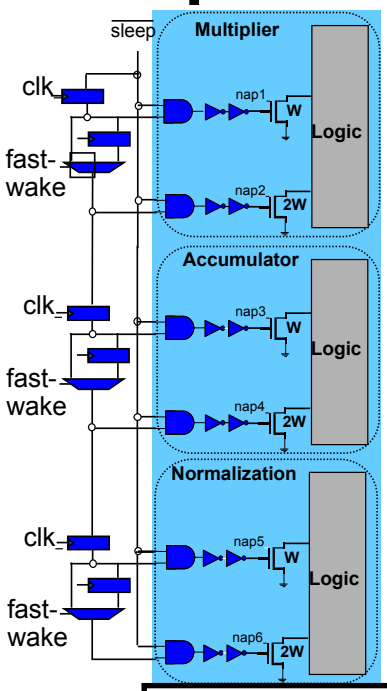
- Dynamic sleep
  - Total measured idle power 13W @ 1.2V
- Regulated sleep for memory arrays
  - State retention

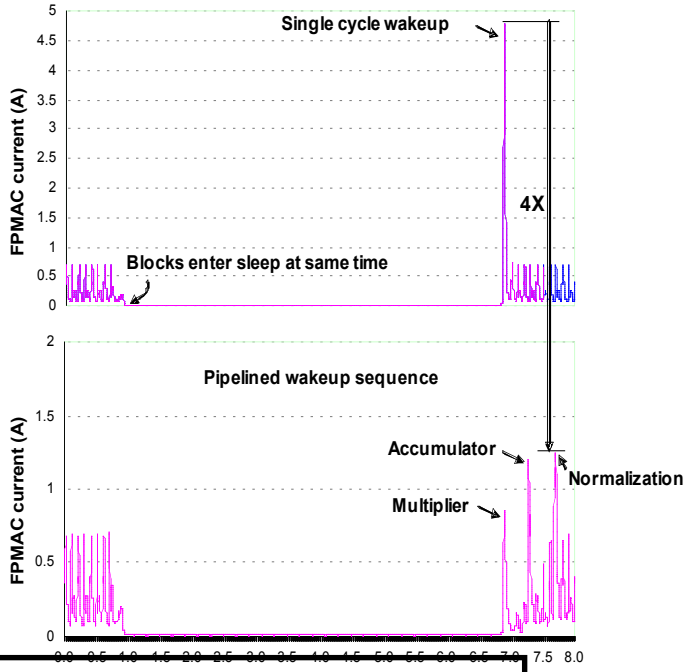




2X-5X leakage power reduction

Pipelined Wakeup from Sleep

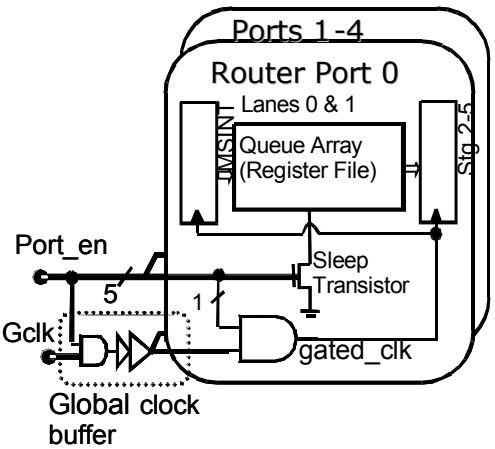


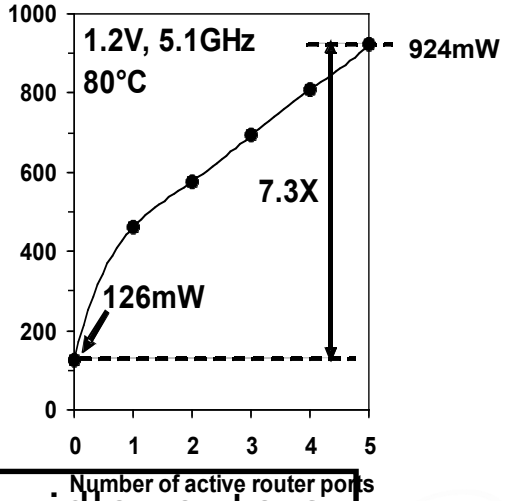


4X reduction in peak wakeup current

Router Power Management

- Activity based power management
- Individual port enables
  - Queues on sleep and clock gated when port idle

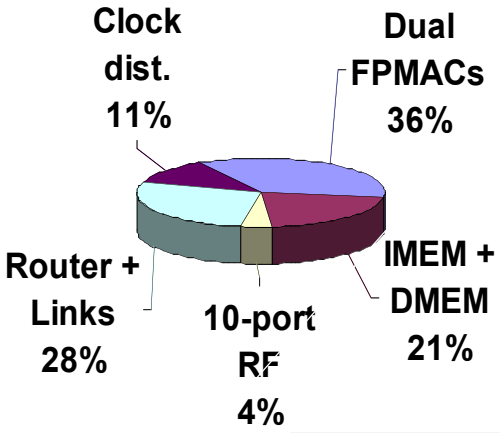




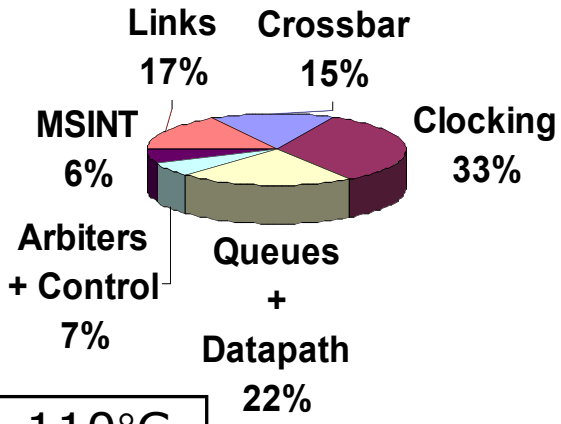
7X power reduction for idle routers

Estimated Power Breakdown

Tile Power Profile



Communication Power Profile



4GHz, 1.2V, 110°C

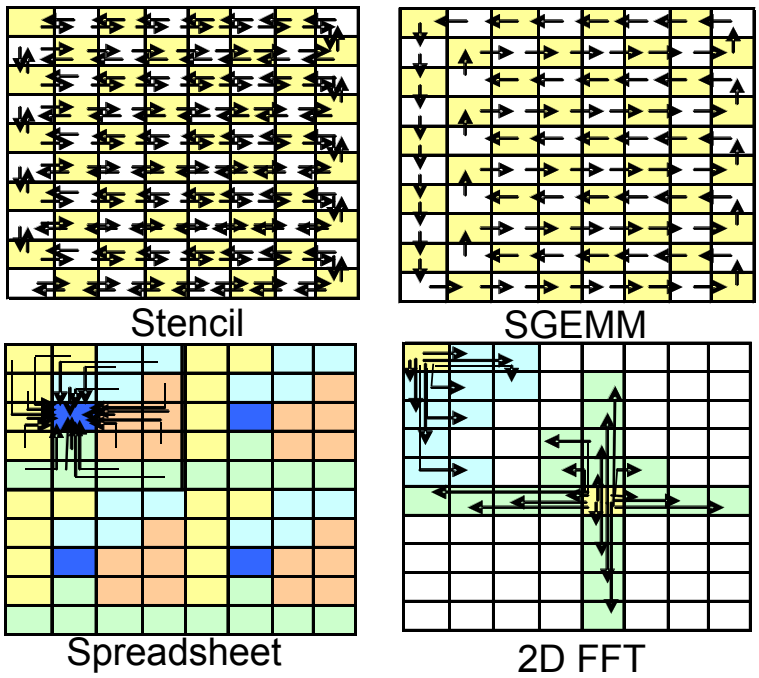
Agenda

- Goals
- Architecture overview
- Processing engine
- Interconnect fabric
- Power management
- Application mapping
- Experimental results

Application Kernels

- Four kernels mapped on to architecture
  - Stencil 2D heat diffusion equation
  - SGEMM for 100x100 matrices
  - Spreadsheet doing weighted sums
  - 64 point 2D FFT (using 64 tiles)
- Kernels were hand coded in assembly code and manually optimized
- Sized to fit in the on-chip local data memory
  - Instruction memory not a limiter

Communication Patterns



Communication overlapped with computation

Agenda

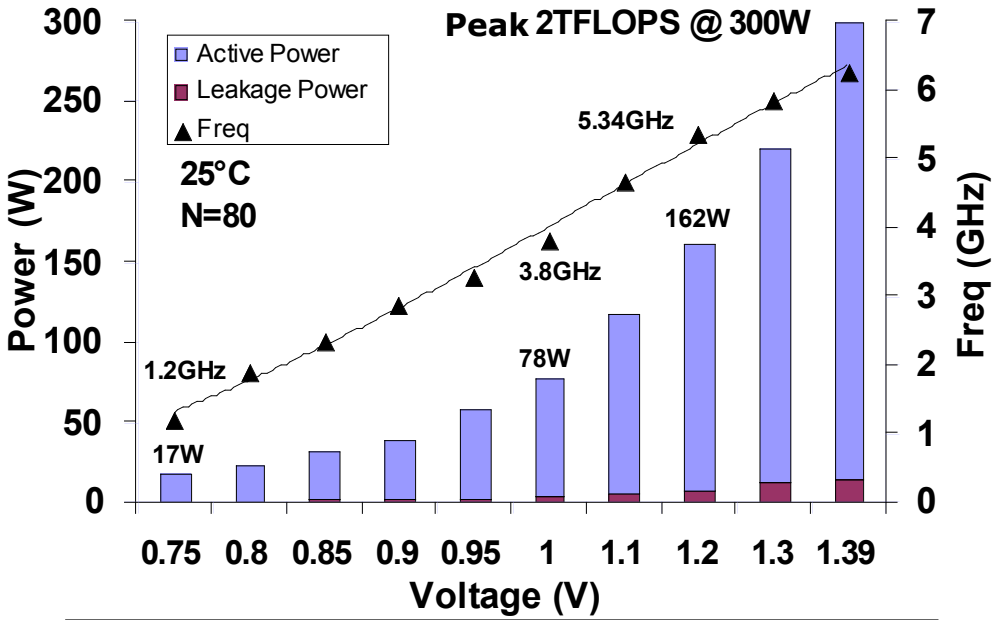
- Goals
- Architecture overview
- Processing engine
- Interconnect fabric
- Power management
- Application mapping
- Experimental results 

Application Performance

Application Kernels	FLOP count	TFLOPS @ 4.27GHz	% Peak TFLOPS	Tiles used
Stencil	358K	1.00	73.3%	80
SGEMM: Matrix Multiplication	2.63M	0.51	37.5%	80
Spreadsheet	62.4K	0.45	33.2%	80
2D FFT	196K	0.02	2.73%	64

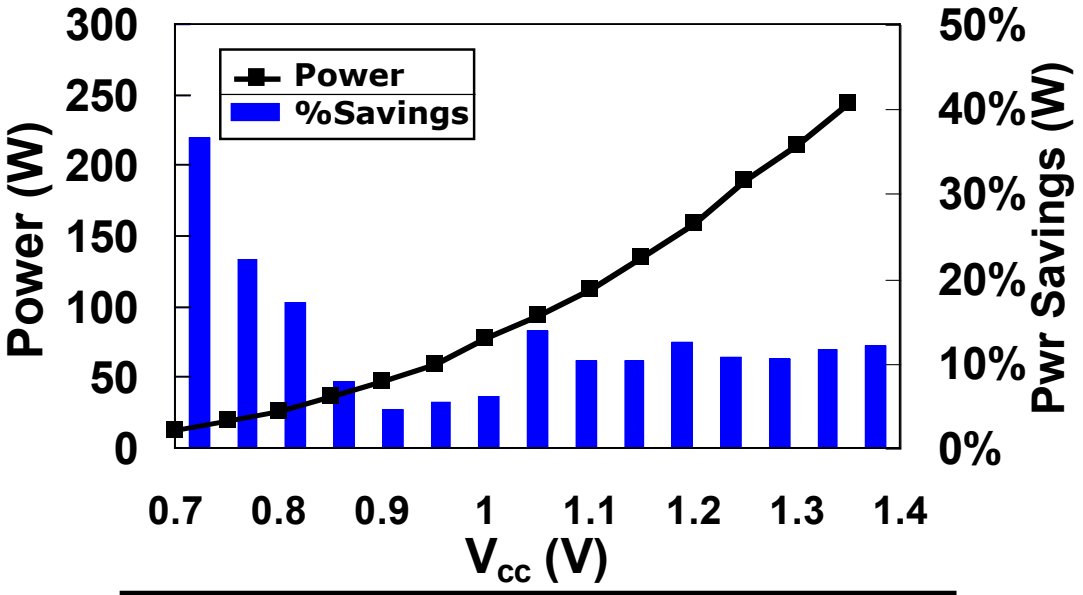
1.07V, 4.27GHz operation 80°C

Frequency vs Power



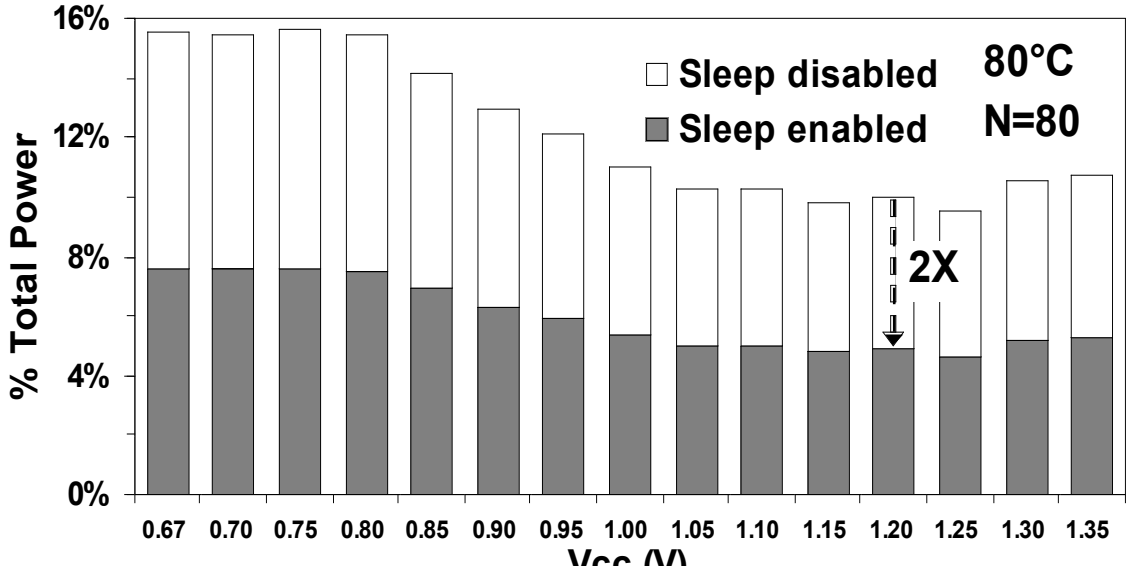
Stencil app peak efficiency:
6 GFLOPS/W to 22 GFLOPS/W

Measured Power Savings



Stencil app: Savings in power with
router power management on

Leakage Savings

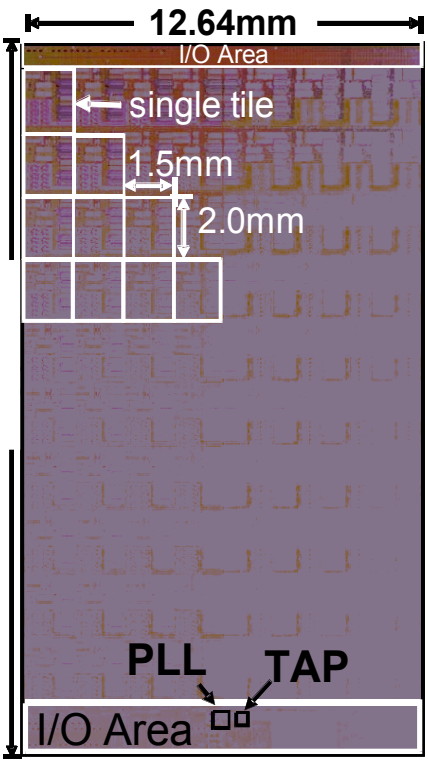


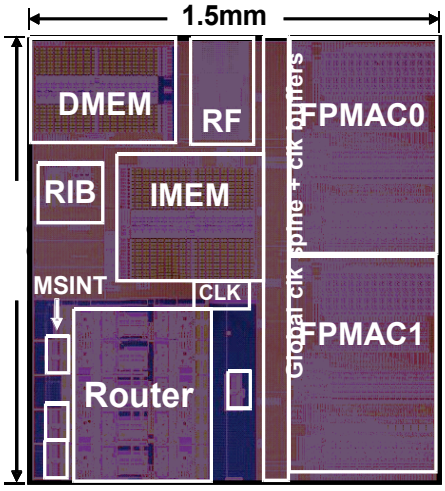
Measured maximum leakage power reduction

Design Tradeoffs

- Core count
  - Determined by die size constraints (300 sqmm) and performance per watt requirements (10 GFLOPS/W)
- Router link width
  - 4 byte data path enables single flit transfer of single precision word
  - Crossbar area and power scales quadratically with link width
- Mesochronous implementation
  - Single clock source with scalable, low power clock distribution
- Backend design and validation benefits
  - Tiled methodology enabled rapid design by small team (less than 400 person-months)
  - Functional first silicon in 2 hours

Die Photo and Chip Details





Technology	65nm CMOS Process
Interconnect	1 poly, 8 metal (Cu)
Transistors	100 Million
Die Area	275mm^2
Tile area	3mm^2
Package	1248 pin LGA, 14 layers, 343 signal pins

Summary

- An 80-Core NoC architecture in 65nm CMOS
  - 160 high-performance FPMAC engines
  - Fast, single-cycle accumulator
  - Low latency, compact router design
  - High bandwidth 2D mesh interconnect
- TFLOPS level performance at high power efficiency
  - Fine grain power management
  - Chip dissipates 97W at 1TFLOPS (Stencil app)
  - Measured energy efficiency of 6 to 22 GFLOPS/W
- Demonstrated peak performance up to 2TFLOPS
- Building blocks for future peta-scale computing

Acknowledgements

- Implementation
  - Circuit Research Lab Advanced Prototyping team (Hillsboro, OR and Bangalore, India)
- Application kernels
  - Software Solutions Group (Santa Clara, CA)
  - Application Research Labs (DuPont, WA)
- PLL design
  - Logic Technology Development (Hillsboro, OR)
- Package design
  - Assembly Technology Development (Chandler, AZ)