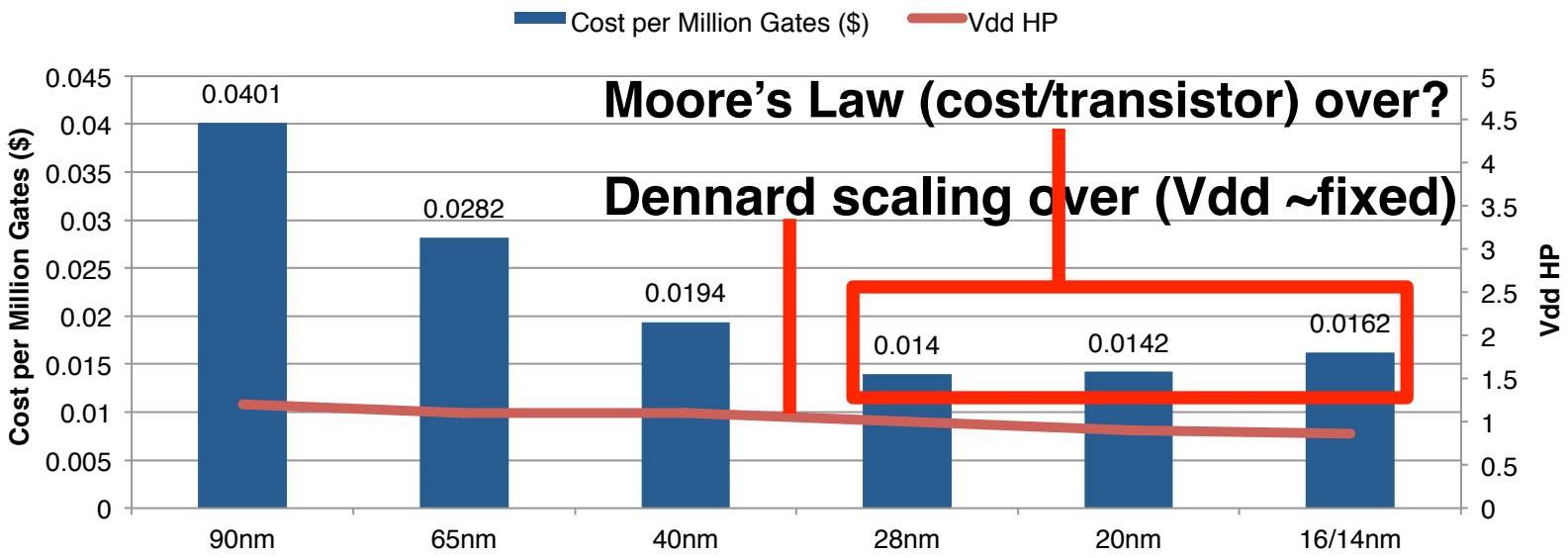
A 45nm 1.3GHz 16.7 Double-Precision GFLOPS/W RISC-V Processor with Vector Accelerators

Yunsup Lee[1], Andrew Waterman[1], Rimas Avizienis[1],
Henry Cook[1], Chen Sun[1,2],
Vladimir Stojanovic[1,2], Krste Asanovic[1]

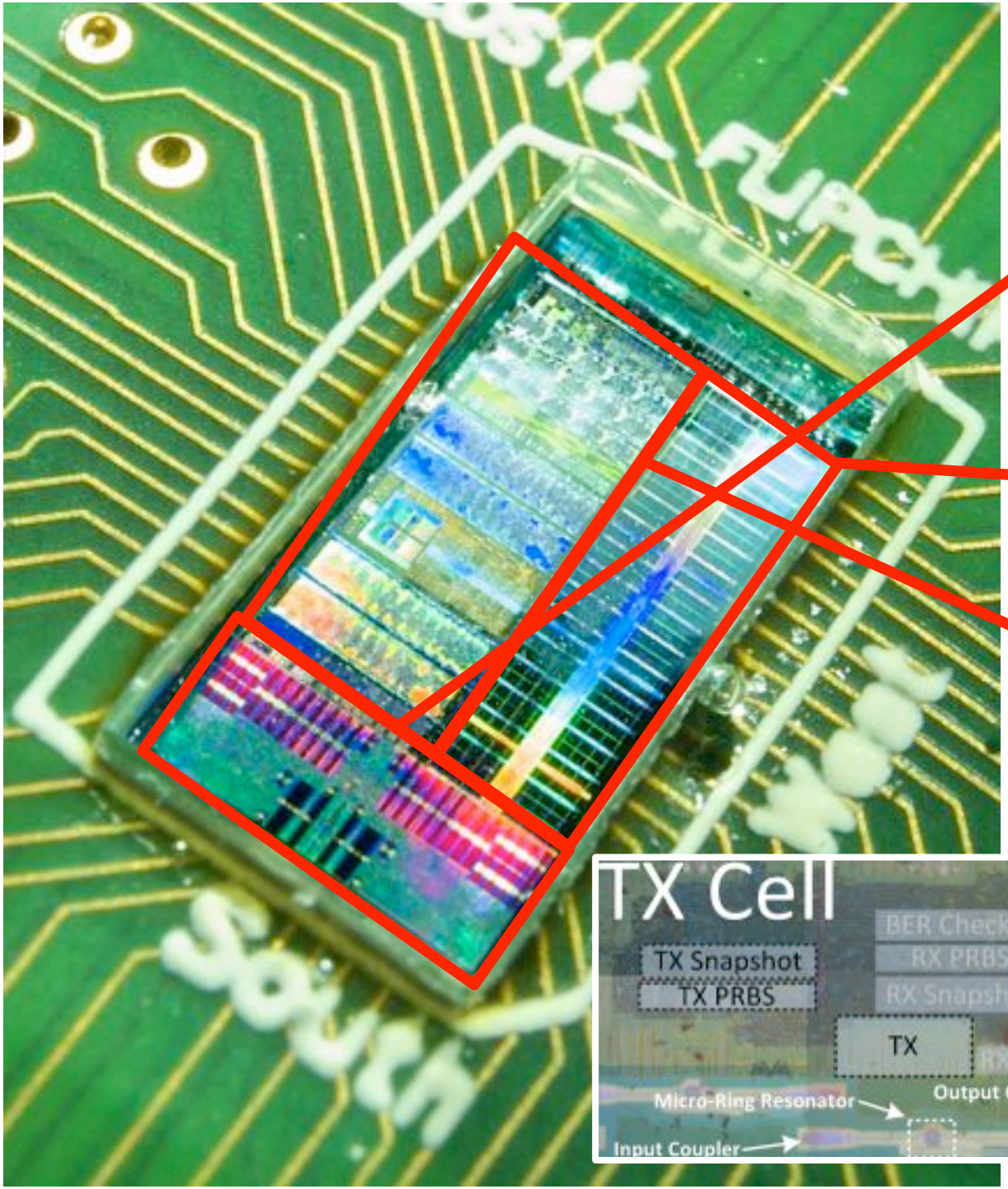
[1]University of California, Berkeley
[2]Massachusetts Institute of Technology

Upheaval in Computer Design



- Energy efficiency constrains everything
  - Incorporate specialized and heterogeneous accelerators into general-purpose processors
  - Write processor generators to express a design space, and do vertically-integrated design space exploration extensively

Source
[1] Why migration to 20nm bulk CMOS and 16/14nm FINFETS is not the best approach for semiconductor industry, IBS, Handel Jones, 2014.
[2] International Technology Roadmap of Semiconductors (ITRS)

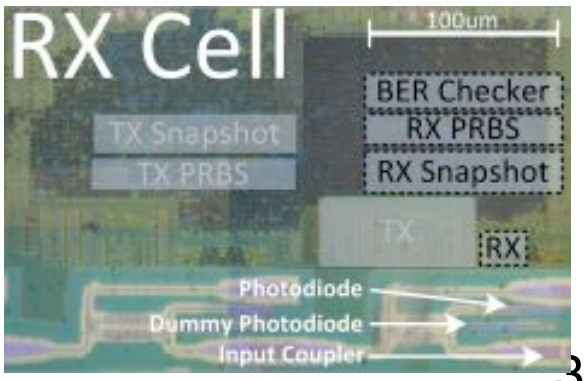


3mm X 6mm Chip
Fabricated in 45nm SOI
75m+ transistors

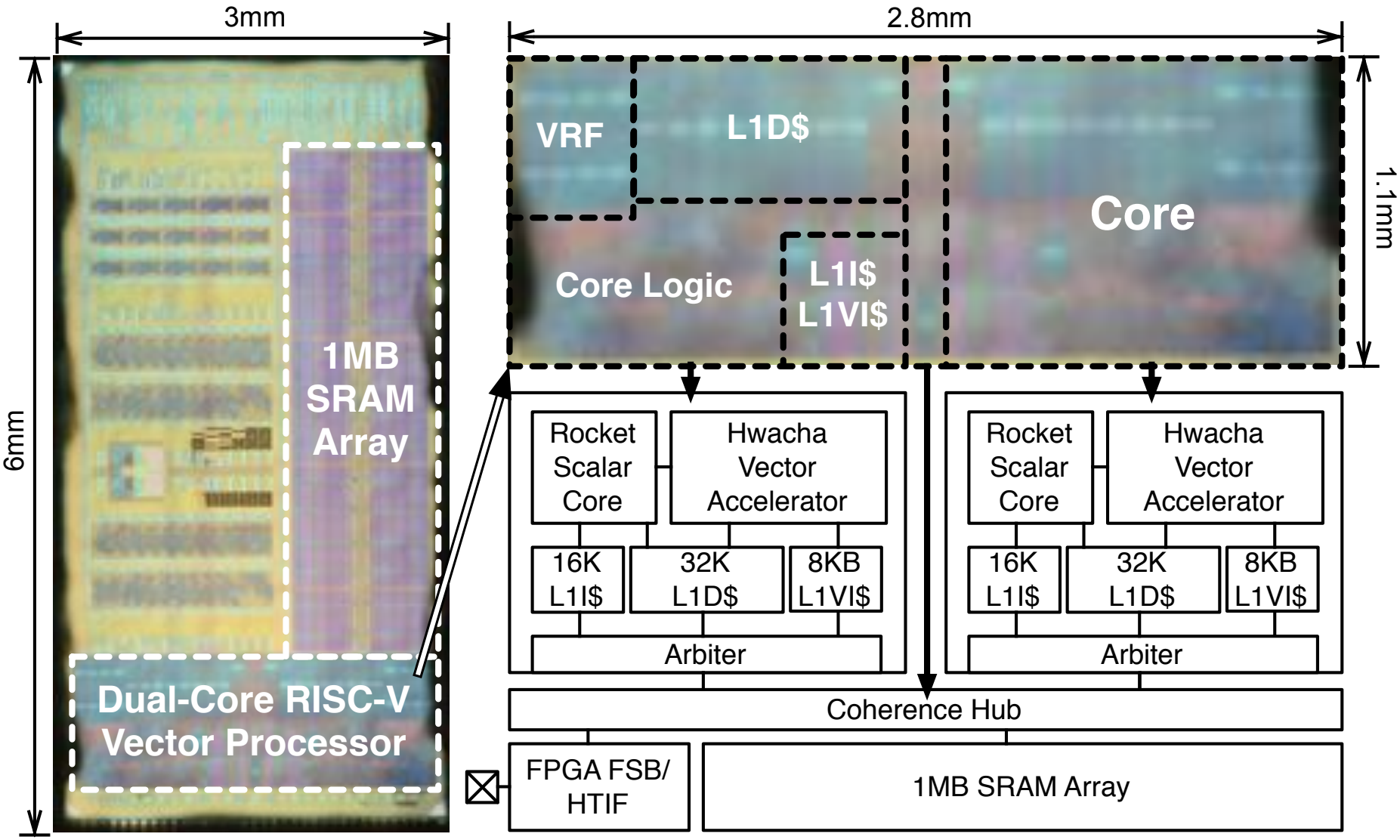
Dual-Core RISC-V Processor with Vector Accelerators

1MB SRAM Memory Structure for Testing

Monolithically-Integrated Silicon Photonic Links
Transmitter: Wade OFC'14
Receiver: Georgas VLSI'14

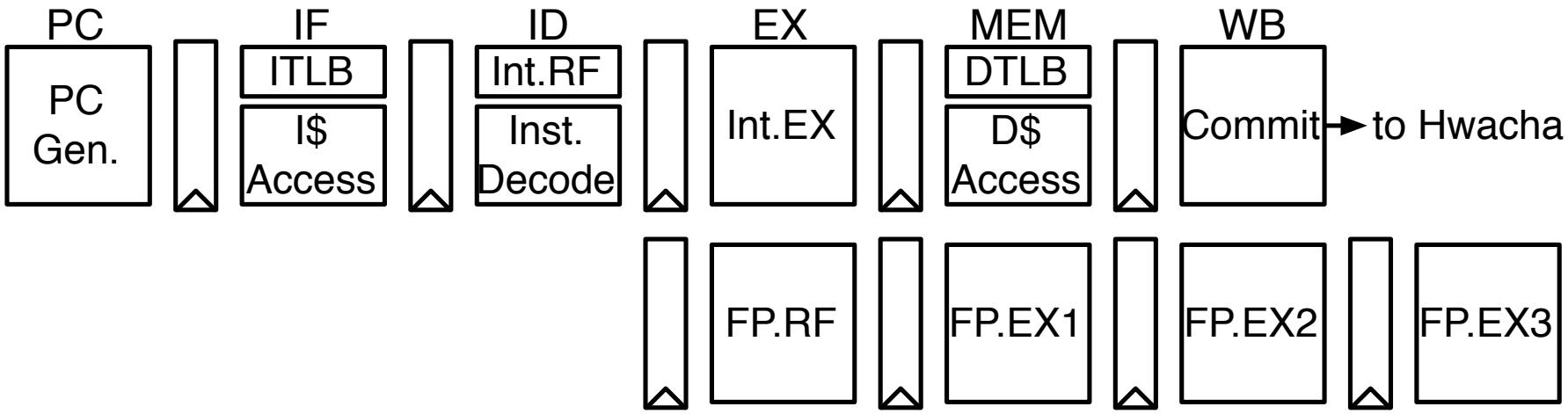


Chip Architecture



- RISC-V is a new, open, and completely free general-purpose ISA
  - Developed at UC Berkeley
- RISC-V designed to be flexible and extensible
  - Better integrate accelerators with host cores
- RISC-V software ecosystem
  - binutils, GCC, Newlib, glibc, GDB, LLVM, Linux, QEMU
- External users contributing to ecosystem

Rocket Scalar Core



- 64-bit 6-stage single-issue in-order pipeline
- Design minimizes impact of long clock-to-output delays of compiler-generated RAMs
- 64-entry BTB, 256-entry BHT, 2-entry RAS
- MMU supports page-based virtual memory
- IEEE 754-2008-compliant FPU
  - Supports SP, DP FMA with hw support for subnormals

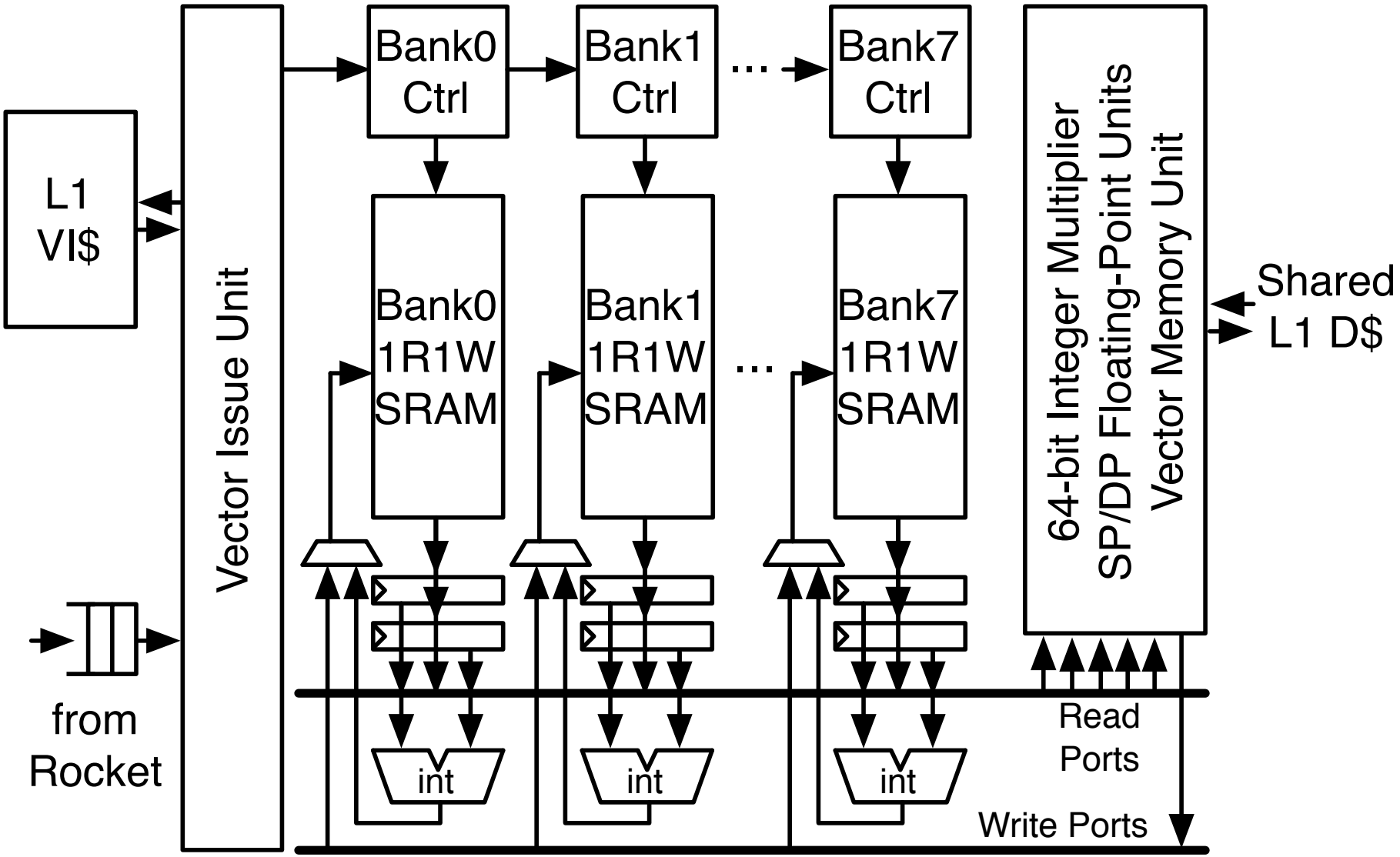
ARM Cortex-A5 vs. RISC-V Rocket

Category	ARM Cortex-A5	RISC-V Rocket
ISA	32-bit ARM v7	64-bit RISC-V v2
Architecture	Single-Issue In-Order	Single-Issue In-Order 6-stage
Performance	1.57 DMIPS/MHz	1.72 DMIPS/MHz
Process	TSMC 40GPLUS	TSMC 40GPLUS
Area w/o Caches	0.27 mm^2	0.14 mm^2
Area with 16K Caches	0.53 mm^2	0.39 mm^2
Area Efficiency	2.96 DMIPS/MHz/mm^2	4.41 DMIPS/MHz/mm^2
Frequency	>1GHz	>1GHz
Dynamic Power	<0.08 mW/MHz	0.034 mW/MHz

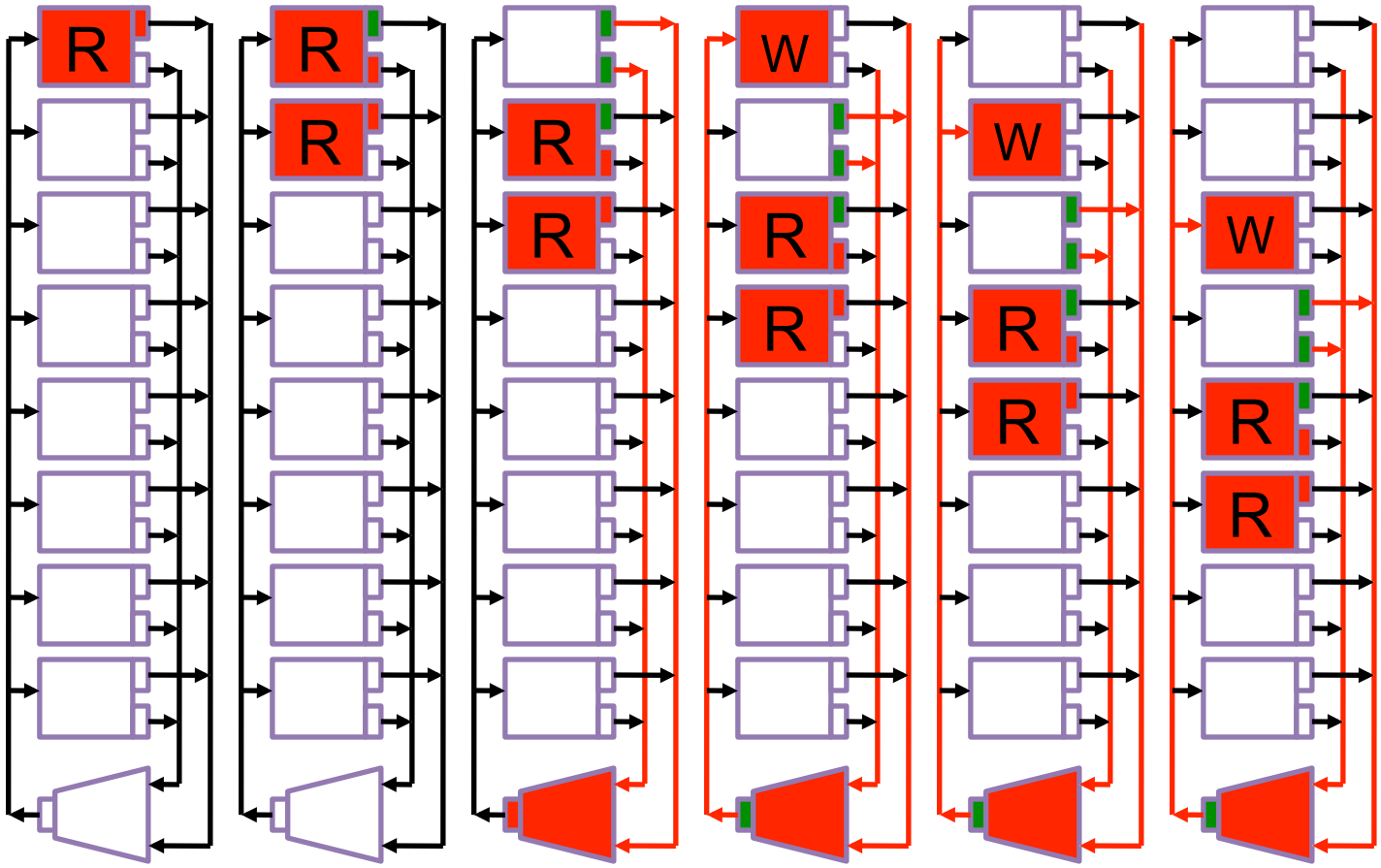
– PPA reporting conditions
  - 85% utilization, use Dhrystone for benchmark, frequency/ power at TT 0.9V 25C, all regular VT transistors

– 10% higher in DMIPS/MHz, 49% more area-efficient

Hwacha Vector Accelerator



Bank Execution Diagram



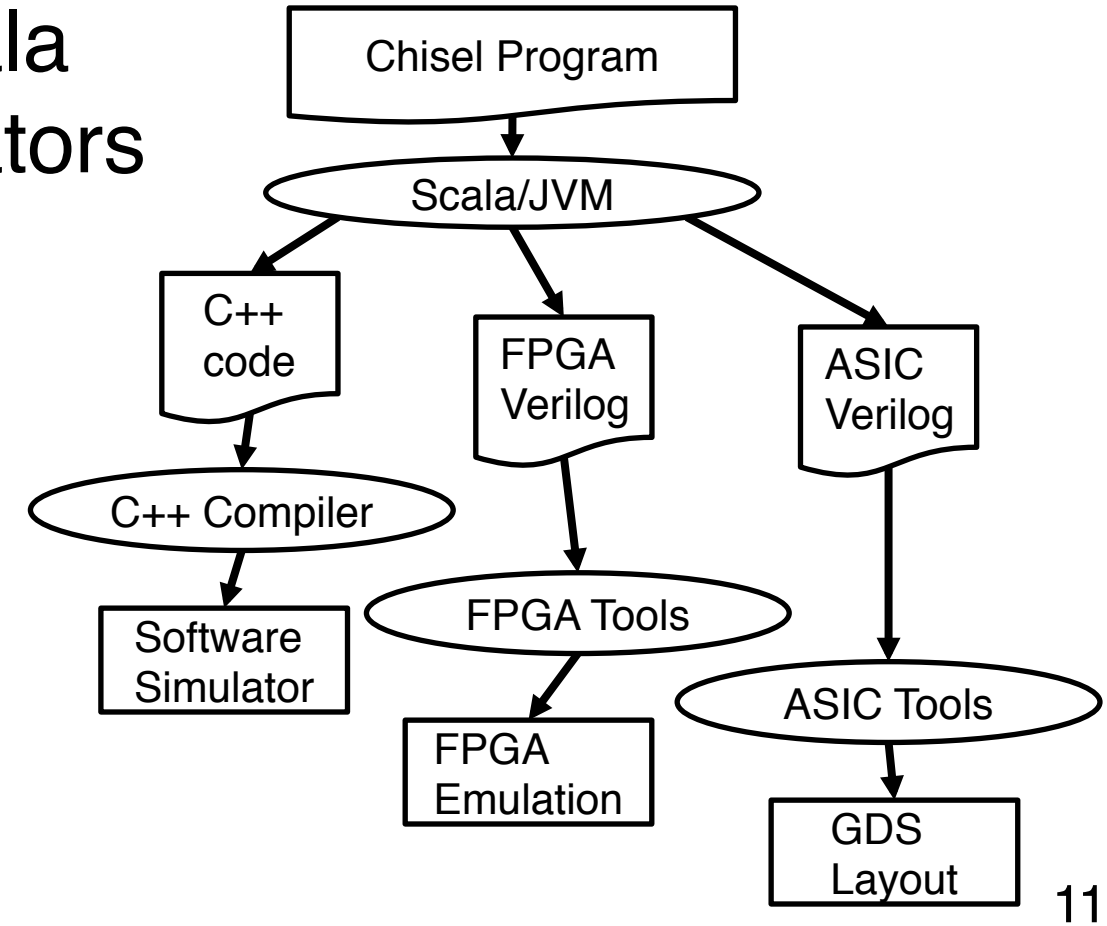
– After a 2-cycle initial startup latency, the banked RF is effectively able to read out 2 operands/cycle.

Processor Generators

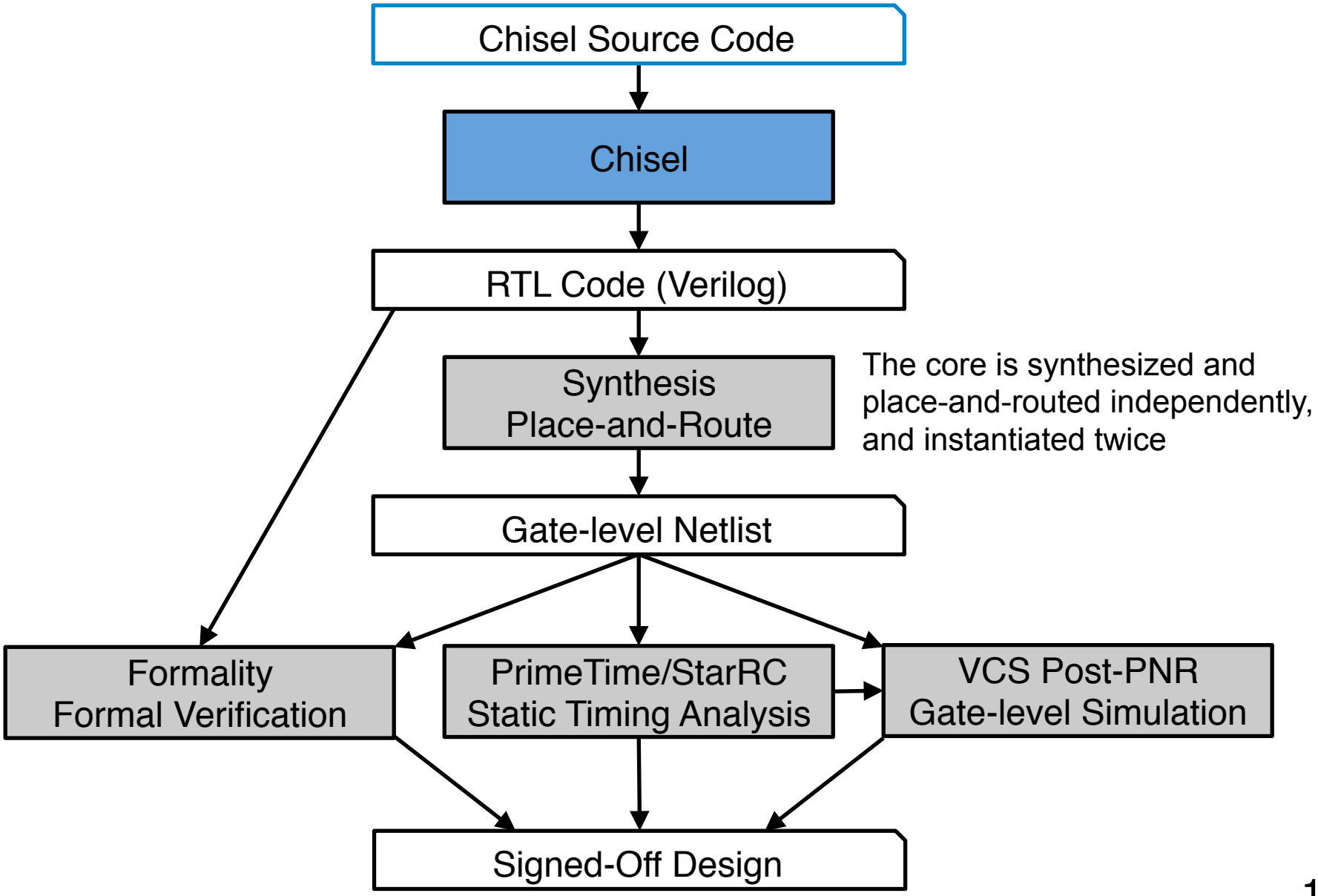
- Express hardware as highly parameterized generators
  - Helps tune the design under different performance, power, and area constraints
- Parameters include:
  - number of cores
  - cache sizes, associativity, number of TLB entries, cache-coherence protocol
  - number of floating-point pipeline stages
  - width of off-chip I/O, and more

Writing Generators with Chisel

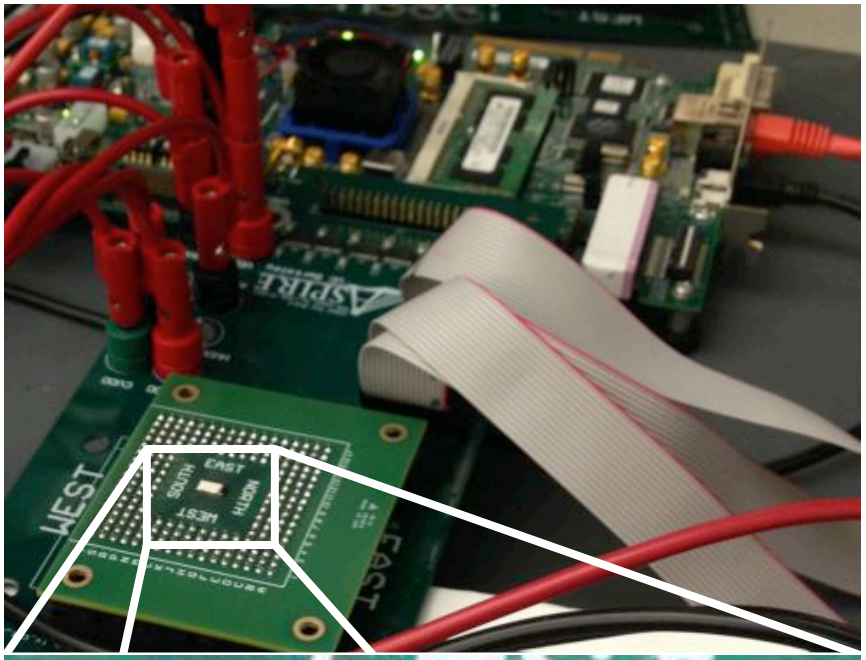
- RTL generator written in Chisel
  - HDL embedded in Scala
- Full power of Scala for writing generators
  - object-oriented programming, functional programming

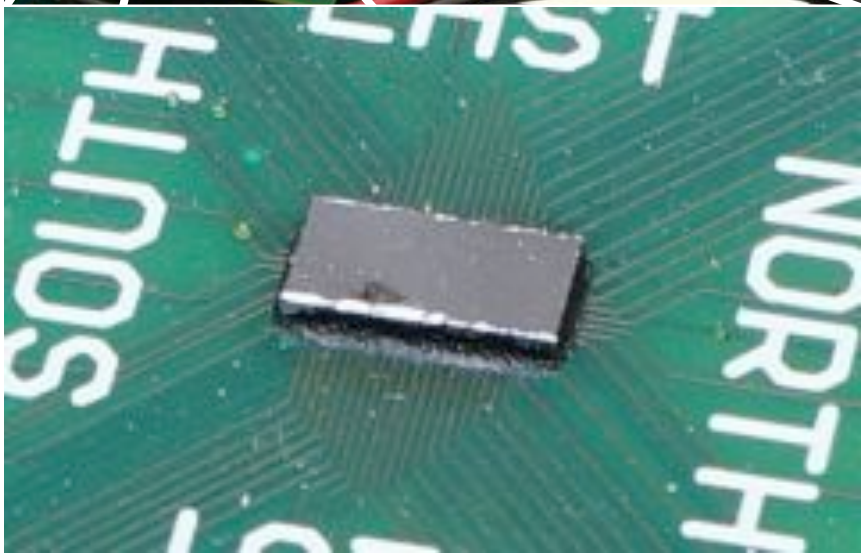


Physical Design Flow



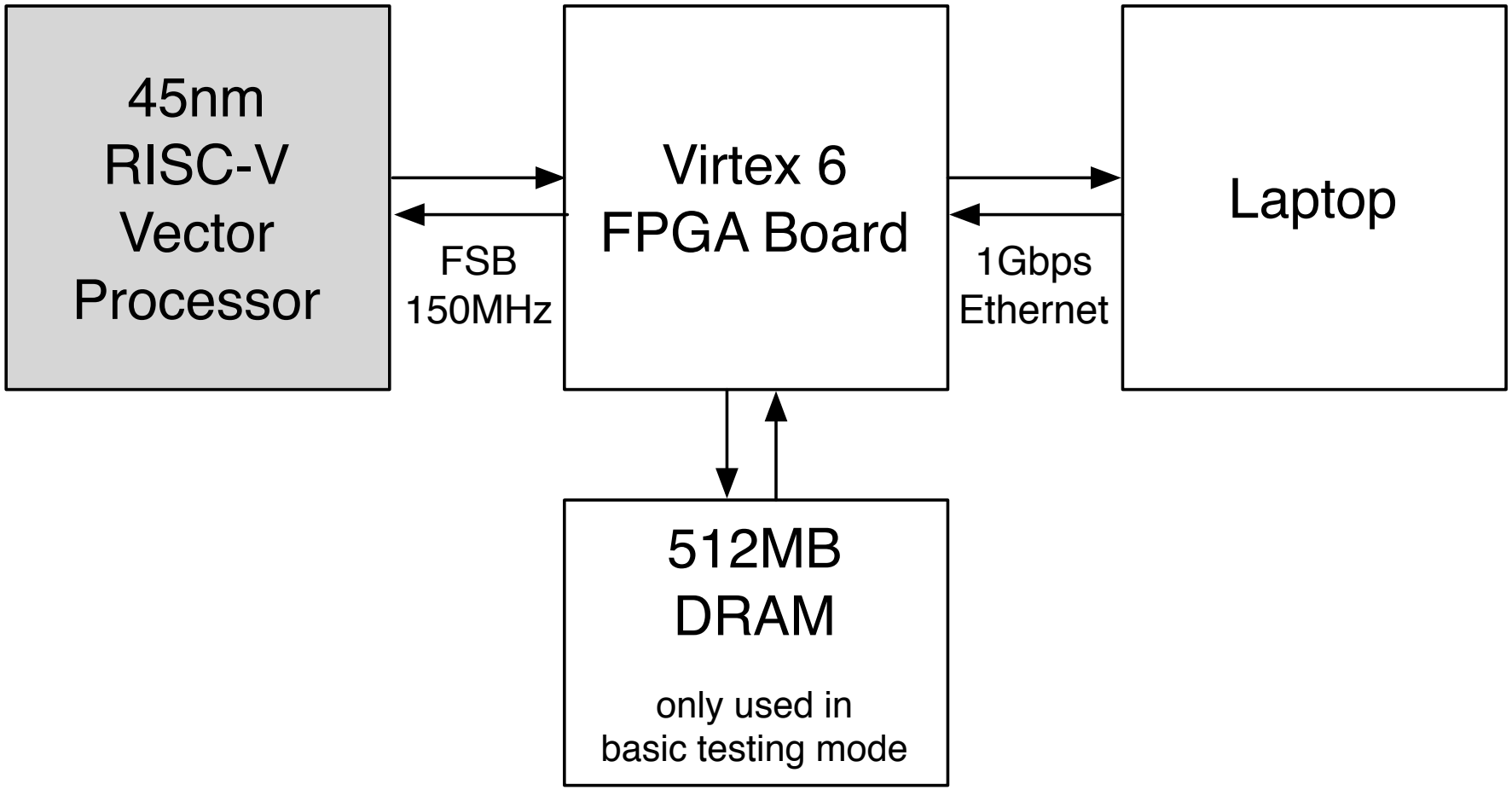
Chip Results





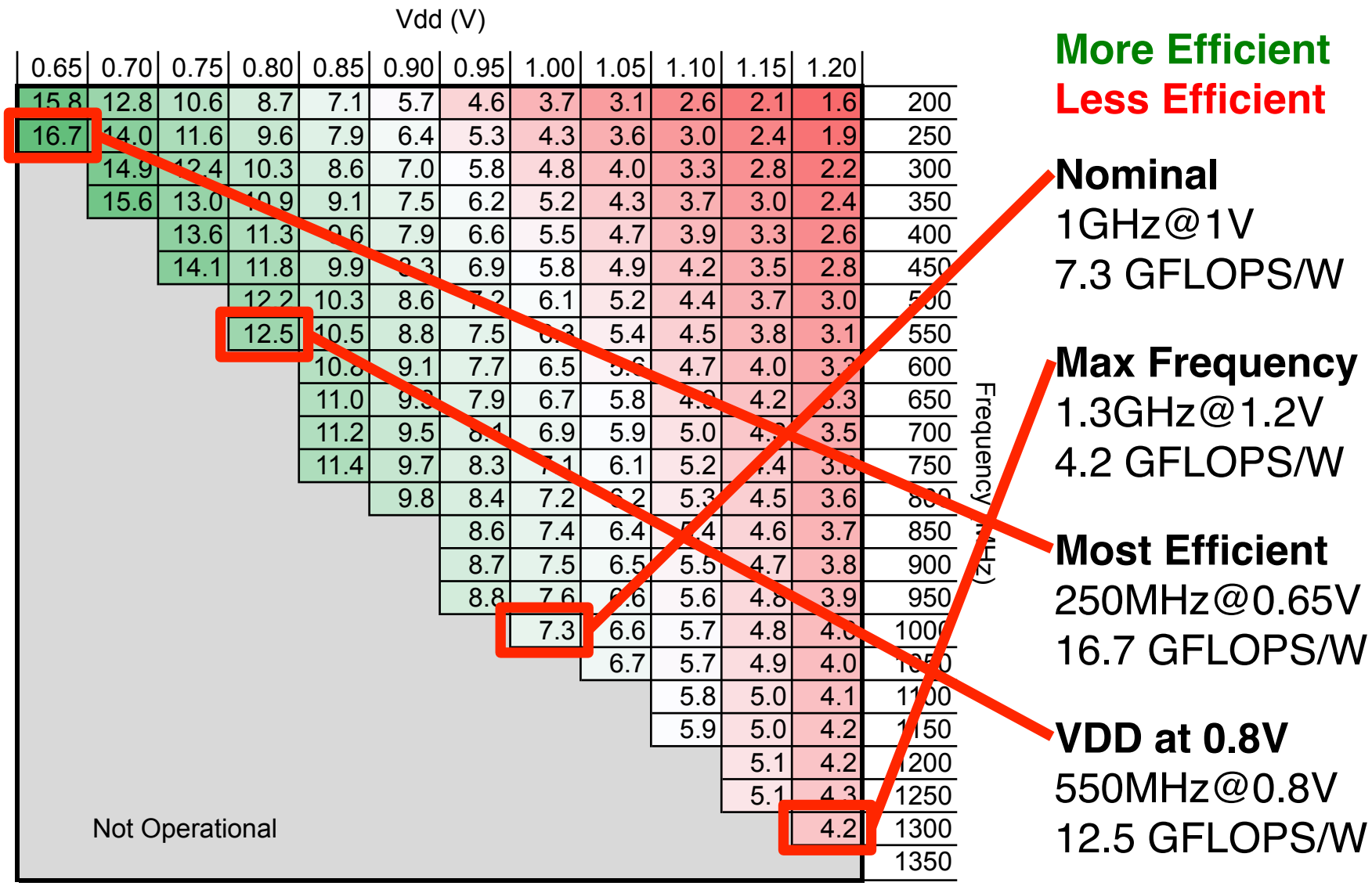
Chip Parameters		
Process	45nm SOI CMOS, 11 metal layers	
Package	C4 area I/O, flip-chip bonded to PCB	
Size	Processor	2.8mm X 1.1mm
	1 Core	1.37mm X 1.06mm
	SRAM Array	1.1mm X 4mm
Standard Cells	Processor	425K (85K flip-flops)
	1 Core	192K (36K flip-flops)
SRAM Bits	Processor	1246K
	1 Core	621K
Frequency	1GHz (Nominal), 250MHz-1.3GHz	
Voltage	1V (Nominal), 0.65V-1.2V	
Power	300mW-430mW (Nominal), 40mW-960mW	

Measurement Setup



Shmoo Plot of DP GFLOPS/W

Running Double-Precision Matrix Multiplication on Vector Accelerator



More Efficient

Less Efficient

Nominal
1GHz@1V
7.3 GFLOPS/W

Max Frequency
1.3GHz@1.2V
4.2 GFLOPS/W

Most Efficient
250MHz@0.65V
16.7 GFLOPS/W

VDD at 0.8V
550MHz@0.8V
12.5 GFLOPS/W

Energy Efficiency Comparison

@0.8V	Frequency (GHz)	64-bit GFLOPS	Power (W)	Efficiency (GFLOPS/W)
Blue Gene/Q	1.60	204.8	29.7	6.9
IBM Cell	3.20	108.8	22.5	4.8
This Work	**0.55**	**1.72**	**0.138**	**12.5**

- BG/Q and IBM Cell fabricated in same 45nm SOI
- Conservatively assume BG/Q and Cell achieves peak GFLOPS, we achieve 78% of peak GFLOPS
- Power numbers only for the core with private caches
  - Blue Gene/Q: Cores dissipate 54% of total power
  - IBM Cell: Assume that cores dissipate 50% of total power
- Why better energy efficiency than others?
  - Simpler, but yet more energy-efficient microarchitecture

More on Comparison

- But BG/Q is clocked 3X faster and Cell is 6X faster?
  - If the end goal is to provide better energy efficiency then use simpler microarchitectures and rely on parallelism for performance.
- But BG/Q and Cell have big on-chip caches? What about I/O power?
  - We only count the power dissipated in the core and the private L1 caches.
- But BG/Q and Cell have 100X more total GFLOPS!
  - Sorry, we only had budget for a small test chip.

Conclusions

- Processor generators written in high-level languages can produce energy-efficient, high-performance hardware
  - Our dual-core RISC-V vector processor achieves 16.7 DP GFLOPS/W at 0.65 V and a maximum frequency of 1.3 GHz at 1.2 V
- Open-source RISC-V ISA can serve as a competitive base ISA for integrating specialized heterogeneous accelerators
- Rocket chip generator and software tools open-sourced at http://riscv.org

Acknowledgment

- DARPA award HR0011-11-C-0100
- DARPA award HR0011-12-2-0016
- Center for Future Architecture Research, a member of STARnet, a Semiconductor Research Corporation program sponsored by MARCO
- NVIDIA graduate fellowship
- ASPIRE Lab industrial sponsors and affiliates Intel, Google, Nokia, NVIDIA, Oracle, and Samsung
- All POEM team members at MIT, UC Berkeley, CU Boulder