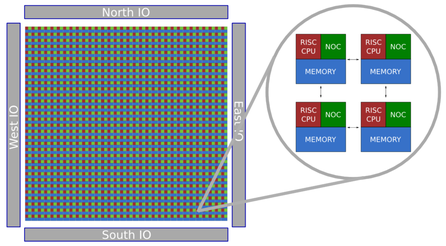
A TFLOPS scale 16nm 1024-core 64-bit RISC-Array Processor

Adapteva, Inc (Funded by DARPA under contract HR0011-15-9-0013 & CRAFT)

Abstract

This paper describes the design of the world's first 1024 core 64-bit System-On-Chip.



Features:

- 5th generation Epiphany processor
- 1024 independent 64-bit RISC processors
- System scaling to 1 billion processors
- 64MB distributed on-chip memory
- 3 x 136-bit wide 2D mesh NOCs
- 2052 Independent power domains
- 1024 programmable I/O pins
- Binary compatibility with Epiphany III & IV
- Compressed 16/32-bit ISA
- Custom ISA extensions for deep learning, communication, and cryptography

Performance: (simulated)

- 2 TFLOPS IEEE 32-bit
- 1 TFLOPS IEEE 64-bit
- 512 Billion Complex MACs/sec
- 128 Tbit/sec Memory Bandwidth
- 6 Tbit/sec NOCbisection bandwidth
- 10 Watt typical power consumption
- 10 mW minimum active power consumption

Supported FOSS Programming Models:

- C/C++ (GCC 7.2)
- OpenCL
- OpenMP
- OpenSHMEM
- MPI
- Erlang
- Python

Epiphany-V Architecture Description

Instruction Set Architecture (Base)

Type	Number	Opcodes
Floating Point	9	FADD, FSUB, FMUL, FMADD, FMSUB, FIX, FLOAT, FMAX, FABS
Integer	9	ADD, SUB, LSR, ASR, LSL, AND, XOR, ORR, BITR
Load/Store	6	LDR, STR, LDRI,STRI, LDRPM,STRPM
Branch	4	B,BL,JALR, JR
Move	5	MOV, MOVI, MOVTS, MOVFS, MOVCC
Control	8	NOP, WAND, GID, GIE, RTI, TRAP, IDLE, BKPT
Synchronization	7	TESTSET, FETCHADD, SWP, CMPSWP, FETCHAND, FETCHOR, FETCHXOR
Total	48	

Memory Architecture

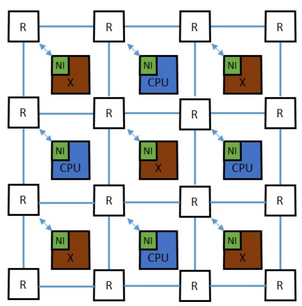
Global Memory Map							
63:62	61:52	51:48	47:44	43:32	31:26	25:20	19:0
RES	Z	ROWHI	COLHI	MEMHI	ROWLO	COLLO	MEMLO

- 32/64 bit memory architecture
- Distributed NUMA shared memory model
- All addresses accessible by all cores
- Flat address map (No hardware cache!)
- 4x16KB local memory banks per node
- Strong local access memory model
- Weak remote access memory model
- Latency = (6 + 1.5 * #NOC hops) * CC

Local Memory Map			
	19:16	Start	End
MEM0	0	0	3FFF
MEM1	0	4000	7FFF
MEM2	0	8000	BFFF
MEM3	0	C000	FFFF

Network-On-Chip

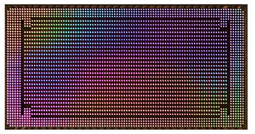
- Three 136-bit 2D mesh NoCs
- Everything is an address & packet
- Distributed x/y routing
- Round robin arbitration
- Single cycle pushback/stall



PACKET FORMAT							
	135:104	103:72	71:40	38:8	7:3	2:1	0
Write32	n/a	D1	D0	DSTLO	CMODE	DMODE	WR
Write64	DSTHI	D1	D0	DSTLO	CMODE	DMODE	WR
Read32	n/a	SRCLO	D0	DSTLO	CMODE	DMODE	WR
Read64	DSTHI	SRCLO	SAHI	DSTLO	CMODE	DMODE	WR

Implementation

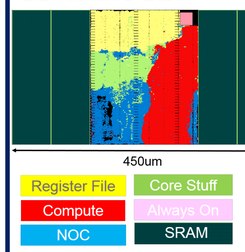
Die Photo



	Value
Process	16nm
Metal Layers	9
Transistors	4.5B
Die Area	117 mm²
FC Bumps	3,460
Bump Pitch	150um
I/O Signals	1,040
Clock Domains	1,152
Voltage Domains	2,052
Frequency	500Mhz

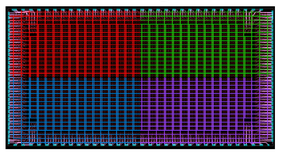
	Area (mm²)	Die Share
SRAM	62.4	53.3%
RegFile	15.1	12.9%
Compute	11.8	10.1%
NOC	12.1	10.3%
"Stuff"	6.5	5.6%
IO Pads	3.9	3.3%

Processor Tile



- NOC baesd platform
- Parametrized Verilog
- 8/16/32/64bit support
- Supports custom cores
- Up to 1 billion cores
- github.com/parallella/oh

Power Management



- 8 external supplies
- Per tile power down
- Split SRAM/logic rails
- 2,052 power domains
- Power = 10mW* cores

CAD Run Times	BlockA (x4)	BlockB (x4)	Tile	Top Level
Synthesis	0.05	0.13	0.4	0
PNR	0.28	1.66	3.66	1
Fill	0.03	0.03	0.066	5
DRC	0	0	0	11
Hours	1.46	7.3	4.1	17

"Server Farm":
- One Xeon server
- 32GB RAM
- One EDA license

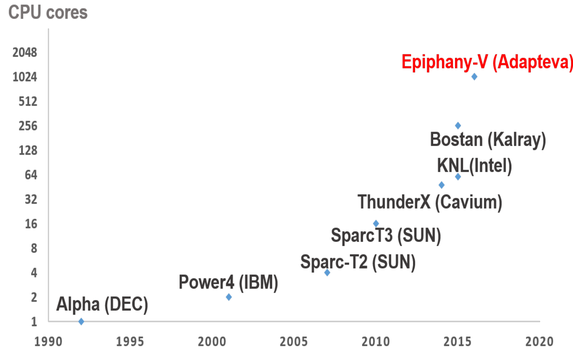
Design Costs

Designer	Task	Hours
Worker A	FPU	200
Worker B	DV Engine	200
Worker C	EDA Services	112
Ola Jeppsson	Simulator	500
Andreas Olofsson	Everything else	4,100

Before & After 16nm

Processor Comparison

	Epiphany-III	Epiphany-IV	Epiphany-V
Year	2010	2011	2016
Technology	65nm	28nm	16nm
Architecture	32 bit	32 bit	64 bit
Cores	16	64	1024
Memory	0.5 MB	2 MB	64 MB
Performance	16 GFLOPS	102 GFLOPS	2048 GFLOPS
Chip Power	~2 Watts	~2 Watts	~10 Watts
Tile Area	0.5 mm²	0.13 mm²	0.1 mm²
Customers	10,000	2 Tier1 Vendors	Evaluation



	P100	KNL	Broadwell	Epiphany-V
Process	16nm	14nm	14nm	16nm
Processors	56	72	24	1024
Area (mm²)	610	683	456	117
Transistors	15.3B	7.1B	7.2B	4.5B
Power (W)	250	245	145	~10
TFLOPS	4.7	3.6	1.3	1
GFLOPS/mm²	7.7	5.27	2.85	8.55
GFLOPS/W	18.8	14.69	9.08	100
SRAM/mm²	0.034	0.05	0.15	0.54
Nodes/mm²	0.09	0.11	0.05	8.75

- **5X improvement in GFLOPS/W**
- **79X improvement in CPU density**

Contact Information:

support@adapteva.com

http://adapteva.com

http://parallella.org