A PetaOp/s is Currently Feasible by Computing in RAM

Duncan G. Elliott

W. Martin Snelgrove†

Christian Cojocaru†

Michael Stumm

Department of Electrical & Computer Engineering, University of Toronto

† Department of Electronics, Carleton University

Building “Smart Memories” with SIMD processors at the sense-amps.

PetaFLOPS Frontier’95

dunc@eecg.toronto.edu (416)978-1659 fax (416)971-2326

A PetaOPS in Which Technology?

Optical
- Fast switching times
- Large scale integration will come

Advanced III-V Semiconductors (e.g. GaAs)
- Fastest semiconductors available
- Used in fast logic and radio circuits
- Manufacturing yield, reliability, and power consumption a problem for high levels of integration

Silicon
- Fast enough and getting faster
- Reliable, highest levels of integration
- Can design for low power (no static power dissipation in CMOS)

What is C•RAM?

- Computation moved from the CPU into the memory
- Computational RAM (C•RAM) = RAM + SIMD
  Single Instruction stream Multiple Data stream processor
- Processing Elements (PEs) at sense amps
  Small area (3-18% increase)
  PetaOPS performance
- Flexible
  Operation controlled by software
- "Shared memory" interface to host processors

DRAM or SRAM?

Same arguments apply to memory and C•RAM

DRAM

- densest
- less expensive per bit
- slow <100ns cycle time readily available

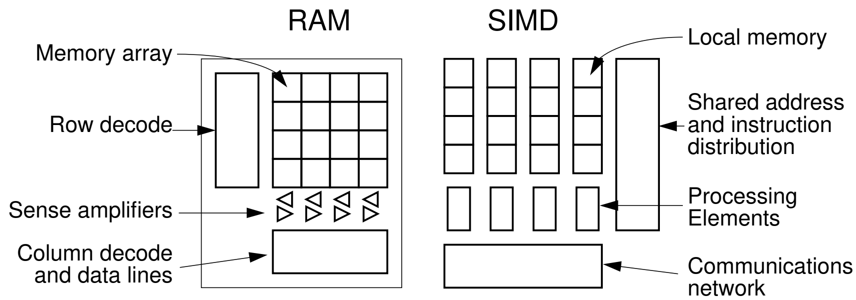
SRAM

- less dense, more chips
- expensive
- fast <10ns cycle time

No memory?

- If an application can use a data-path only systolic array then it doesn't need significant amounts of memory or, for that matter, C•RAM

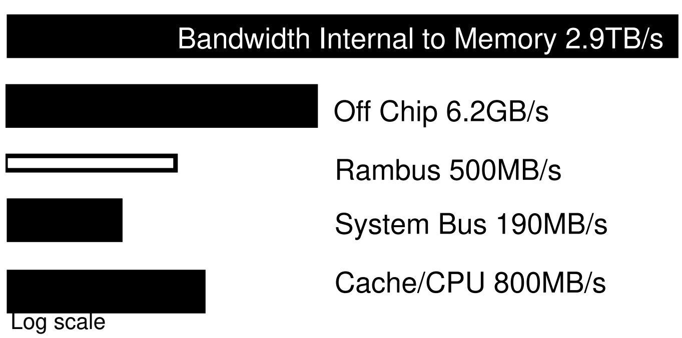
Compare RAM and SIMD



• Memory column	Local Memory
• Row address	Common local memory address
• Data lines	Communication
• Accesses every column	Accesses every local memory

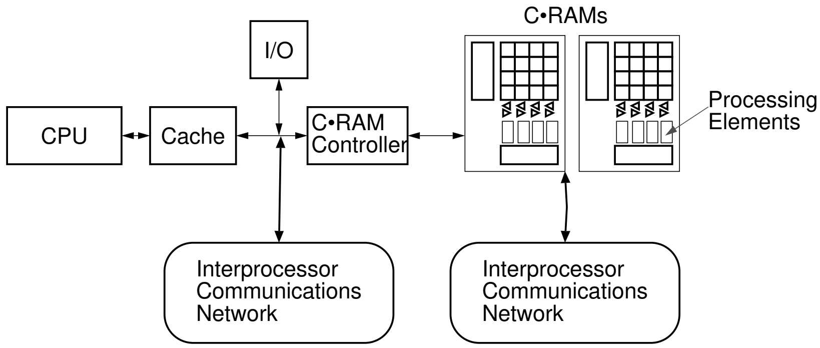
Motivation

- $3\frac{1}{2}$ orders of magnitude more memory bandwidth inside the chip (256MB of 16Mb 50ns DRAM, 64bit bus, 100MHz CPU)



- Video RAMs, Synchronous Mode DRAMs, RAMBUS clock the data at higher rates but are still pin limited
- C•RAM exploits a 2Mb datapath without going off-chip

A C•RAM computer



- Add PEs to memory at the sense amps
- Exploit highest memory bandwidth

Data Bandwidth Requirements for a PetaOPS

32 bit, 3 operand operations require 12PB/s data bandwidth

Off-chip computing

- Assume processor registers and on-chip cache reduce the data bandwidth requirement by a factor of 24
- Requires 0.5PB/s memory bandwidth
- 2.5 million SRAM chips (16Mb, 10ns, 16 data pins per chip)
- 40 million bus wires

Computing in the memory (C•RAM)

- Assume we are register poor, need 24PB/s memory bandwidth
- Need as few as 15 thousand SRAM chips
- More on account of geometries, number of active arrays, power

Power Consumption

Conventional memories waste memory bandwidth which wastes power

- Power is consumed by every column of memory when a row is accessed
- In standard DRAMs, 0.4% to 0.1% of retrieved bits are delivered off-chip
- C•RAM can utilize 25% to 100% of active columns

Off-chip signal power

- Consider power used in off-chip memory access for a petaOPS
  0.5PB/s, 3.3V swing, 50pF load ⇒1 megawatt just to drive a bus!
- Contrast collecting the bandwidth in the memory at the sense amps
  12PB/s, 3.3V with centre precharge, 400fF bitline pair ⇒ 24kW
  memory voltages are dropping: 3.3V commodity, 1.2V published, 1V soon
- With C•RAM, only interprocessor communications and shared instruction travel off-chip during computation

Why SIMD?

- Doesn't SIMD limit what kinds of applications can be run?
  Yes
- Don't some PEs sit idle in conditional code?
  Yes

- High performance per transistor (per silicon area, per volume, per $)
  Build in more peak performance to offset utilization.
- Memory structure provides a long word (e.g. 1Kb - 4Kb)
  Don't throw away bandwidth by using part of word
  Common row address, can't use MIMD
  Don't change memory architecture and add overhead
  ⇒ Use SIMD
- Simple synchronization and communication
- Shared instruction decode allows a 75 transistor PE

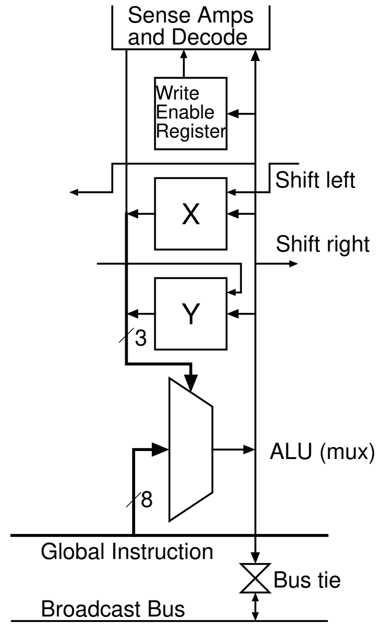
Architecture

Processing Element

- ALU is a multiplexer
  Memory and 2 registers as inputs
- Result to registers or memory
- Write enable permits conditional operations
- Bit-serial arithmetic
- Registers double as left-right communication
  1D on chip, 2D inter-chip
- Bi-directional driver to Broadcast Bus
- Instruction fed from off chip

Memory

- Conventional memory interface is maintained for host
- Row address is local memory address
- Each PE sees its column(s) as local memory



C•RAM Chips

Prototype

- 64PEs × 128bits in 1.2µm CMOS
- PEs are 9% of chip area (and cost)
- 75 transistor PE; width of a memory column

High speed [Christian Cojocaru]

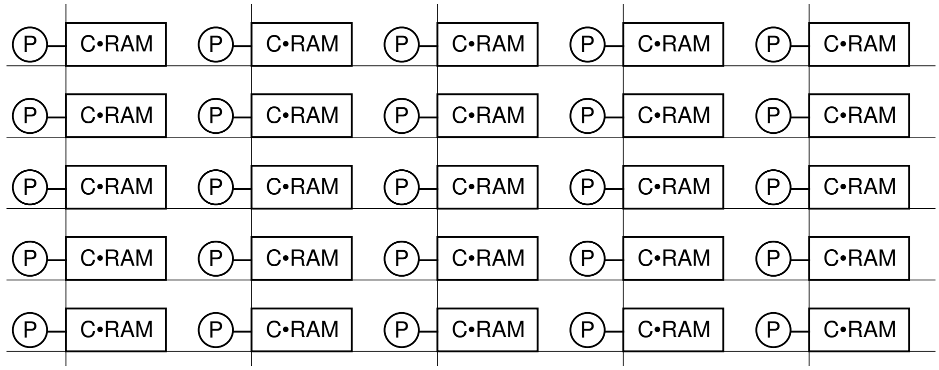
- 64PEs × 512bits in 0.8µm BiCMOS
- PEs are 3% of chip area

Proposed High Density

- 8192PEs × 2048bits in a 16Mbit DRAM process
- PE & logic add 18% to chip area
- “Price of commodity DRAM plus ε”

- 32K PEs × 512bits in a 16Mbit SRAM process

Proposed SIMD-MIMD Hybrid



- 4096 superscaler microprocessors with C•RAM for memory
- MIMD performs portions of application not suited to SIMD
- Single or multiple “SIMD” instruction streams so application can best utilize PEs

Peak SIMD Performance

4096 microprocessors each with 256MB $\Rightarrow$ 1TB SRAM
20ns read/ALU/write cycle

- 860 peta-operations/second boolean (but who cares)
- 46 petaOPS string compare, 8 bit character
- 6.7 petaOPS 32 bit addition
- 2.6 petaOPS multiply-accumulate 8/16 bit
- 0.4 petaFLOPS multiply, single precision

Performance

Application	Performance (petaOPS)
16×16 Convolution	3.6
Vector Quantization	3.4
LMS Best-Fit Associative Search and Update	3.2

Simulated for 1TByte of 20ns cycle C•RAM

Faster Multiplies & FP

- Problem with our bit-serial architecture: multiplicand and partial sums held in memory

- Gang together PEs in parallel under software control
- Arbitrary precision
- Added registers
- Multiplies no longer need intermediate memory accesses
- Christian Cojocaru is working on such a chip

- More architectural space to be explored

Conclusions

- Focus on memory bandwidth and power consumption leads us to bring the consumers of the memory bandwidth — the PEs — into the memory
- A petaOPS can be achieved with the C•RAM architecture and current silicon IC technology