An Ultra-High Performance Architecture for Embedded Defense Signal and Image Processing Applications

Stewart Reddaway and **Pete Rogina**
World*Scape* Defense Company
Email: {sfr, pete} @wscapeinc.com

Ken Cameron, Simon McIntosh-Smith, and David Stuttard
ClearSpeed Technology, Limited
Email: {ken, simon, daves} @clearspeed.com

Michael Koch, Rick Pancoast, and **Joe Racosky**
Phone: (856) 797-8700
Fax: (856) 797-0026
Email: {Michael.j.koch, rick.pancoast, joseph.r.racosky} @lmco.com

Abstract:

This briefing describes the development of a novel, ultra-high performance next-generation Single-Instruction/Multiple-Data (SIMD) processing architecture originally designed to realize immersive, photo-realistic 3-D simulations. This low-power, Multi-Threaded Array Processor (MTAP) architecture provides for hundreds and ultimately thousands of processing elements, each with optional floating point hardware, to perform data parallel processing on image and signal processing applications as well as for compression, encryption, search, and general sensor processing applications. The technology is supported by a flexible development environment, including assembly language and C-based language support, as well as a cycle accurate simulator, with plans to develop industry standard API Libraries based upon VSIPL and, ultimately, HPEC-SI. This new technology, being developed by World*Scape* and ClearSpeed, has been shown to provide ten to one hundred times the overall performance of PowerPC or Pentium-based architectures, especially when performing image and signal processing functions, such as FFTs or filters. In general, the architecture has been shown to provide significant throughput, size, and power advantages for embedded processing applications.

ClearSpeed Technology Limited is developing the MTAP architecture that provides a scalable array of Processing Elements (PEs) on a single die. Currently 64 and 256 PE devices are planned, although the array can scale to 1,000s of PEs. The technology is complemented by a scalable packet switched bus architecture called ClearConnect that has been designed to support the high bandwidths required for many applications. The technology is proven in silicon and is capable of delivering per device peak performance of over 100 GFLOPS while dissipating less than 5 Watts. The processor is supported by a professional Software Development Kit (SDK) and includes an optimizing C compiler, graphical debugger and a full suite of supporting tools and libraries.

Report Documentation Page		*Form Approved* *OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **20 AUG 2004**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**

4. TITLE AND SUBTITLE **An Ultra-High Performance Architecture for Embedded Defense Signal and Image Processing Applications**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **WorldScape Defense Company; ClearSpeed Technology, Limited**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001694, HPEC-6-Vol 1 ESC-TR-2003-081; High Performance Embedded Computing (HPEC) Workshop (7th)., The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**33**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

World*Scape* Defense Company has been developing key algorithms and library functions such as FFTs and FIR filters which efficiently utilize the architecture and floating point per PE hardware to gain exceptional performance at very low power dissipation levels. Specific application work, supported by the Office of Naval Research, has been undertaken for radar processing with raw throughput numbers for functions such as FFTs, complex multiplies, filters, etc. significantly higher than other industry standard processing and DSP platforms. This briefing also describes new levels of benchmark performance for FFT per second per watt that provide the basis of plans for embedded SAR processing systems on small UAV's. High level C and VSIPL library support are planned and currently under development.

Lockheed Martin Naval and Electronic Surveillance Systems has been trained in the use of the SDK, and has ported some key, high-performance application benchmarks, such as radar pulse compression, for performance comparison with general purpose processing architectures. Results have shown the potential for considerable performance enhancement for airborne, shipboard, ground-based and undersea tactical signal and image processing systems.

In this briefing, we describe an embedded processing architecture that promises a performance advantage over conventional general-purpose processors of one or two orders of magnitude. Finally, the results of a DoD benchmark algorithm run on the cycle-accurate simulator in summer, 2003, will be presented and compared with general- purpose processor performance.

An Ultra-High Performance Architecture for Embedded Defense Signal and Image Processing Applications

Authors

Ken Cameron (ken@clearspeed.com)

Mike Koch (michael.j.koch@lmco.com)

Simon McIntosh-Smith(simon@clearspeed.com)

Rick Pancoast (rick.pancoast@lmco.com)

Joe Racosky (joseph.r.racosky@lmco.com)

Stewart Reddaway (sfr@wscapeinc.com)

Pete Rogina (pete@wscapeinc.com)

David Stuttard (daves@clearspeed.com)

September 24, 2003

Agenda

Overview

Architecture

Benchmarking

Applications

Summary

Q & A

Overview

✦ Work Objective

- **Provide benchmark for Array Processing Technology**

  – **Enable embedded processing decisions to be accelerated for upcoming platforms (radar and others)**

  – **Provide cycle accurate simulation and hardware validation for key algorithms**

  – **Extrapolate 64 PE pilot chip performance to WorldScape 256 PE product under development**

  – **Support customers strategic technology investment decisions**

- **Share results with industry**
  – **New standard for performance AND performance per watt**

Architecture

Overview

Architecture

Benchmarking

Applications

Summary

Q&A

✦ *ClearSpeed's* Multi Threaded Array Processor Architecture – MTAP
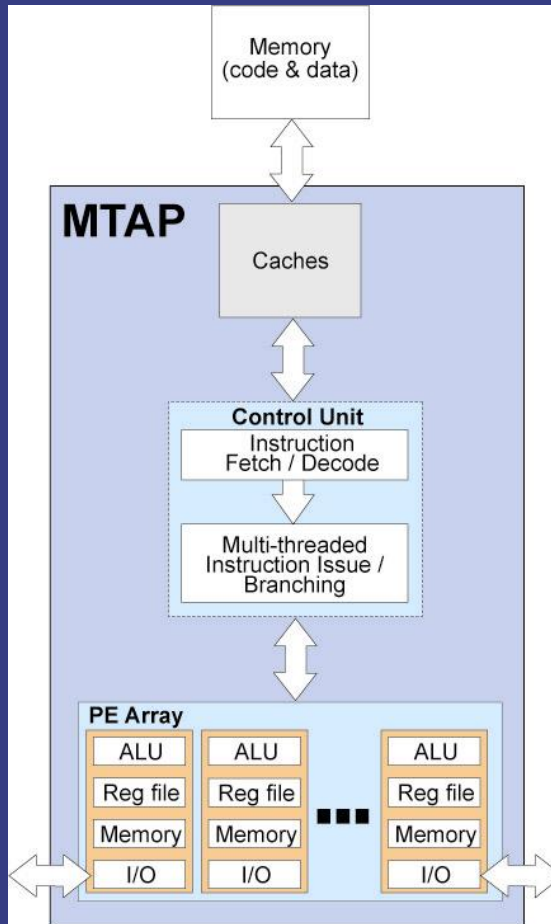
- **Fully programmable in C**

- **Hardware multi-threading**

- **Extensible instruction set**

- **Fast processor initialization/restart**

Architecture

Overview

Architecture

Benchmarking

Applications

Summary

Q&A

✦ *ClearSpeed's* Multi Threaded Array Processor Architecture – MTAP (cont.)
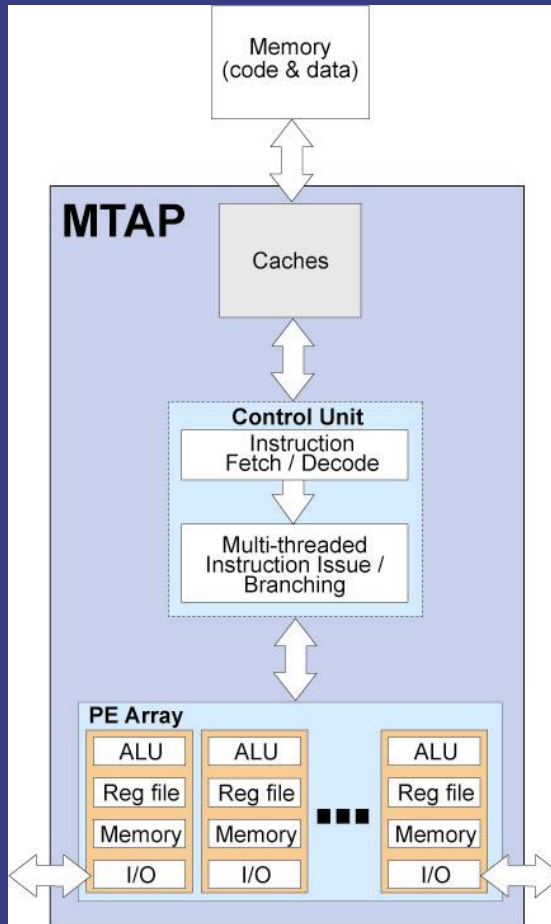
- **Scalable internal parallelism**
  - **Array of Processor Elements (PEs)**
  - **Compute and bandwidth scale together**
  - **From 10s to 1,000s of PEs**
  - **Built-in PE fault tolerance, resiliency**
- **High performance, low power**
  - **~10 GFLOPS/Watt**
- **Multiple high speed I/O channels**

✦ Processor Element Structure
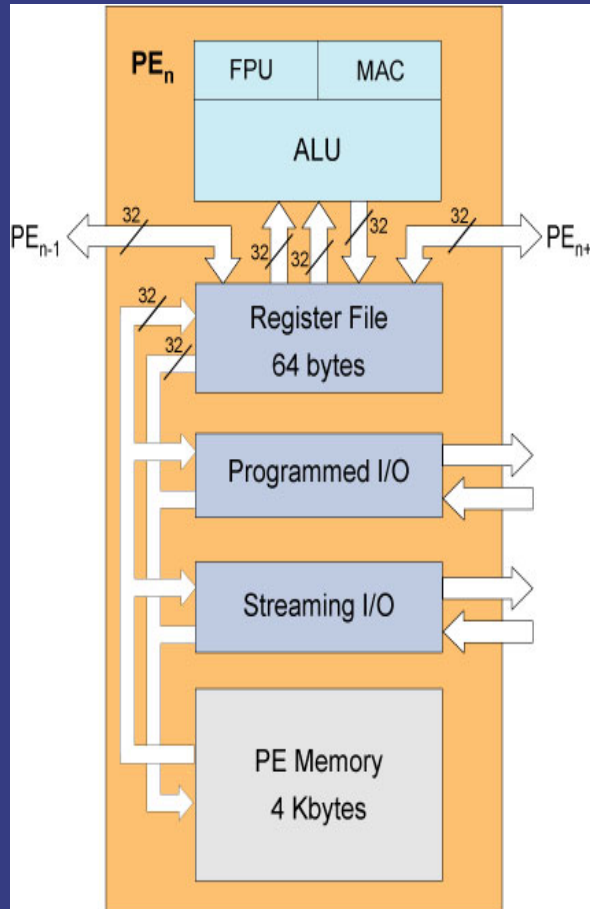
- **ALU + accelerators: integer MAC, FPU**
- **High-bandwidth, multi-port register file**
- **Closely-coupled SRAM for data**
- **High-bandwidth per PE DMA: PIO, SIO**
  - **64 to 256-bit wide paths**
- **High-bandwidth inter-PE communication**

Architecture

✦ **Processor Element Structure (cont.)**
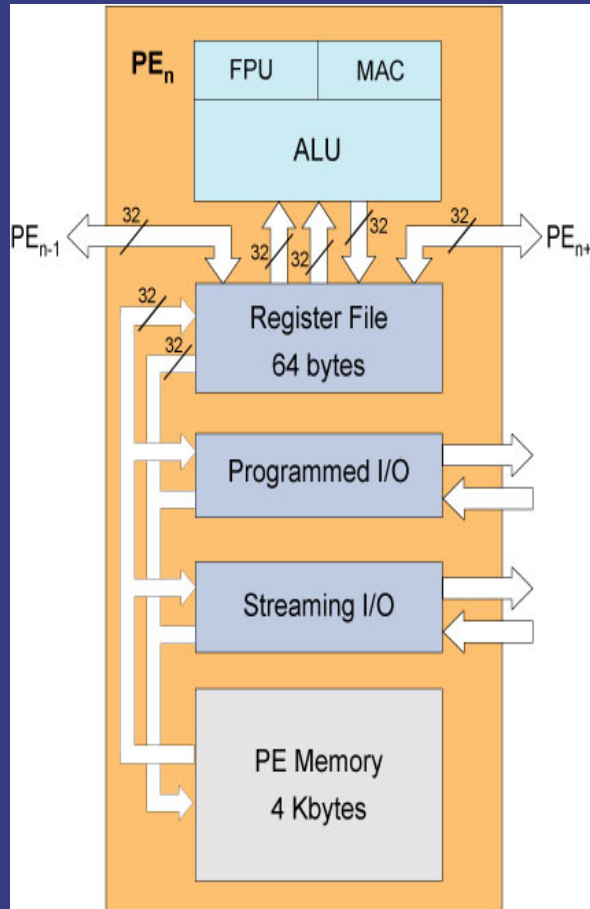
- **256 PEs at 200MHz:**
  - **200 GBytes/s bandwidth to PE memory**
  - **100 GFLOPS**
  - **Many GBytes/s DMA to external data**

- **Supports multiple data types:**
  - **8, 16, 24, 32-bit, ... fixed point**
  - **32-bit IEEE floating point**

Architecture

✦ **Architecture DSP Features:**

- **Multiple operations per cycle**
  - **Data-parallel array processing**
  - **Internal PE parallelism**
  - **Concurrent I/O and compute**
  - **Simultaneous mono and poly operations**
- **Specialized per PE execution units**
  - **Integer MAC, Floating-Point Unit**
- **On-chip memories**
  - **Instruction and data caches**
  - **High bandwidth PE "poly" memories**
  - **Large scratchpad "mono" memory**
- **Zero overhead looping**
  - **Concurrent mono and poly operations**

Architecture

✦ **Software Development:**

- **The MTAP architecture is simple to program:**
  - **Architecture and compiler co-designed**
  - **Uniprocessor, single compute thread programming model**
    - **Array Processing: one processor, many execution units**
- **RISC-style pipeline**
  - **Simple, regular, 32-bit, 3 operand instruction set**
- **Extensible instruction set**
- **Large instruction and data caches**
- **Built-in debug: single-step, breakpoint, watch**

Architecture

Overview

Architecture

Benchmarking

Applications

Summary

Q&A

- ✦ **Software Development Kit (SDK):**
  - **C compiler, assembler, libraries, visual debugger etc.**
  - **Instruction set and cycle accurate simulators**
  - **Available for Windows, Linux and Solaris**
  - **Development boards & early silicon available from Q4 03**

- ✦ **Application development support:**
  - **Reference source code available for various applications**
  - **Consultancy direct with ClearSpeed's experts**
  - **Consultancy and optimized code from ClearSpeed's partners**

Architecture

✦ *ClearSpeed* **debugger:**

Benchmarking

✦ **FFTs on 64 PE pilot chip and 256 PE processor**

- **Benchmark for FFT→Complex Multiply→IFFT**
  - **FFT benchmark example is 1024-point, complex**
  - **IEEE 32-bit floating point**

- **Each FFT mapped onto a few PEs**
  - **Higher throughput than one FFT at a time**

- **In-situ assembler codes**
  - **Single 1024-point FFT or IFFT spread out over 8 PEs**
  - **128 complex points per PE**
  - **Output bit-reversed mapping in poly RAM**
  - **IFFT input mapping matches bit-reversed FFT output**

Benchmarking

- ✦ **FFTs on 64 PE pilot chip and 256 PE processor**
  - **128 complex points per PE enables**
    - **High throughput**
    - **Enough poly RAM buffer space for multi-threading to overlap I/O and processing**

  - **Complex multiply code**
    - **Multiplies corresponding points in two arrays**
    - **Enables convolutions and correlations to be done via FFTs**

  - **Performance measured on Cycle Accurate Simulator (CASIM)**
    - **Individual FFT**
    - **IFFT**
    - **Complex Multiply**
    - **Iterative loops including I/O**
    - **FFT – CM – IFFT (using a fixed reference function)**

Benchmarking

Overview

Architecture

Benchmarking

Applications

Summary

Q&A

✦ **Detail on the current FFT codes**

- **4 complete radix-2 butterfly instructions in microcode**

    - **Decimation-in-time (DIT) - twiddle multiply on one input**

    - **DIT variant - multiply inserts an extra 90 degree rotation**

    - **Decimation-in-frequency (DIF) - multiply on one output**

    - **DIF variant - extra 90 degree rotation inserted**

    - **These instructions take 16 cycles**

Benchmarking

✦ **Detail on the current FFT codes (cont.)**

- **Twiddle factors pre-stored in poly RAM**
  - **Use of 90 degree variants halves twiddle factors storage**
  - **IFFT uses same twiddle factors**
  - **Fastest codes require 68 twiddle factors per PE**
- **I/O routines transfer data between poly and mono RAM**
- **Mono data is in entirely natural order**
- **Complete set of 1024 complex points per FFT per transfer**

✦ **Most Load/Store cycles are done in parallel with the butterfly arithmetic**

✦ **Moving data between PEs costs cycles**

Benchmarking

✦ Measured Performance

- **Performance is measured in GFLOPS @200 MHz**
  - Single FFT or IFFT counted as 5n*(log(n)) FLOPs
  - Complex Multiply is 6 FLOPs per multiply

Code	Cycles	GFLOPS 256 PE	256 PE batch size	GFLOPS 64 PE pilot chip	Pilot chip batch size
FFT	13.9k	23.5	32	5.9	8
IFFT	14.5k	22.6	32	5.6	8
CM	1.7k	23.1		4.5	
FFT-CM-IFFT	30.1k	23.1	32	5.8	8
FFT-CM-IFFT w I/O	32k	21.7	32	5.4	8

Benchmarking

✦ **Measured Performance** (cont.)

- **I/O transfers are to on-chip SRAM (mono memory)**
  - For the pilot chip, this I/O is ~75% of processing time
  - Pilot chip bandwidth to off-chip memory is 50% lower

- **256 PE product will have ~3 GB/s off-chip mono transfers**
  - Data transfer will be 100% overlapped with processing

Benchmarking

✦ Performance analysis (raw and per watt)

- **256 PEs @200MHz: peak of 102.4 GFLOPS**
  - **64 PE pilot chip @200MHz: peak of 25.6 GFLOPS**

- **Current code achieving 23.5 GFLOPS (23% of peak)**

- **In many applications, Performance/Watt counts most**
  - **256 PE product will be ~5 Watts based on GDSII measurements and other supporting data**

- **256 PE extrapolated FFT performance (200 MHz)**
  - **~23.5 GFLOPS becomes ~4.7 GFLOPS/Watt**
  - **~0.9 M FFTs/sec/Watt (1024-point complex)**

Overview
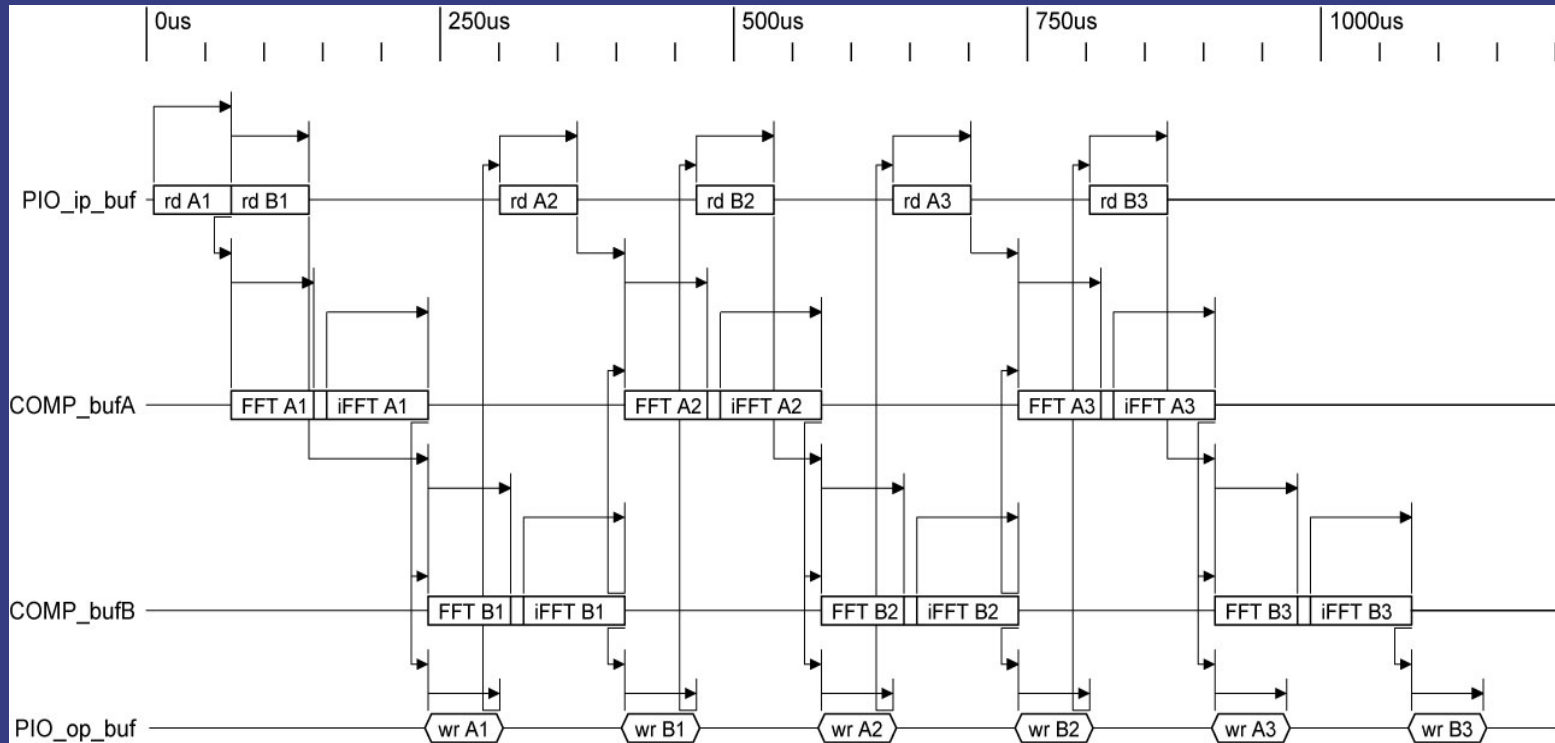
Architecture

Benchmarking

Applications

Summary

Q&A

Benchmarking

✦ Application Timing Diagram

Cycle accurate simulations show that the FFT-CM-iFFT Compute and I/O phases completely overlap when double buffering is employed

Benchmarking

✦ Other size FFTs

- **Spreading one FFT over more PEs increases data movement**
  - **Cost: lower peak GFLOPS**
  - **Benefit: Enables larger FFTs or lower latency per FFT**

Example (1024-point complex FFT on 64-PE pilot chip):

Reducing batch size from 8 (8 PEs/FFT) to 1 (64 PEs/FFT), changes performance from ~5.7 to ~2.5 GFLOPS, but compute time reduces by ~2.6x

Benchmarking

✦ **Other size FFTs (cont.)**

- **8k-point FFT performance extrapolation:**
  - ~2.9 GFLOPS on pilot chip (batch size = 1)
  - ~11.6 GFLOPS on 256 PE (batch size = 4)

- **128-point FFT performance extrapolation:**
  - ~31 GFLOPS on 256 PE product (batch size = #PEs)

256 PE product under development can deal with FFTs up to 32k points without intermediate I/O

However, >8k points may use intermediate I/O for speed

This enables unlimited FFT sizes

Benchmarking

- ✦ **Library FFT development**
  - **World*Scape* developing optimized FFT (and other) libraries**
    - **Start with components callable from Cn**
      - **Within-PE multiple short FFTs**
      - **Inter-stage twiddle multiplies**
      - **Across-PE data reorganization**
  - **VSIPL interface planned to maximize portability**
  - **Seeking industry and University partners**

Benchmarking

✦ **Faster in the Future**

- **New microcoded instructions**
  - **Better handling of FP units' pipeline for multiple butterflies**
    - **Split into Startup, multiple Middle, and End instructions**
    - **Each Middle does an extra butterfly faster than now**
  - **Separate multiplies from add/subtracts**
    - **Enables higher FFT radices**
    - **Saves first stage multiplies**

- **Speedup**
  - **Within-PE: ~100%, to ~62 GFLOPS for 256 PEs (~60% of peak)**
  - **1024-point FFT in 8 PEs: ~60%, to ~36 GFLOPS**

Benchmarking

Overview

Architecture

Benchmarking

Applications

Summary

Q&A

✦ **Technology Roadmap**

- **Faster clock speeds**

- **Faster/smaller memory libraries**

- **More PE's/chip**

- **Scalable I/O and chip-chip connections**

- **More memory/PE & mono memory**

- **Custom cell implementation**
  - **<50% size & power (leverage from parallelism)**

- **Future fabrication processes**
  - **Embedded DRAM, smaller line widths….**

Applications

Overview

Architecture

Benchmarking

Applications

Summary

Q&A

✦ **FFT/Pulse Compression Application**

- **1K complex 32-bit IEEE floating point FFT/IFFT**

- **Lockheed Martin "Baseline"**
  – **Mercury MCJ6 with 400-MHz PowerPC 7410 Daughtercards with AltiVec enabled, using one compute node**

  – **Pulse Compression implementation using VSIPL: FFT, complex multiply by a stored reference FFT, IFFT**

✦ FFT/Pulse Compression Application (cont.)

- **ClearSpeed CASIM (Cycle Accurate SIMulator)**
  - 64 PEs simulated at 200 MHz

  - 1K FFT or IFFT on 8 PEs with 128 complex points per PE

  - Pulse Compression using custom instructions: FFT, complex multiply by a stored reference FFT, IFFT

  - 32-bit IEEE standard floating point computations

Applications

✦ Performance Comparison

Function	Mercury G4 (LM Measured)	Mercury G4 (Published)	World*Scape* / ClearSpeed 64 PE Chip	Speedup vs. Mercury G4 (LM Measured)	Speedup vs. Mercury G4 (Published)
FFT	39.0 µs	25.0 µs*	8.7 µs	4.5 X	2.9 X
Complex Multiply	10.8 µs	N/A	1.1 µs	7.7 X	------
IFFT	34.7 µs	N/A	9.2 µs	3.8 X	------
Pulse Compression	152.2 µs	N/A	20.0 µs	7.6 X	------

* Mercury published figure for time to complete a 1K FFT on 400-MHz PowerPC 7410

Applications

Overview

Architecture

Benchmarking

Applications

Summary

Q&A

✦ Power Comparison

Processor	Clock	Power	FFT/sec /Watt	PC/sec/ Watt
Mercury PowerPC 7410	400 MHz	8.3 Watts	3052	782.2
World*Scape*/ ClearSpeed 64 PE Chip	200 MHz	2.0 Watts**	56870	24980
Speedup	----	----	18.6 X	31.9 X

**Measured Using Mentor Mach PA Tool

Applications

✦ **Power Comparison (cont.)**

World*Scape*/ClearSpeed 64 PE Chip	Total Cycles**
FFT	14067
Complex Multiply	1723
IFFT	14654
Pulse Compression	32026

** As simulated using CASIM at 200 MHz

Summary

- ✦ **Massively Parallel Array Processing Architecture (MTAP) presented**

  - **Programmable in C**
  - **Robust IP and tools**
  - **Applicability to wide range of applications**

- ✦ **Inherently scalable**
  - **On chip, chip-to-chip**

- ✦ **Low Power**

- ✦ **New standard for performance AND performance per watt**

Overview

Architecture

Benchmarking

Applications

Summary

Q&A