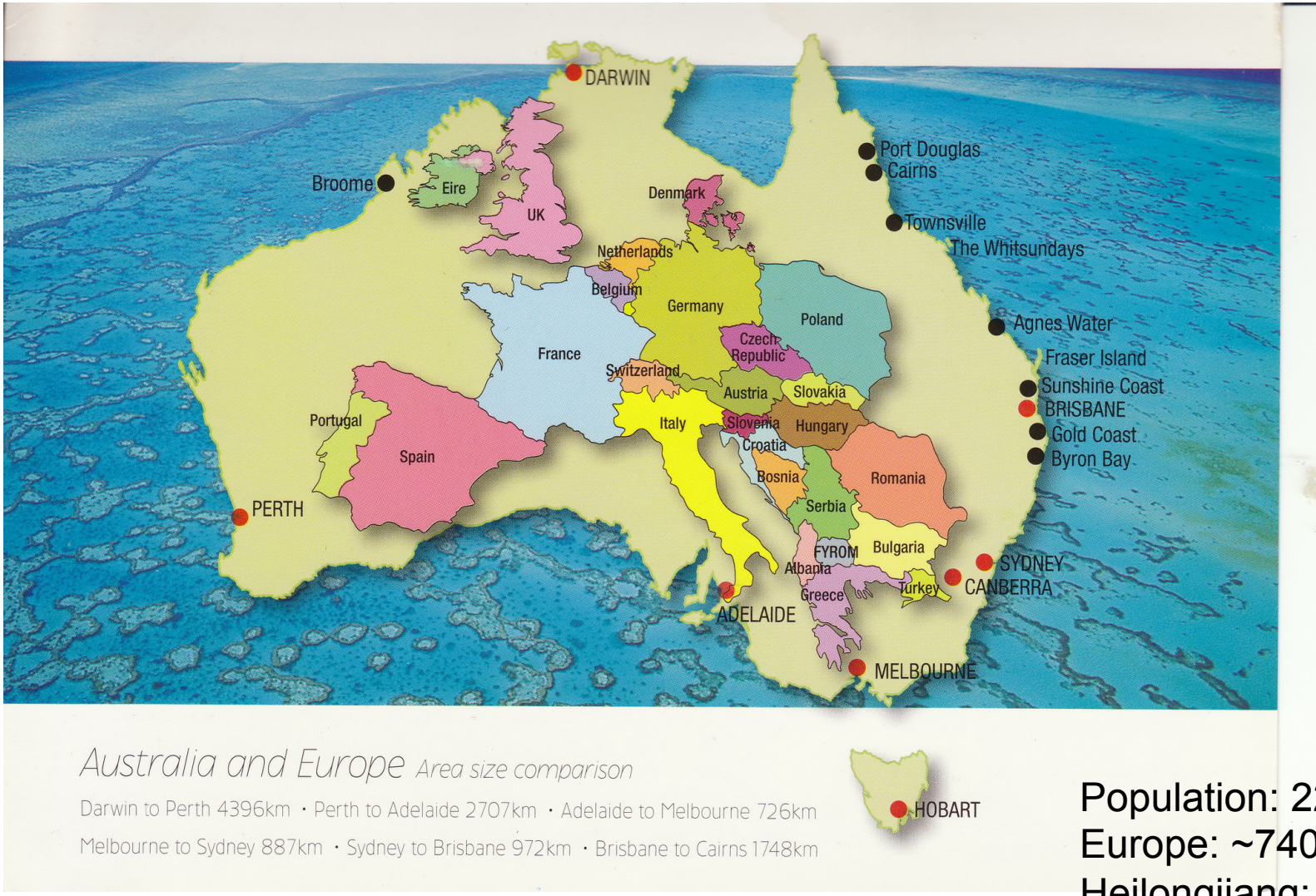
Reconfigurable Computing

Philip Leong (philip.leong@sydney.edu.au)
School of Electrical and Information Engineering

http://www.ee.usyd.edu.au/~phwl

THE UNIVERSITY OF SYDNEY

Australia



Population: 22M
Europe: ~740M (2011)
Heilongjiang: 38M

THE UNIVERSITY OF SYDNEY

Sydney

Population: 4.6M

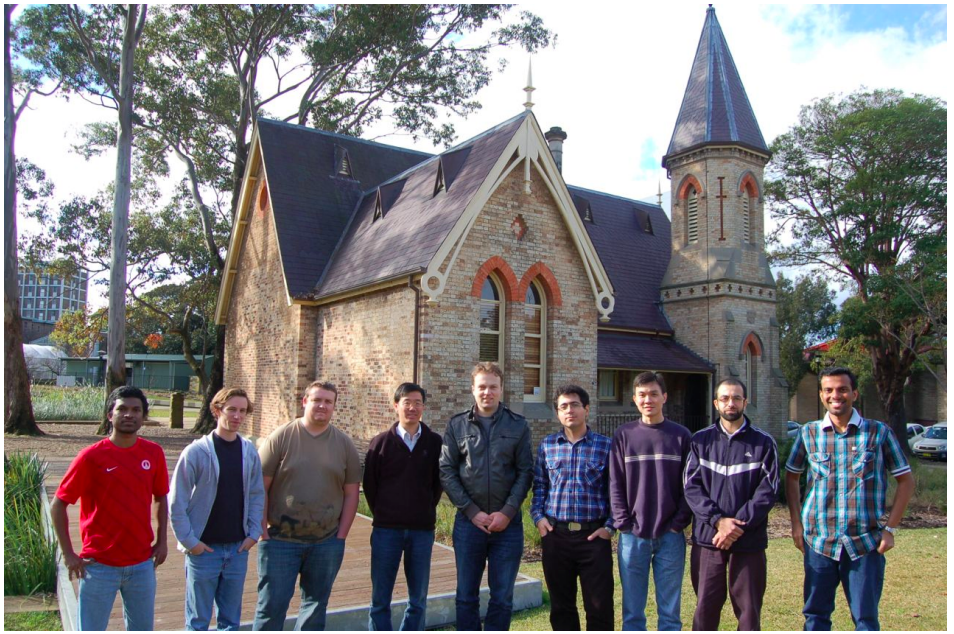
The University of Sydney

(Old part of campus – not our building)

Population: 49,000
from 130 countries

- Work focuses on using parallelism to solve computationally demanding problems
  - Develop novel computer architectures and computing techniques.
  - Understand tradeoffs between ASIC, FPGA, GPU and microprocessor technologies
  - Improve designer productivity

- Applications
  - Computational Finance
  - Signal Processing
  - Biomedical Engineering



This course

1. Introduction to Reconfigurable Computing
   - what is reconfigurable computing, applications, areas for research
2. Abstractions for Implementation
   - microcode, map-reduce
3. Monte Carlo Simulation
   - uniform and Gaussian random number generation, Monte Carlo arithmetic
4. FPGA Architecture
   - floating-point FPGAs, process variation

Introduction to Reconfigurable Computing

- **FPGAs**
- Reconfigurable computing
- Applications

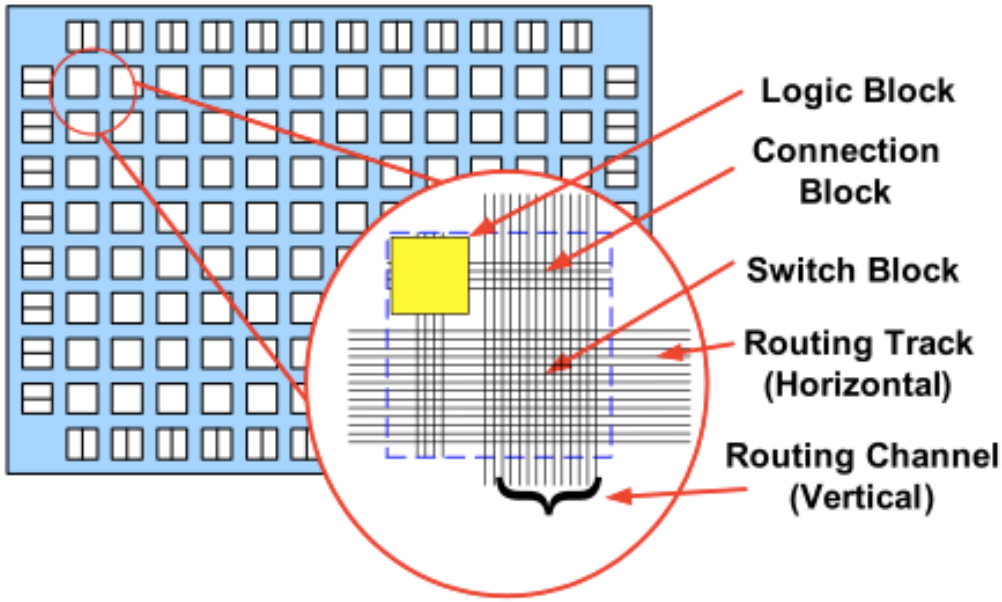
Integrated Circuits

Most electronics rely on application-specific ICs (ASICs) for perf, cost and P

Source: Arvind MIT

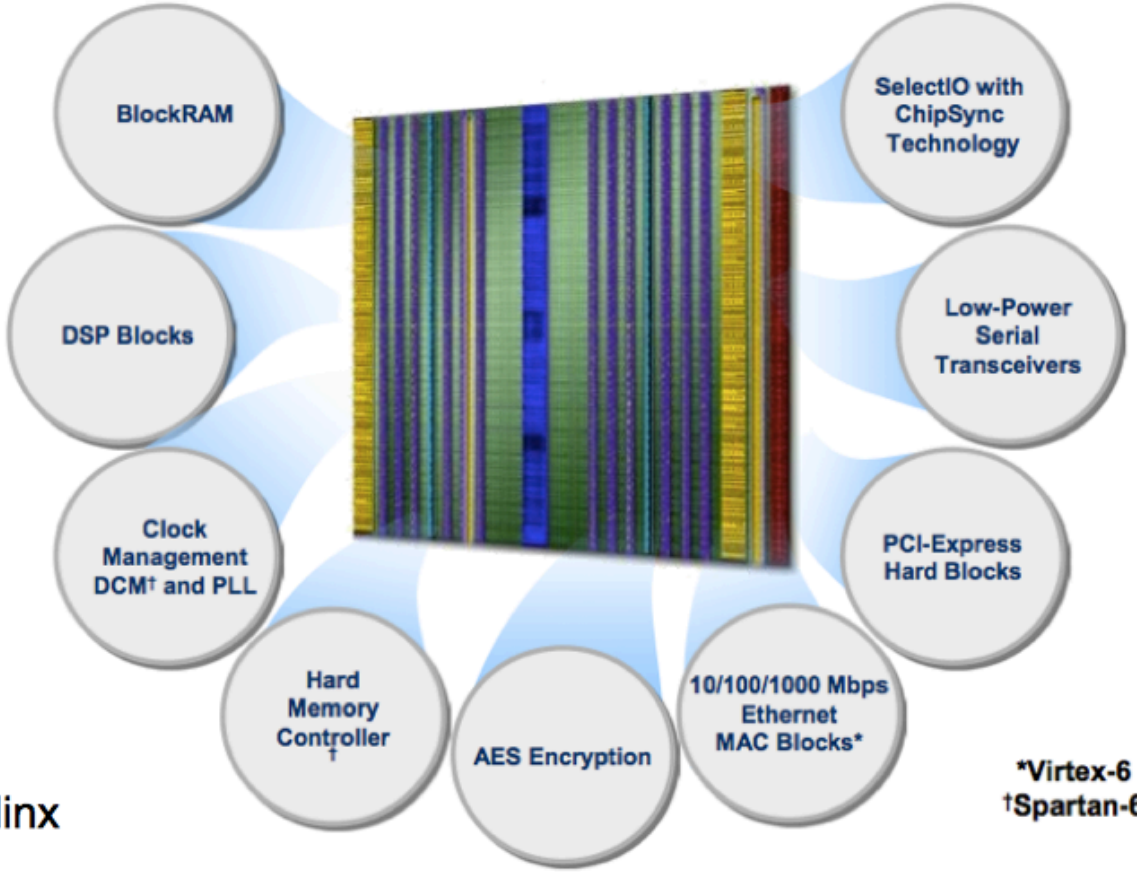
FPGA

- A generalised ASIC
  - Logic blocks for digital operations
  - Programmable interconnect for routing
- Arbitrary digital circuits can be implemented
- Functionality downloaded to FPGA memory (in seconds)



Source: Steve Wilton (UBC)

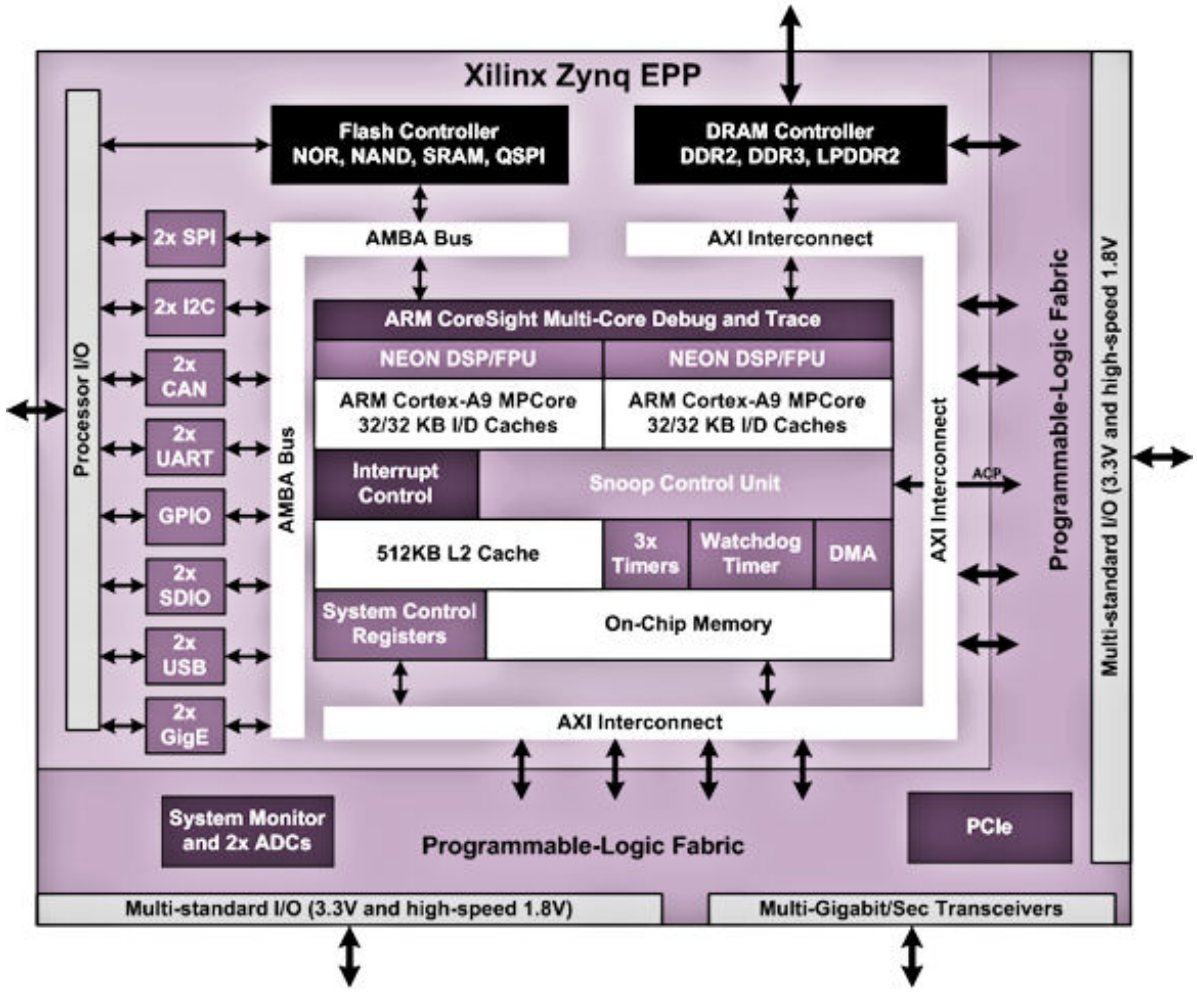
FPGA Embedded Blocks



Source: Xilinx

Hard IP blocks for widely-used functions: faster, more efficient, lower power
Careful choice: every user must pay for these functions, whether used or not

Zynq (ARM+ Reconfigurable Fabric)

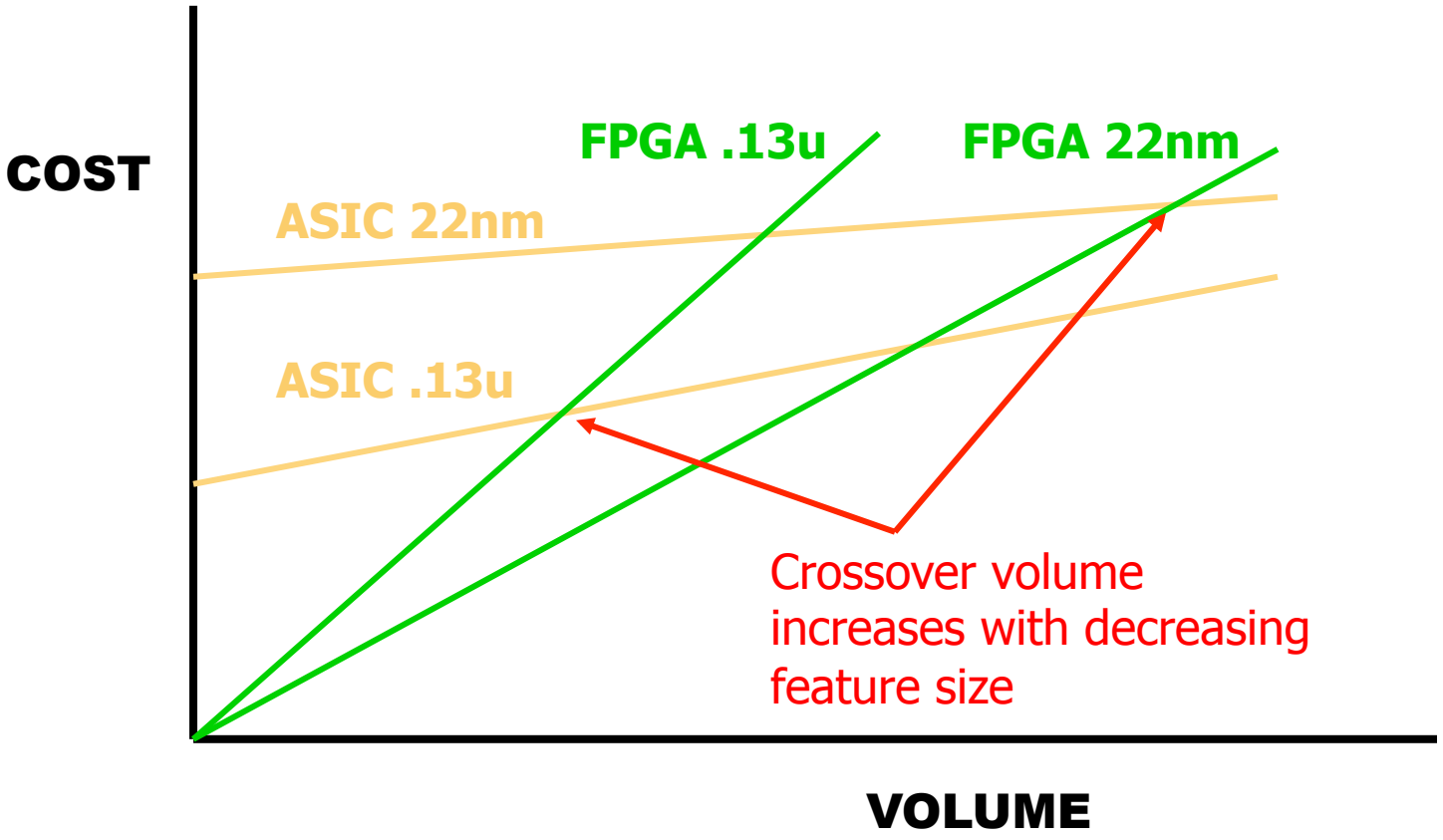


FPGA Families

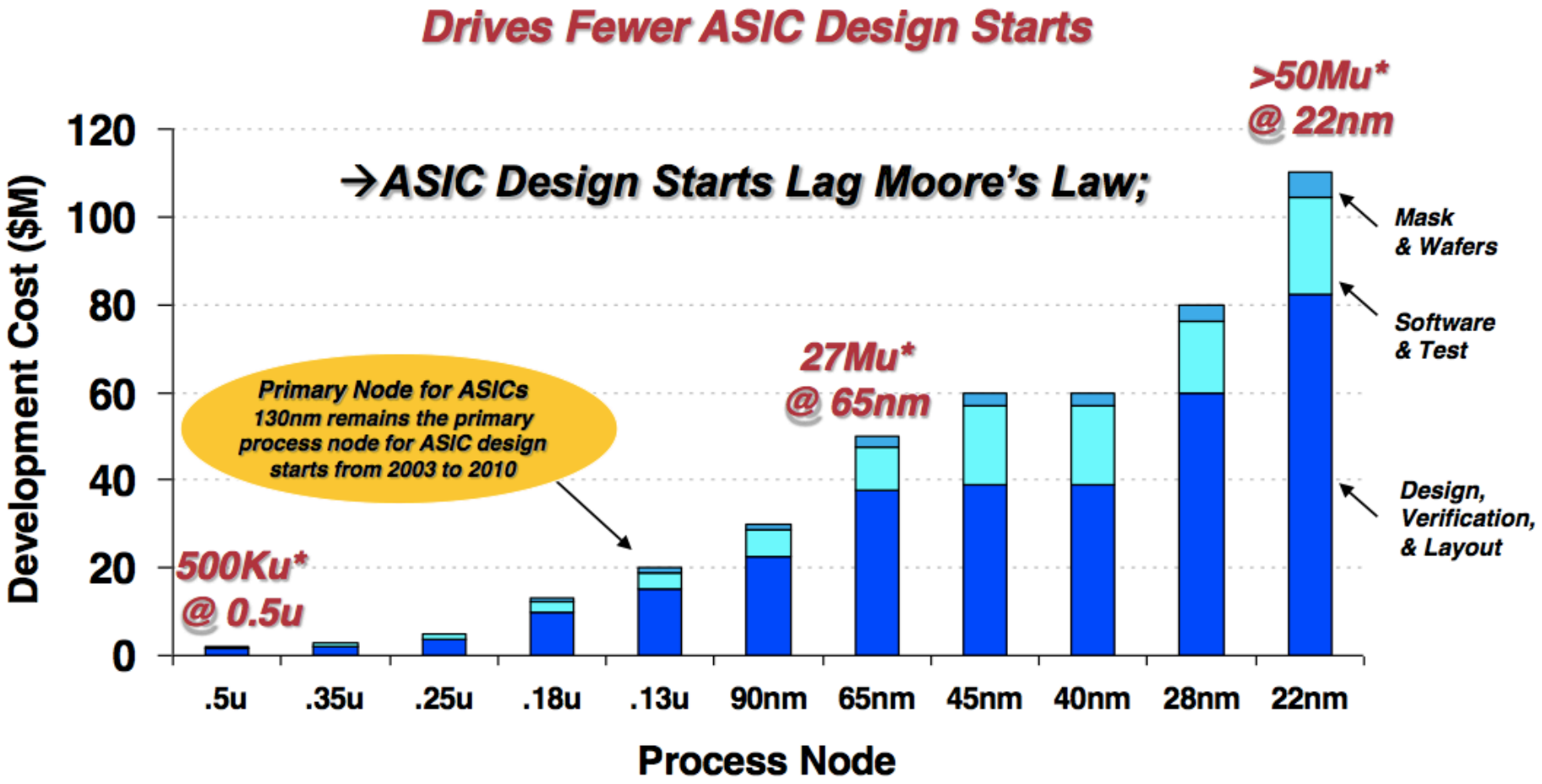
Xilinx 7-series FPGAs, 28nm

DSP Slices	740	1,920	3,600
DSP Performance (symmetric FIR)	930GMACs	2,845GMACs	5,335GMACs
Transceiver Count	16	32	96
Transceiver Speed	6.6Gb/s	12.5Gb/s	28.05Gb/s
Total Transceiver Bandwidth (full duplex)	211Gb/s	800Gb/s	2,784Gb/s
Memory Interface (DDR3)	1,066Mb/s	1,866Mb/s	1,866Mb/s
PCI Express® Interface	x4 Gen2	x8 Gen2	x8 Gen3
Analog Mixed Signal (AMS)/XADC	Yes	Yes	Yes
Configuration AES	Yes	Yes	Yes
I/O Pins	500	500	1,200

ASIC vs FPGA Cost



ASIC Development Costs Today



Source: Altera

Return on Investment Analysis

ASIC / ASSP (65nm) Development Cost **$55M**

20% of Revenue on R&D

Revenue Target **$275M**

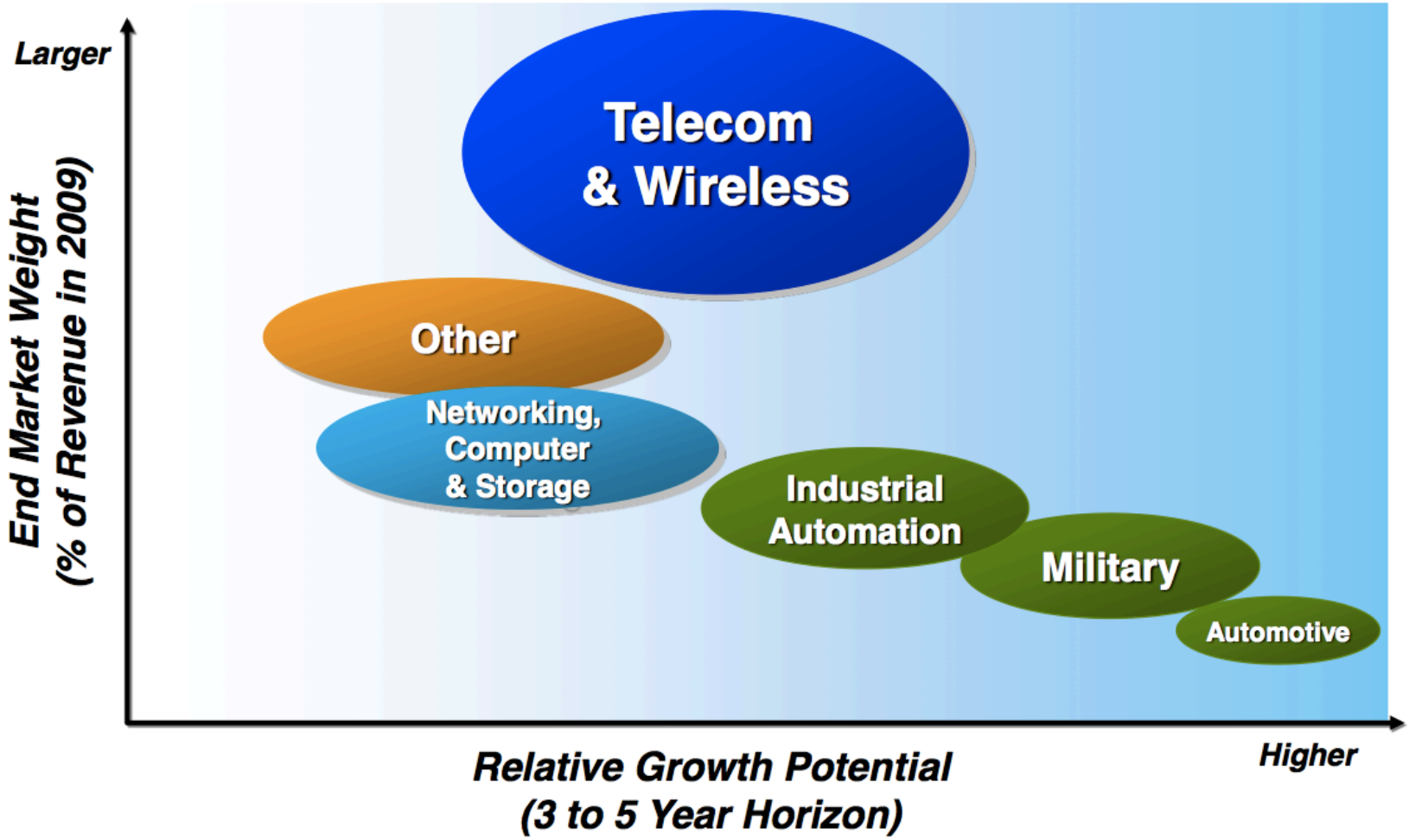
$10 to $50 Device ASP

Unit Volume Req'd ..... ***5Mu to 27Mu***

Very Few High Volume Applications Justify ASIC / ASSP Development

Source: Altera

Application Domains



Source: Altera

Typical High Performance Commercial Applications

Application

Optical Transport
OTU Transponder

40GbE/100GbE Switch

Radar

Requirements

- >350 MHz performance
- 28 Gbps transceivers
- 10GBASE-KR backplane support
- High-performance on-chip memory
- High-performance and flexible memory controller
- Hard system-level IP for bandwidth
- High precision DSP

Solution

ALTERA. Stratix V

Process: 28HP
- >350 MHz performance
- Lowest power in its class
- Up to 1.1M LEs on a monolithic die

Transceiver: 14.1 Gbps/28 Gbps

Product Architecture:
- Soft memory controller supports 800MHz DDR3 DIMM
- 2,560 M20K memory blocks
- 54x54 variable precision DSP

System IP:
- PCIe Gen3 x8, 40 GbE/100 GbE, Interlaken

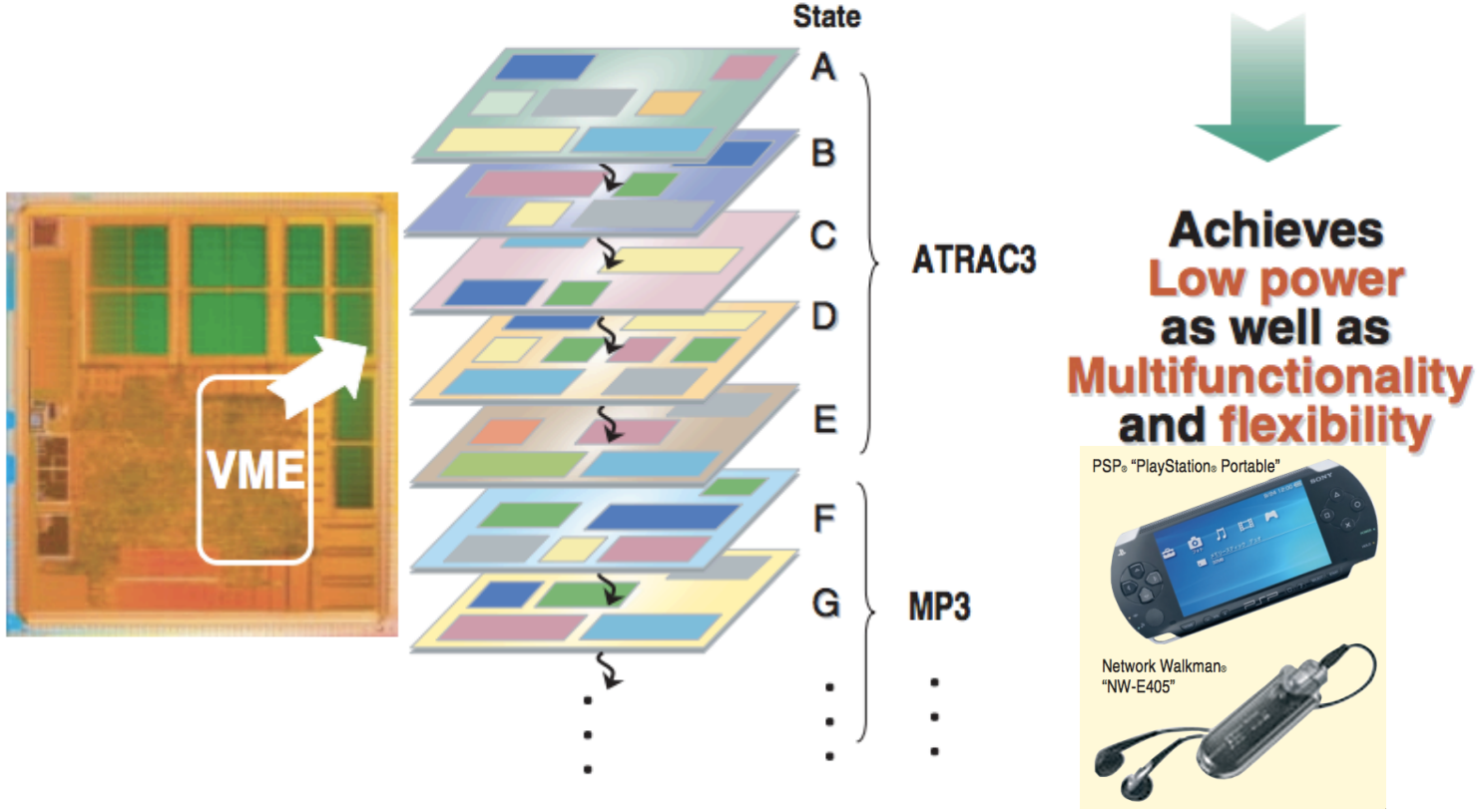
Source: Altera

Sony Virtual Mobile Engine 1

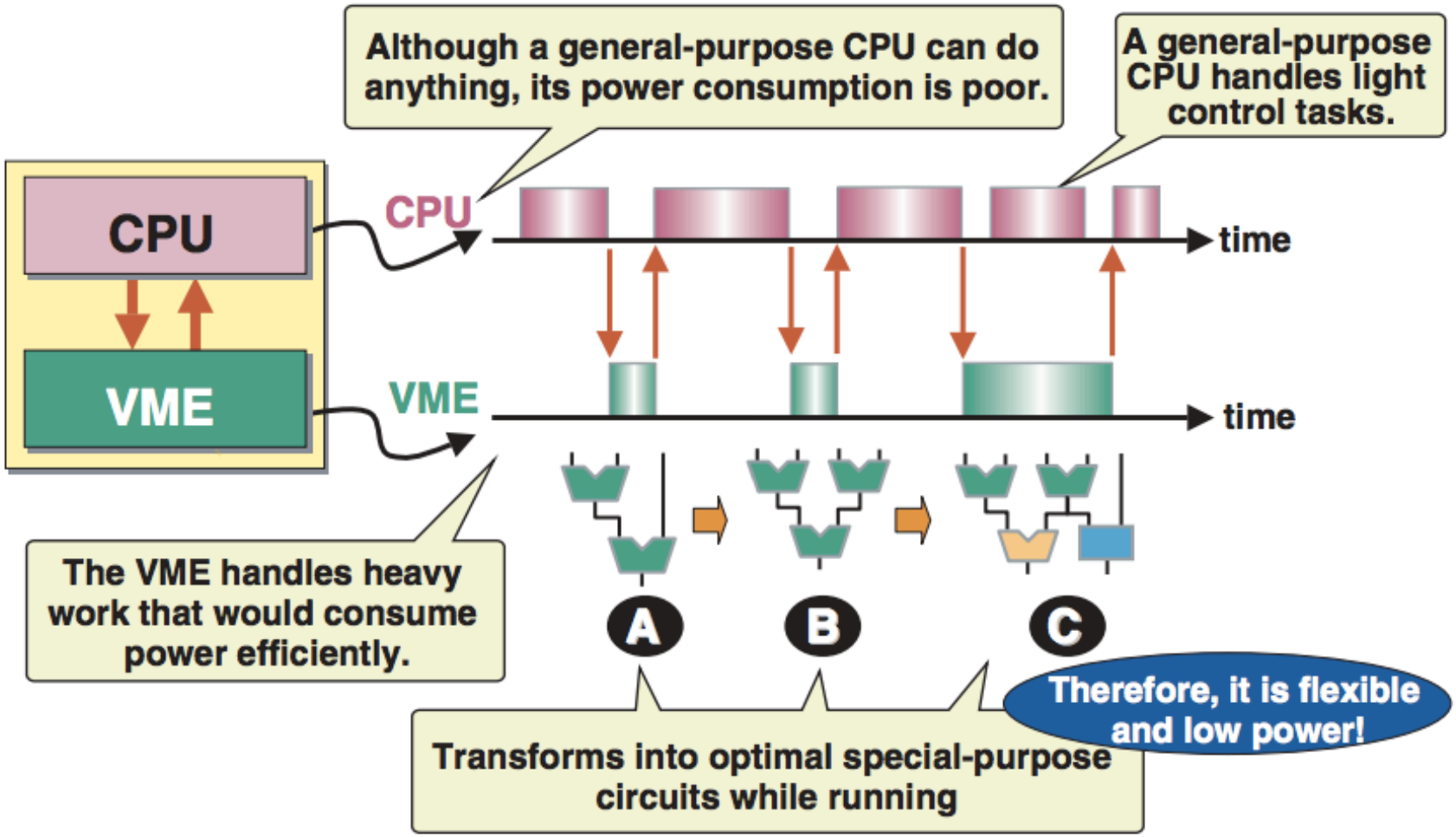
Dynamic Reconfiguration

Semiconductor devices that can optimally transform themselves in software
(Reconfigurable technology)

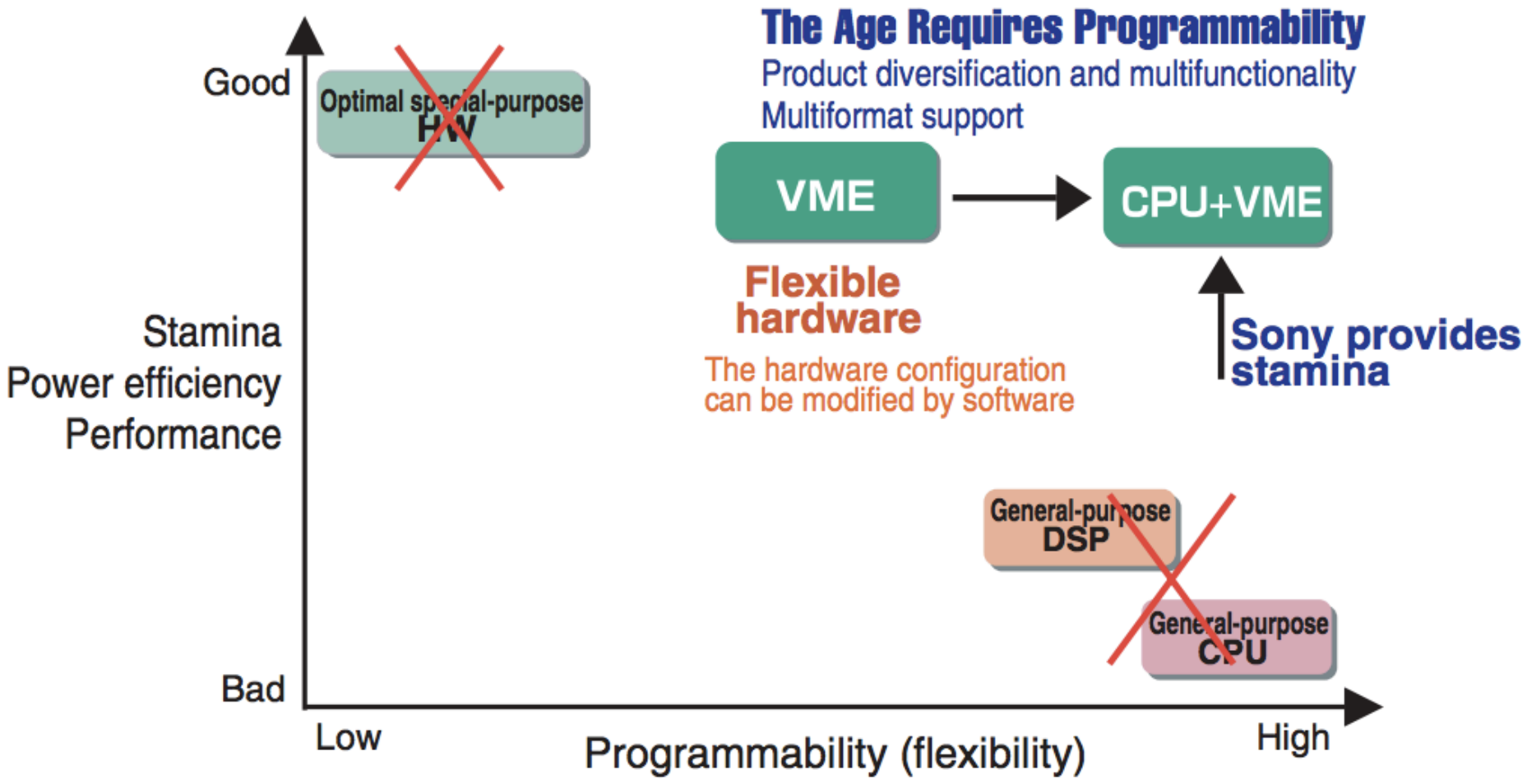
– Conceptual examples –



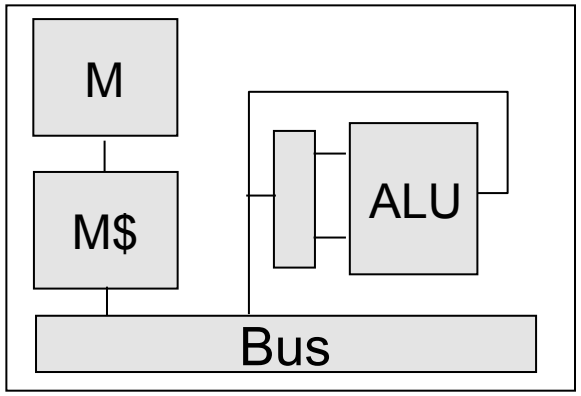
Sony Virtual Mobile Engine 2



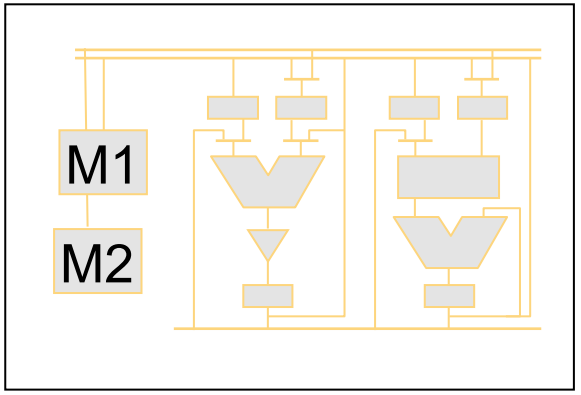
Flexibility vs Performance



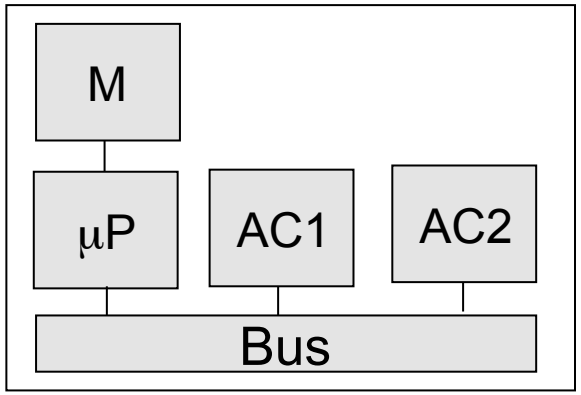
Architectural Choices



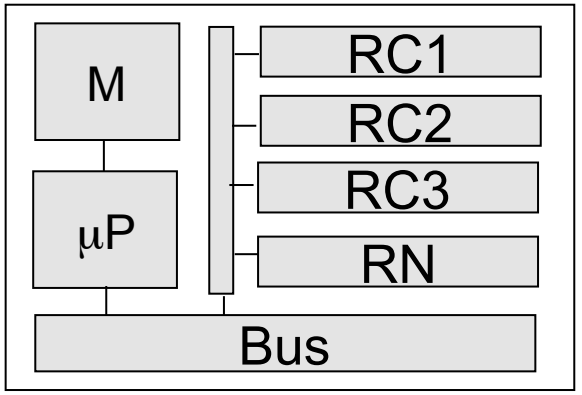
General-purpose processor



Application-specific processor



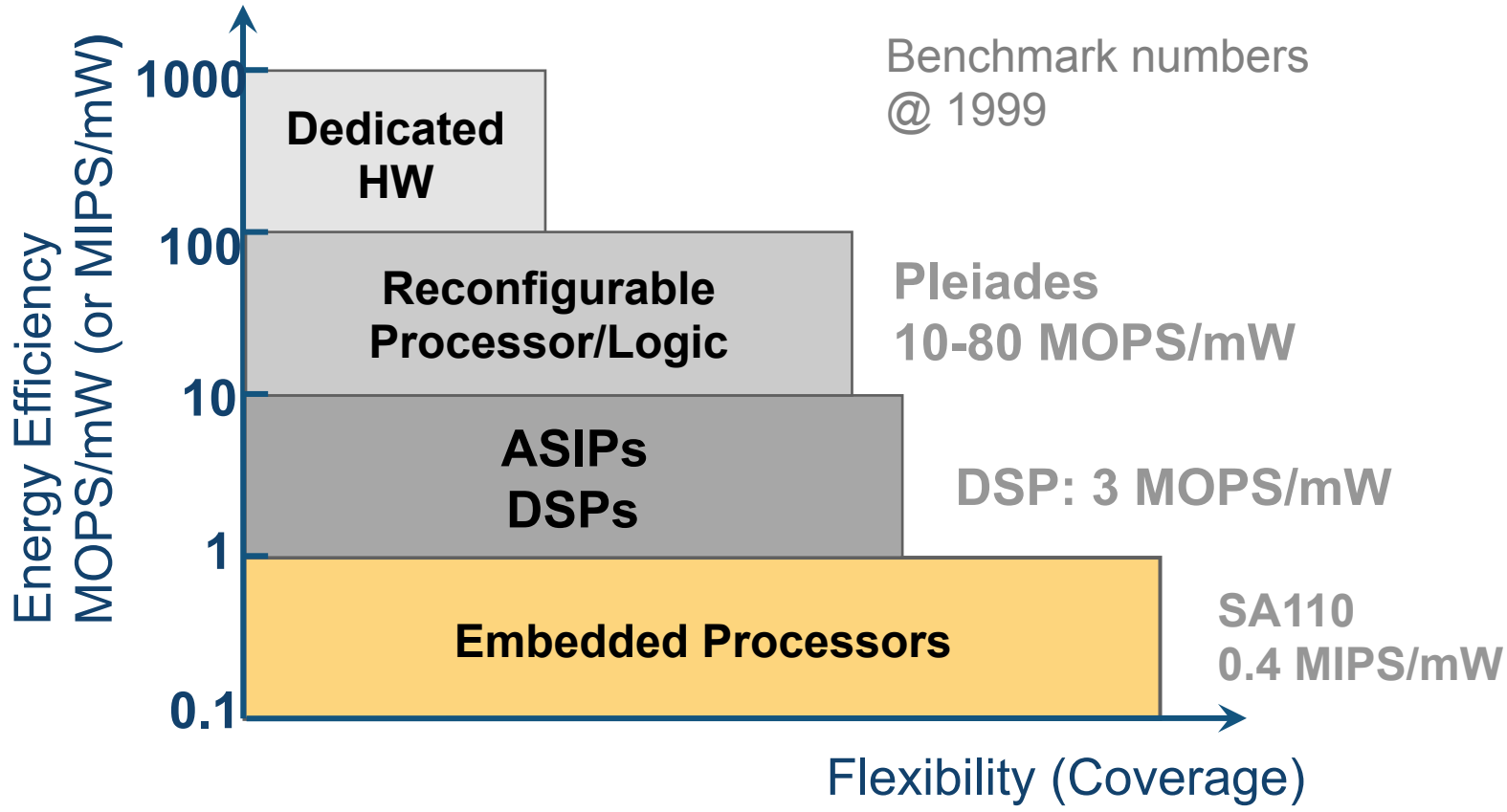
Dedicated accelarators



Reconfigurable processor

Source: Rabaey UCB

Flexibility vs Energy

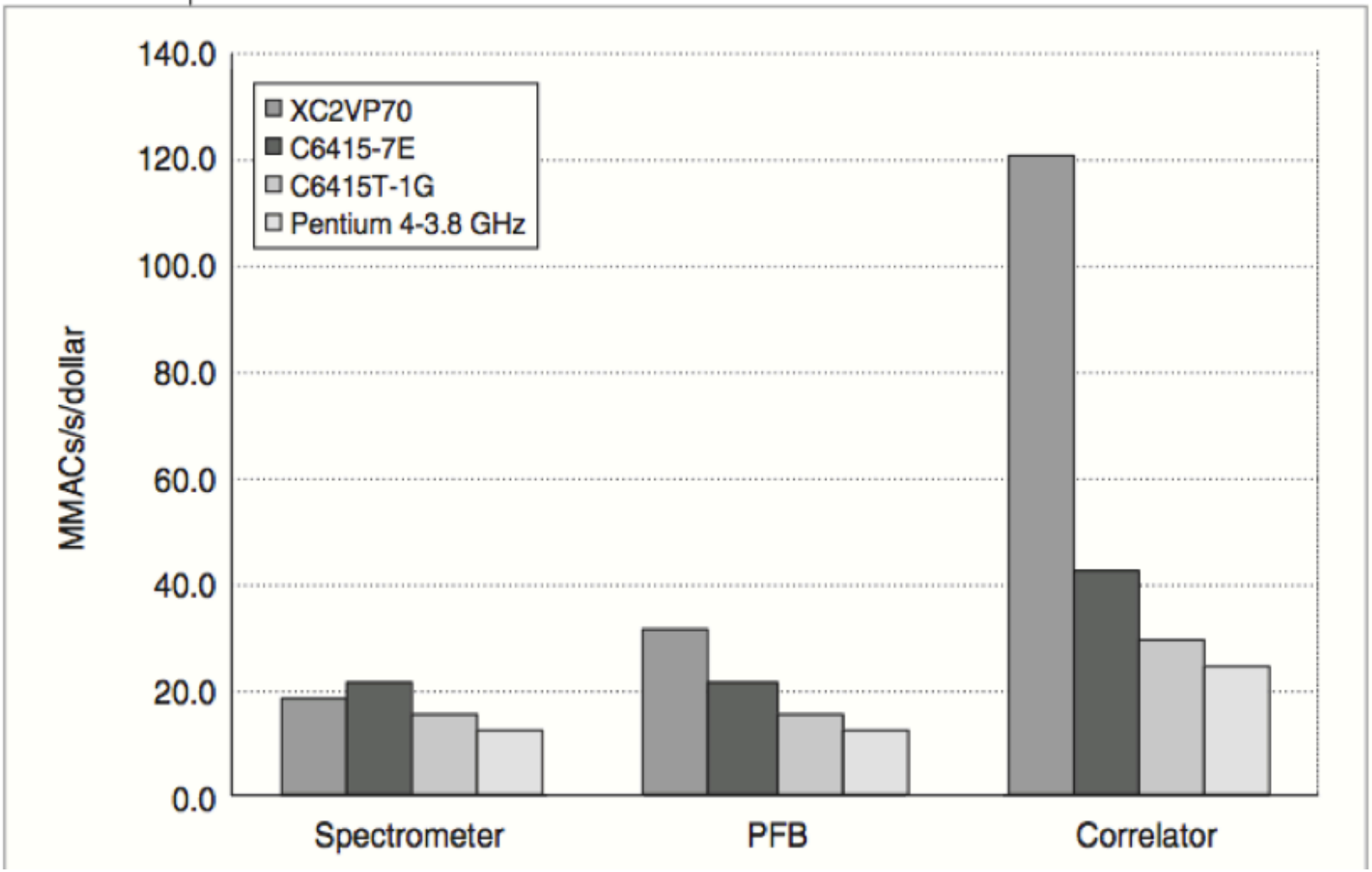


Approximately three orders of magnitude in inefficiency from general-purpose to dedicated!

Source: Rabaey UCB

FPGA vs DSP and CPU Cost Comparison

Berkeley BEE2 cost comparison (FPGA, DSP1, DSP2, uP)



Tools

- Traditionally designed using ASIC development tools
  - VHDL/Verilog very low level
- Recent advances
  - Vivado HLS
  - OpenCL
- Extensive module generators and libraries e.g. filters, fft, floating-point, maths coprocessors, soft processors, network controllers, memory controllers, I/O controllers …
- Still an active research topic

	Hand-coded VHDL	Vivado HLS C
Design Time (weeks)	12	1
Latency (ms)	37	21
Memory (RAMB18E1)	134 (16%)	10 (1%)
Memory (RAMB36E1)	273 (65%)	138 (33%)
Registers	29686 (9%)	14263 (4%)
LUTs	28152 (18%)	24257 (16%)

Resource utilization example: hand coded versus Vivado HLS.

THE UNIVERSITY OF SYDNEY

Comparison of FPGAs with uP and ASIC

- Compared with uP and DSP
  - higher speed, lower power, smaller variance in execution time
  - Longer development times, higher cost per unit
- Compared with ASICs
  - Lower initial cost
- Rides Moore's Law, development costs amortised over users
  - Faster time to market, lower risk
  - Can be customised to problem in ways not possible with ASICs

Overview

- FPGAs
- **Reconfigurable computing**
- Applications

Reconfigurable Computing

- Application of FPGA devices to computing problems

Acceleration

- FPGAs allow computational problems to be accelerated through
  - Parallelism
  - Customisation
  - Integration

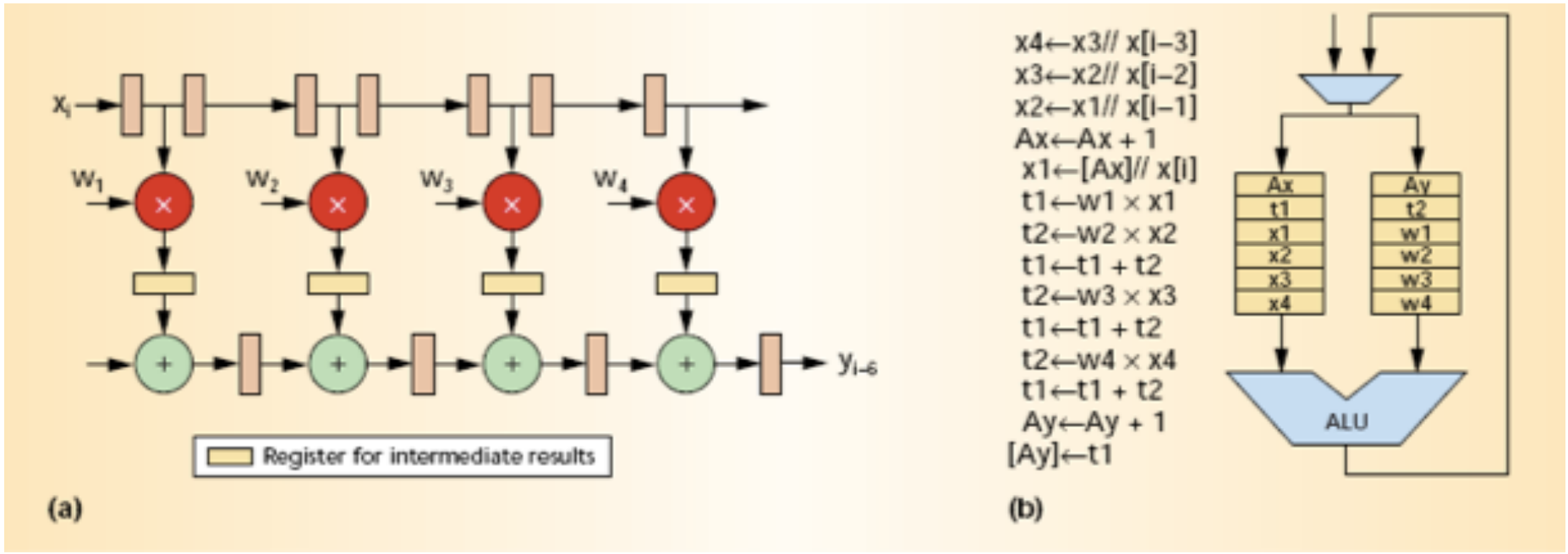
Parallelism

- Do what would take many cycles on uP in fewer cycles (instruction level parallelism)
- Do many independent tasks/threads/processes in parallel (multiprocessor)
- Tradeoff latency with throughput by doing things in stages (pipelining)

http://fernandoexperiences.blogspot.com.au/

Parallelism Example

- Microprocessor: data passed sequentially to computing unit
- FPGA & ASIC: spatial composition of parallel computing units (multiple muls, pipelining)
- E.g. 4-tap FIR filter, FPGA 1 output per cycle, uP takes multiple cycles
- Lower power and higher speed



Source: DeHon “The Density Advantage of Configurable Computing”

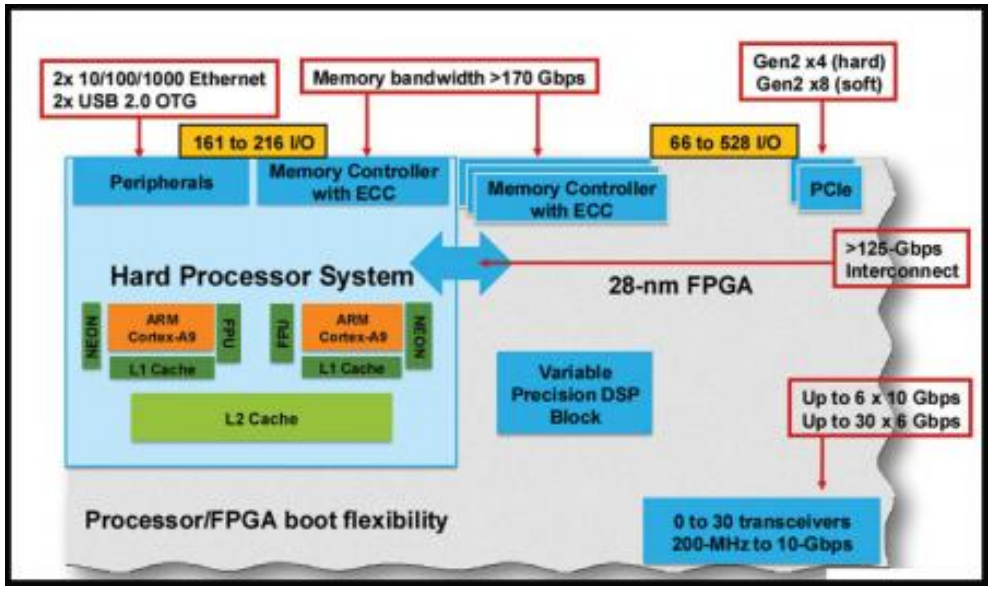
Customisation

- More specific functions can be implemented more efficiently
- Too expensive to design ASIC to perform very specialised function
- FPGAs can be heavily customised due to their programmability i.e. only do one thing efficiently
  - Tradeoffs between speed and accuracy can be exploited, on uP, only get single or double; char, short or long
  - General operators can be replaced with specific ones
- E.g. Chip which only encrypts for a specific password

Integration

› Networking, chip IO and computation on same device

- Reduction of buffering can help latency
- Single chip operation massive interconnect within chip exploited
- Multiple (small) memories within FPGA offer enormous memory bandwidth



Overview

- FPGAs
- Reconfigurable computing
- **Applications**

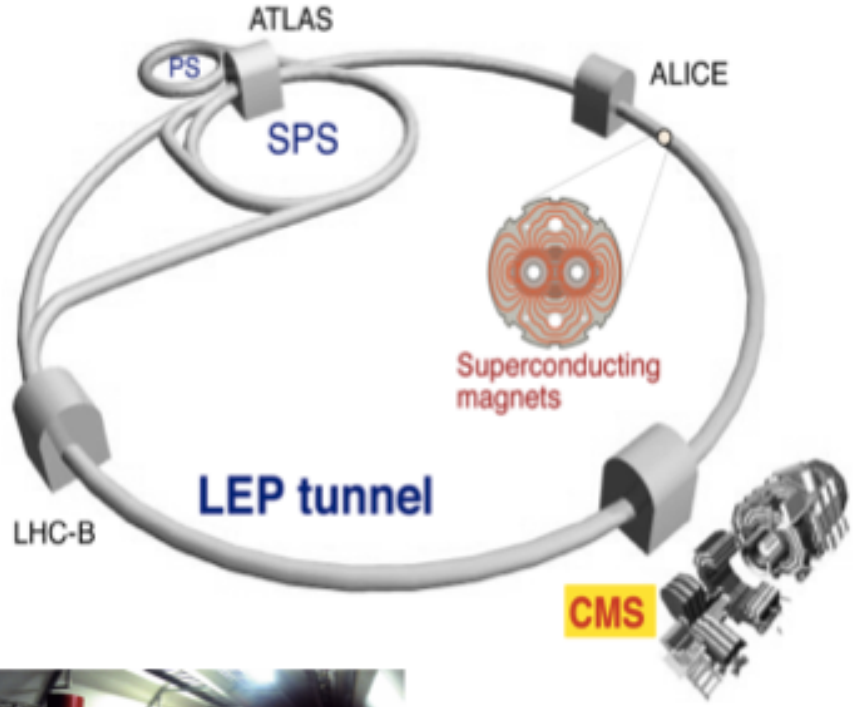
BMW Williams Formula 1 Team 2003

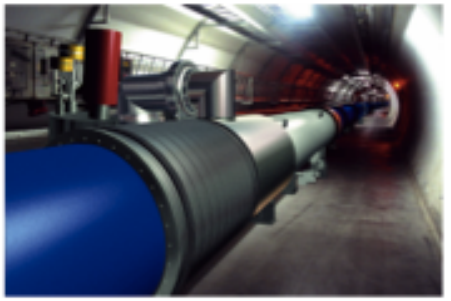
- Vehicle Control Module uses Virtex-II devices
  - gearbox, differential, traction control, launch control and telemetry
- High speed real-time control and DSP application

Source: BMW Williams

CERN Large Hadron Collider

- Compact Muon Solenoid
  - 10^{15} collisions per second
  - Few interesting events ~ 100 Higgs events per year
  - 1.5Tb/s real-time DSP problem
  - More than 500 Virtex and Spartan FPGAs used in real-time trigger





Source: Geoff Hall, Imperial College

Square Kilometer Array

- Square Kilometre Array (SKA) will be one of the largest and most ambitious international science projects ever devised (€1.5 billion).
- CSIRO Developing Australian SKA Pathfinder (ASKAP), a $150M next- generation radio telescope using FPGA technology for the data collection & processing

Source: CSIRO

THE UNIVERSITY OF SYDNEY

Other RC Applications

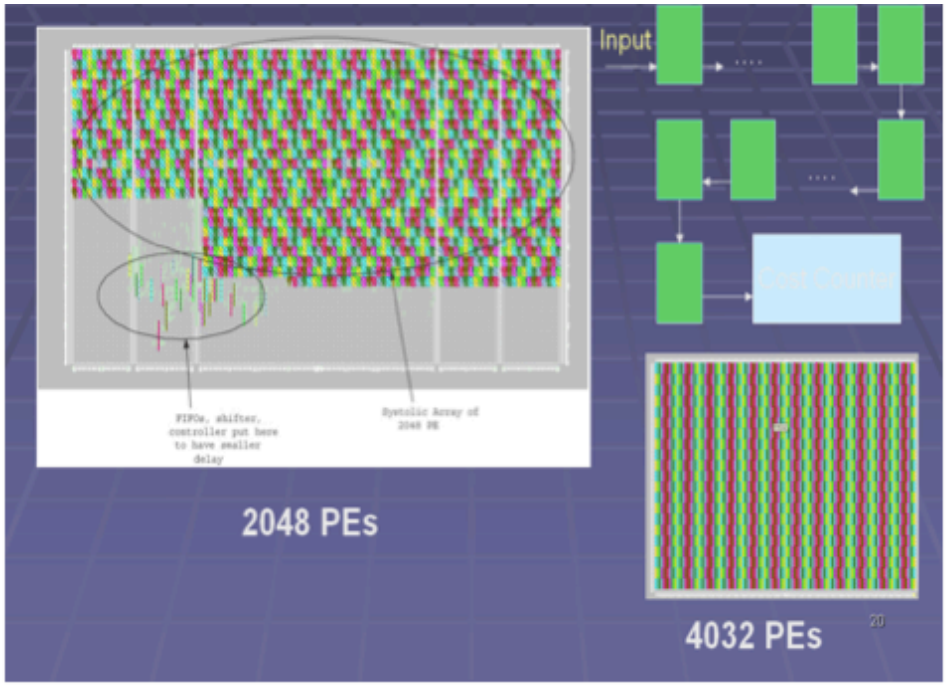
- Applications suited to acceleration
  - seismic processing astrophysics FFT
  - adaptive optics (transforming to frequency domain and removing telescope image noise)
  - biotech applications such as BLAST, Smith Waterman and HMM
  - financial modeling
- Functions well suited to FPGA acceleration
  - searching & sorting
  - signal processing (audio/video/image manipulation)
  - encryption
  - error correction
  - coding/decoding
  - packet processing
  - random-number generation for Monte Carlo simulations

Source: cray.com

Overview

- FPGAs
- Reconfigurable computing
- **Applications (some of our work)**

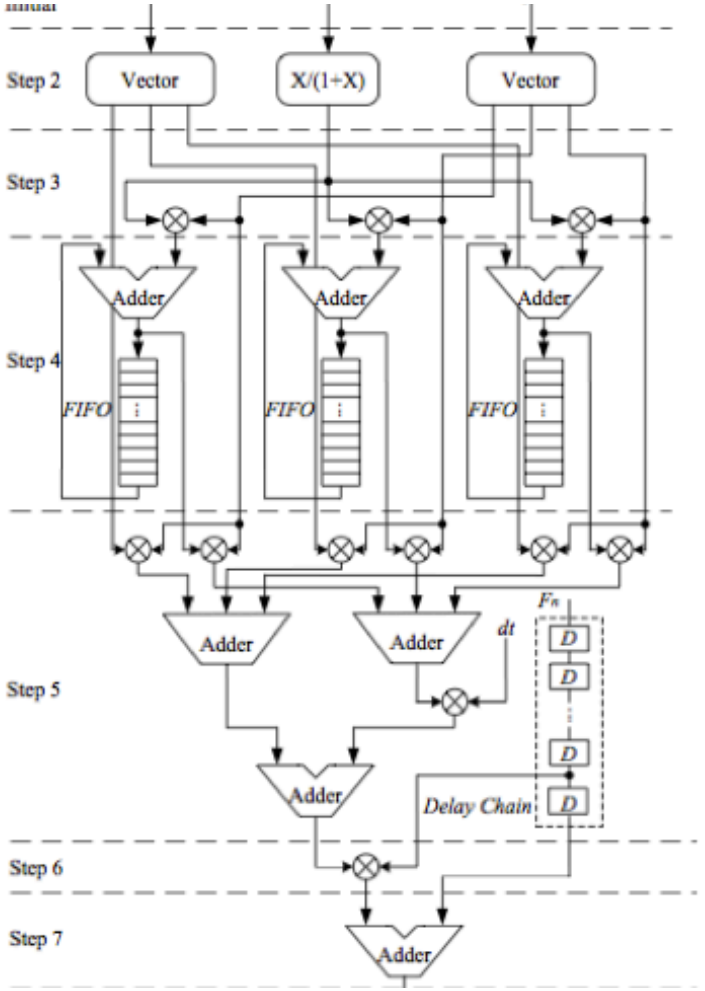
DNA Sequence Alignment (FPL03)



System	Device	Performance (billion CUPS)
Splash 2	16× XC4010	2.7
Paracel	ASIC	1.9
Celera (software)	800 DEC Alphas	0.3
Our work	XCV1000	742

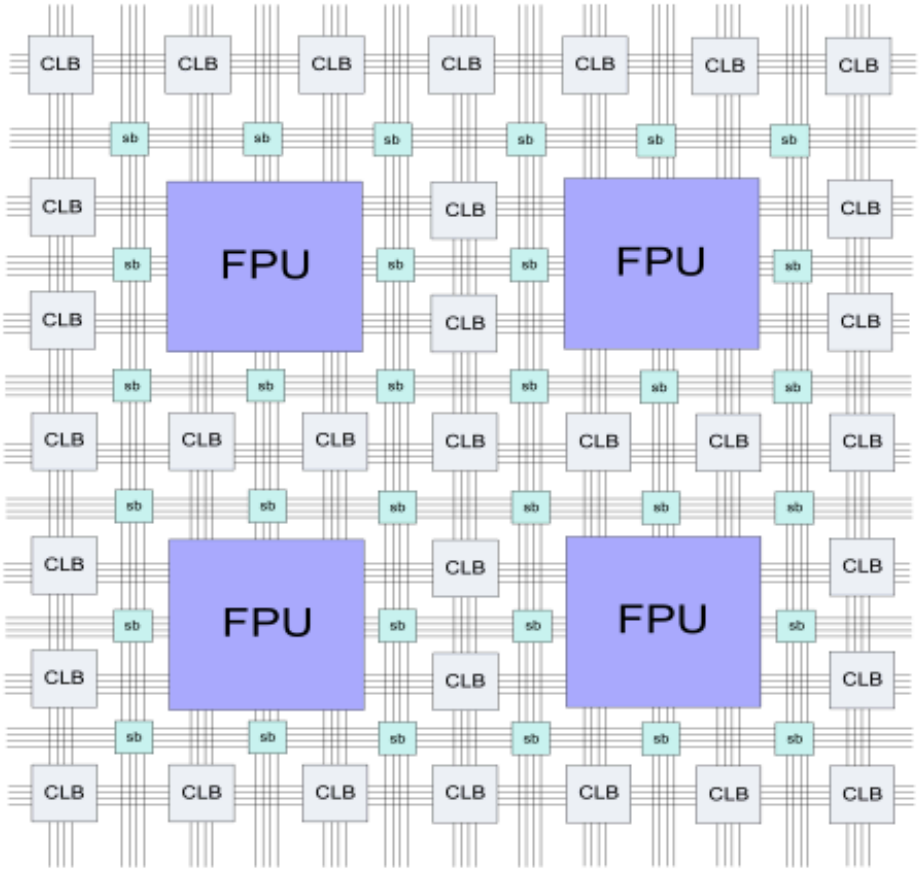
Acceleration of Derivative Pricing (FPT05)

- Pricing of interest rate derivatives computationally expensive
- Pipelined BGM implementation 25x speedup over PC (much lower power)
  - Novel Gaussian random number generators
- First work in this area, banks now making their own implementations



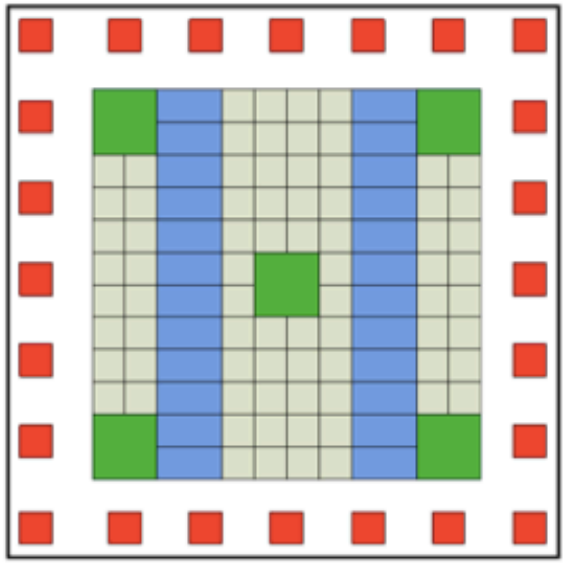
Floating Point FPGAs (FPL07)

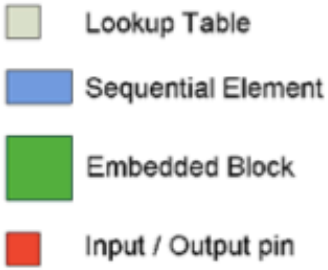
- Fixed/Floating point FPGAs
- Specialising reduces flexibility but improves design time, density, speed and power

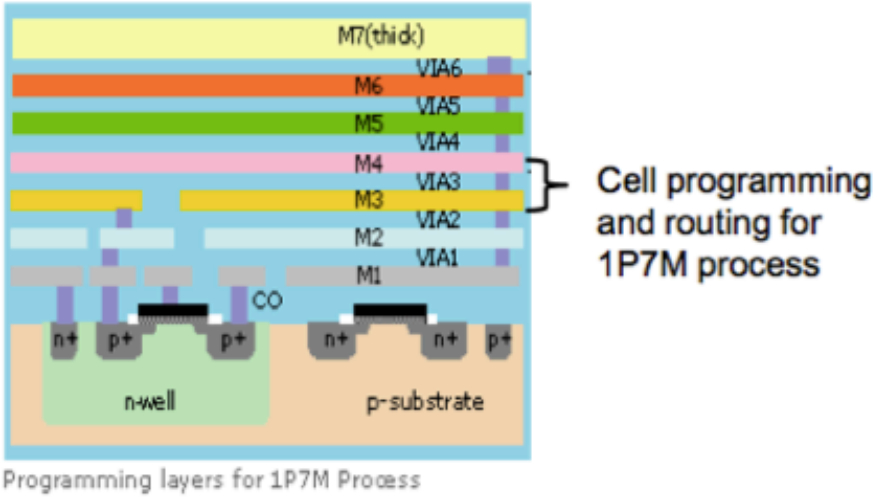


Structured ASIC I (DAC10 user track)

- ASIC mask costs becoming prohibitively high
- Structured ASIC: Cadence tools, 3 masks





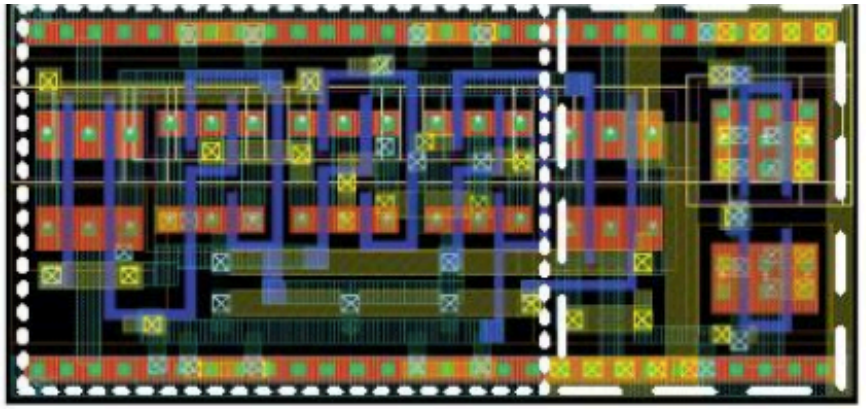


Programming layers for 1P7M Process

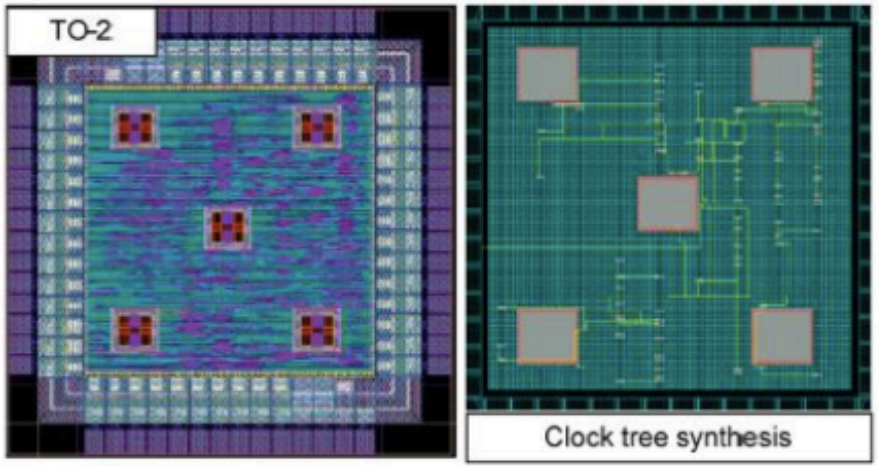
Source: http://www.faraday-tech.com/html/products/structuredASIC.html

Structured ASIC II (DAC10 user track)

- Active dynamic LED backlight controller
  - UMC 0.13um 1P8M2T, 100MHz, 3.6K gates, 0.16 mm^2
  - 0.7x speed and 4x area of ASIC
  - Fully tested

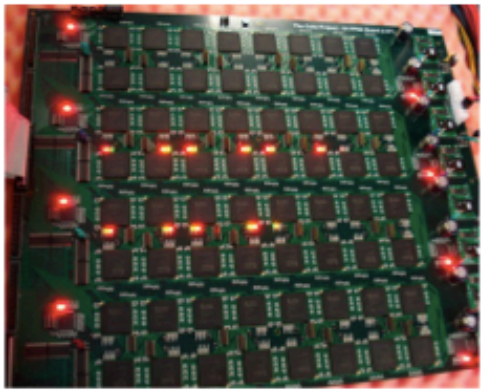


Screen capture of configurable cell in transmission gate technology



THE UNIVERSITY OF SYDNEY

Cube System I (SPL09)



Massively RC system for cryptanalysis; simulation; regular expression engine; network scanner; multimedia; ...

Cube System II (SPL09)

- Used for RC4 key search
- 96 cores on XC3S4000
- Cube system was 8x faster, 86x less power, 20x less volume than equivalent Intel Xeon cluster

THE UNIVERSITY OF SYDNEY

Our Research Areas

- Tools
  - Ways to analyze floating point precision requirements
- Applications
  - Machine learning
  - Signal Processing
- Architectures
  - Floating point FPGAs
  - Runtime reconfiguration: virtual hardware
  - Late binding to address process variation

THE UNIVERSITY OF SYDNEY

Current Research

- Hedging of foreign exchange risk
- Spherical Microphone Array
- Rounding Error Analysis
- Online Machine Learning
- SKAMP

THE UNIVERSITY OF SYDNEY

Overview

- **Hedging of foreign exchange risk**
- Spherical Microphone Array
- Rounding Error Analysis
- Online Machine Learning
- SKAMP

Hedging of Foreign Exchange Risk

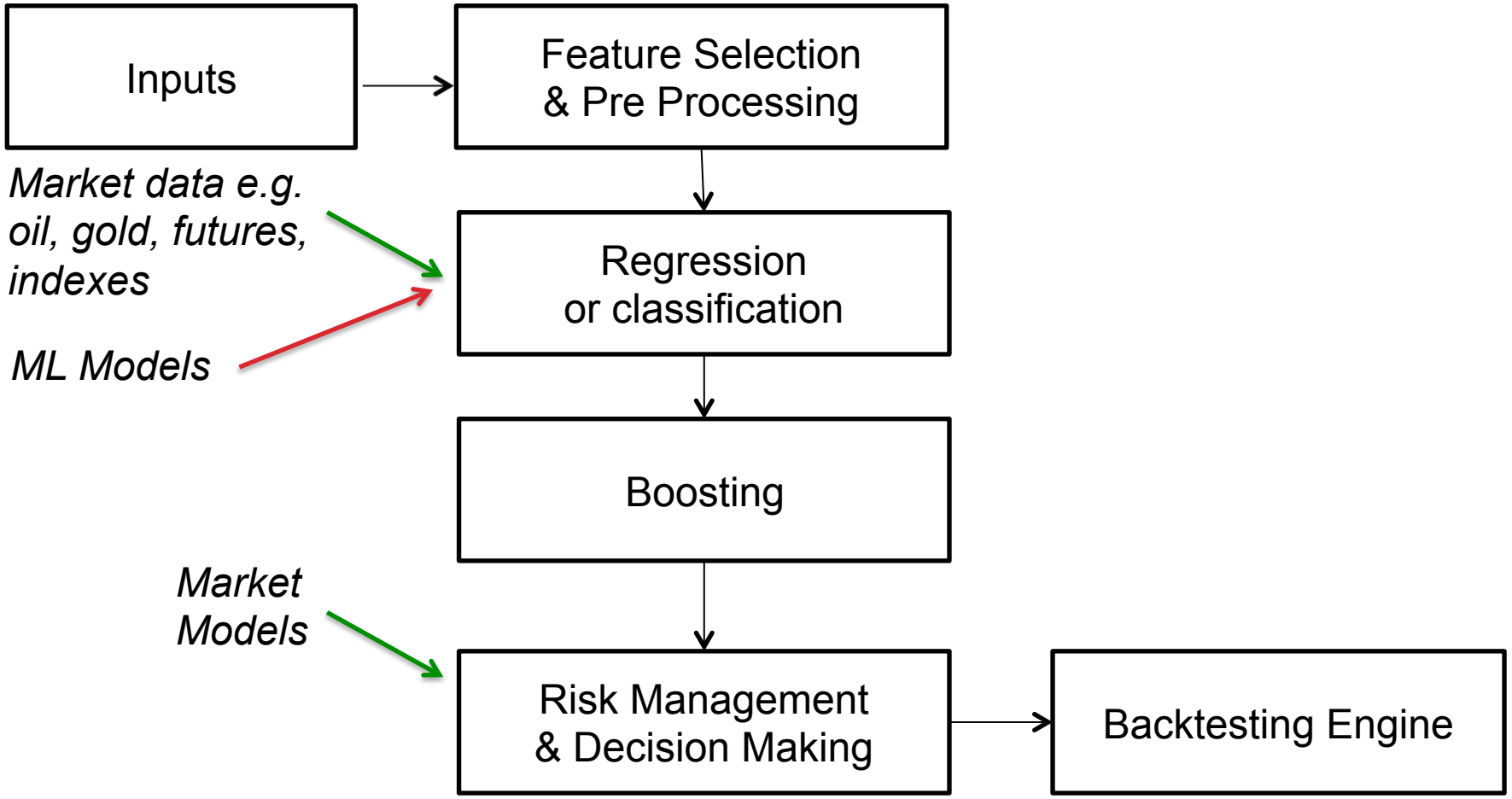
- Project starting 2012 sponsored by Westpac
- Apply parallel computing and machine learning techniques to better understand and manage exposure to FX risk
  - Software environment for the testing of risk management strategies
  - Interface to scalable cloud computing resources
  - Predict customer flow and exchange rates
  - Develop hedging strategies and market models
- Project goals are to develop software system but we are also looking at opportunities for FPGA-based acceleration
- Enable Australian banks to better quantify and manage risk, making them more competitive in global FX markets

THE UNIVERSITY OF SYDNEY

The Problem

- Alice buys $0.969M AUD using $1M USD
- Bank buys $1M USD
- AUDUSD exchange rate falls
- Bank loses money (if position large)
- Need to understand and hedge risk

System Architecture

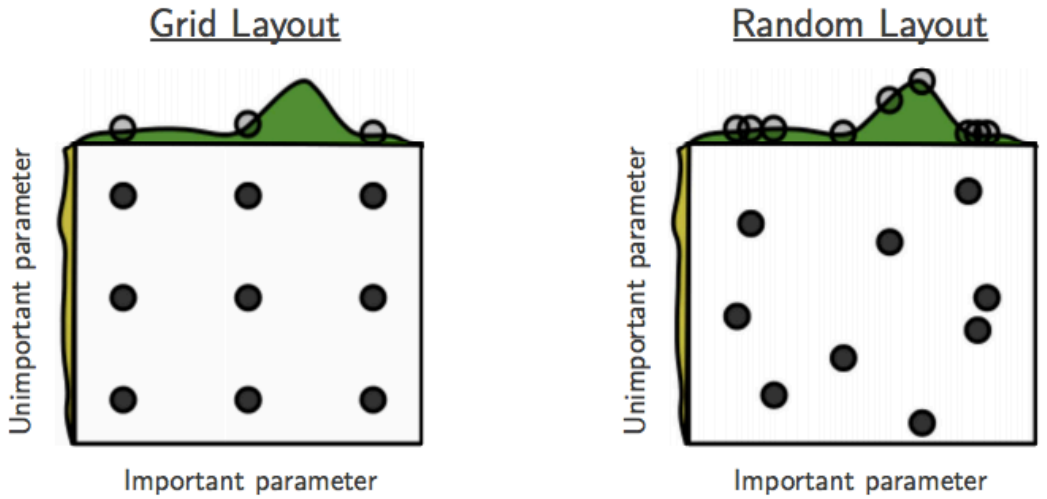


› Cloud Computing Used to Parallelise Simulations

Machine Learning Research

› Currently working on

- Feature selection: generate large feature set and select most appropriate ones – RC?
- Irregular time series: point process models for non-periodic time series.
- Hybrid models: Cluster data automatically before regression/classification, random hyperparameter search



Overview

- Hedging of foreign exchange risk
- **Spherical Microphone Array**
- Rounding Error Analysis
- Online Machine Learning
- SKAMP

Spherical Microphone Array

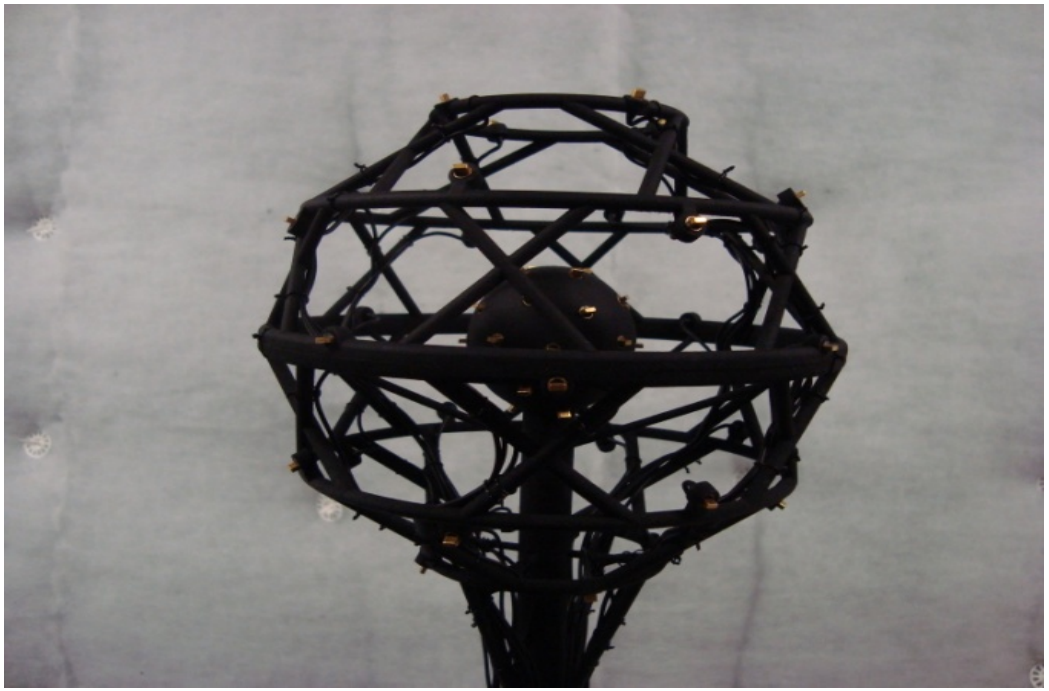


Figure: CARLab dual concentric spherical microphone array

1) First dual concentric SMA of its kind.

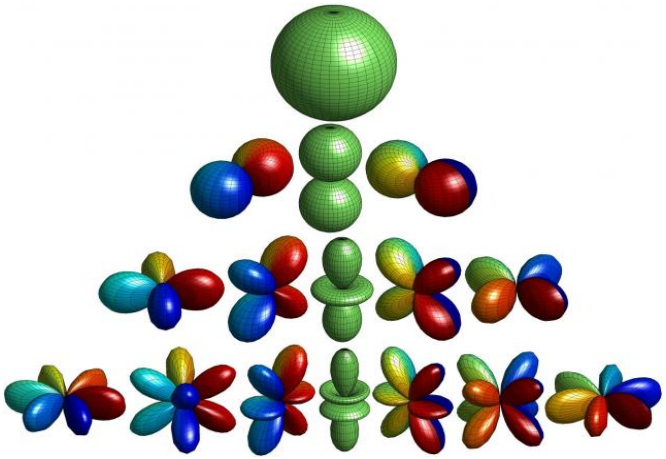
2) 32 microphones on the inner sphere and 32 microphones on the outer sphere.

3) Sound field analyses in the spherical harmonic domain demonstrates interesting advantages for *sound separation, sound localisation, and acoustic holography*.

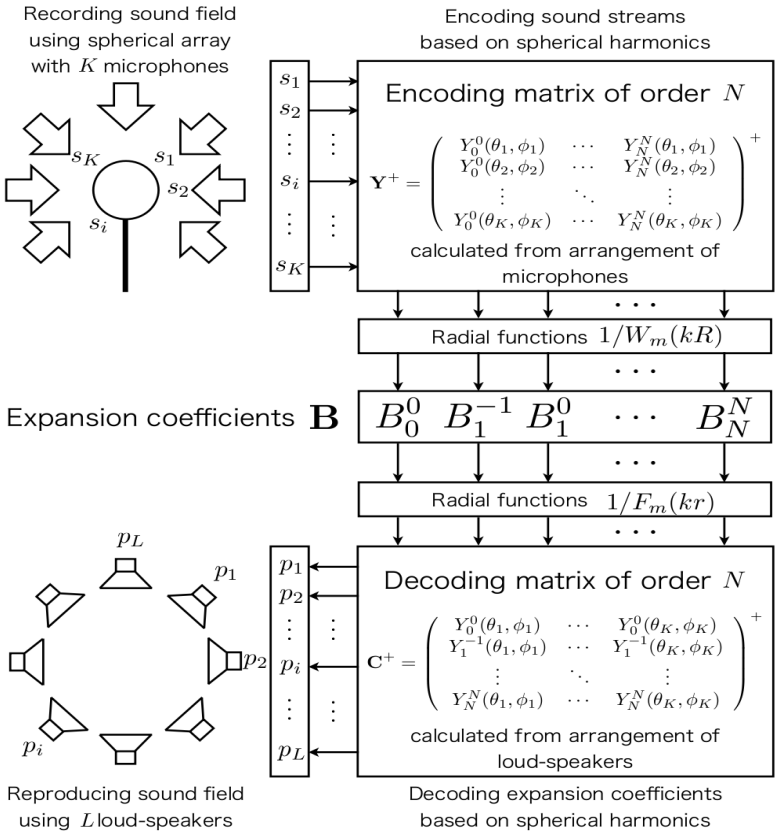
http://www.ee.usyd.edu.au/carlab

Higher order Ambisonics (HOA)

Ambisonics is a series of recording and replay techniques using multichannel mixing technology.



3rd Order ➡ **16 Channels (64 microphones)**



HOA is well suited for network transmission. It is scalable, in which transmission rate can be adapted to the available network bandwidth.

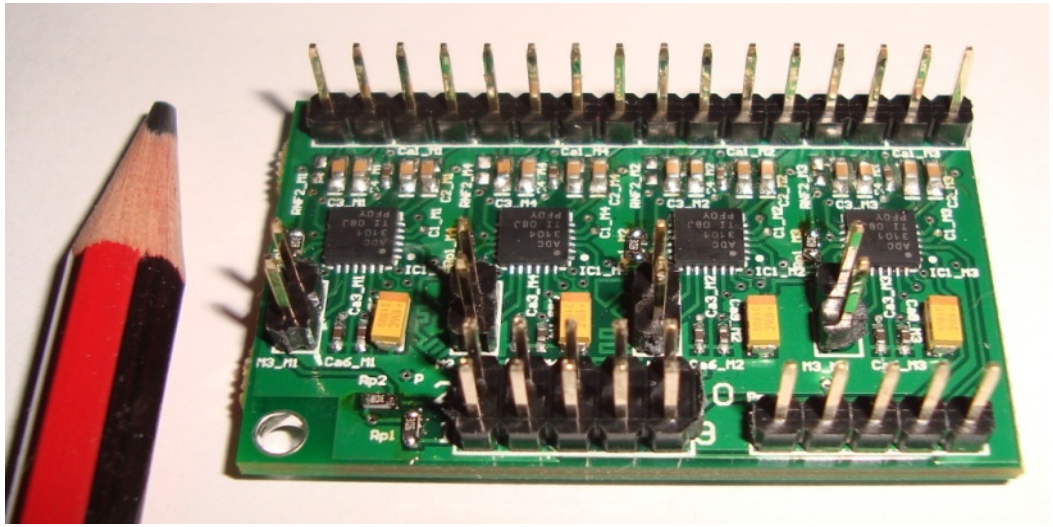
[T. Okamoto et al., 2010]

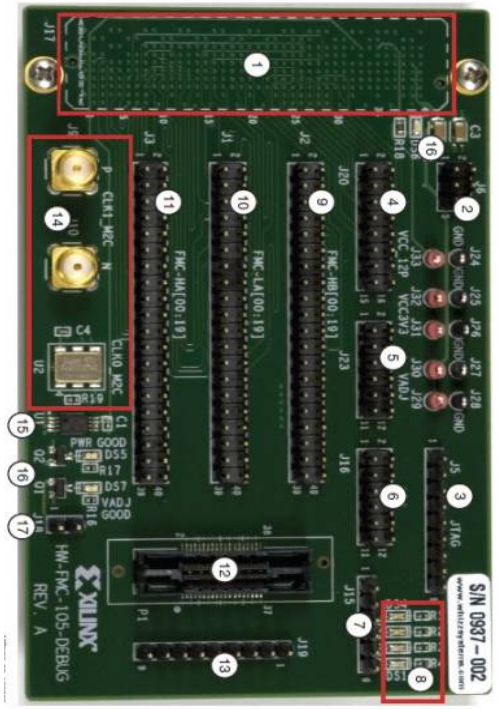
Existing Systems

- Existing systems are high power, bulky and expensive.
- Real-time performance of compressed sensing cannot currently be achieved.
- Developing portable HOA system.

Portable SMA System

System being developed to implement SMA-based audio signal processing on FPGAs.





THE UNIVERSITY OF SYDNEY

Overview

- Hedging of foreign exchange risk
- Spherical Microphone Array
- **Rounding Error Analysis**
- Online Machine Learning
- SKAMP

Rounding Error Analysis (MCA)

- Floating point: $x \oplus y = fl[\, x + y \,] = (x + y)\,(1 + \delta)$
- MCA: $x \odot y$ = round(randomize(x) • randomize(y))
- Executions give different answers => Monte Carlo simulation
- Mean is associative, standard error detects cancellation

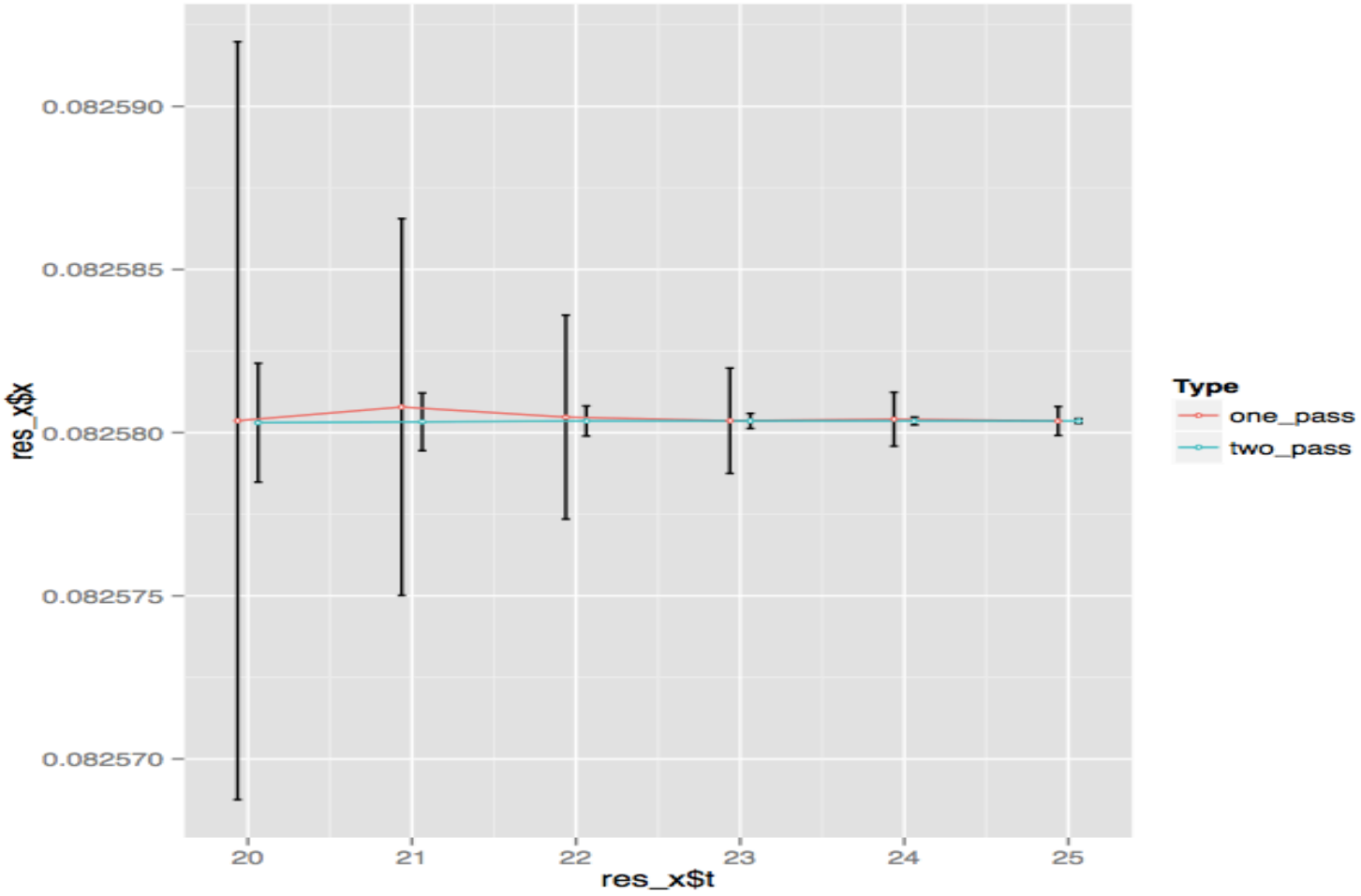
run	r1	r2
1	1240.86	.000198747
2	1240.86	.000248582
3	1240.86	.000251806
4	1240.86	.000177380
5	1240.86	.000203571
computed average:	1240.86	.000216017
standard deviation:	.000	.000032739
standard error:	.000	.000014641

Table 4: Roots of $7\,x^2 - 8686\,x + 2 = 0$, computed with single precision MCA.

Rounding Error Analysis (Source-Source Compiler)

- Developed MCA analysis using the Berkeley CIL Compiler
- 120 lines of ML code allows all FP calls to be intercepted and sent to a library
- MCA library developed
- Other applications of this compiler

Sample Run (Standard Deviation)



THE UNIVERSITY OF SYDNEY

Overview

- Hedging of foreign exchange risk
- Spherical Microphone Array
- Rounding Error Analysis
- **Online Machine Learning**
- SKAMP

Online Machine Learning

- Most ML batch based e.g. SVM requires new convex optimisation problem to be solved upon receiving a new input
- Kernel recursive least squares [Engel et. al. 2004]
  - sequential sparsification process admits into the kernel representation a new input sample only if its feature space image cannot be sufficiently well approximated by combining previously admitted samples

KRLS Architecture

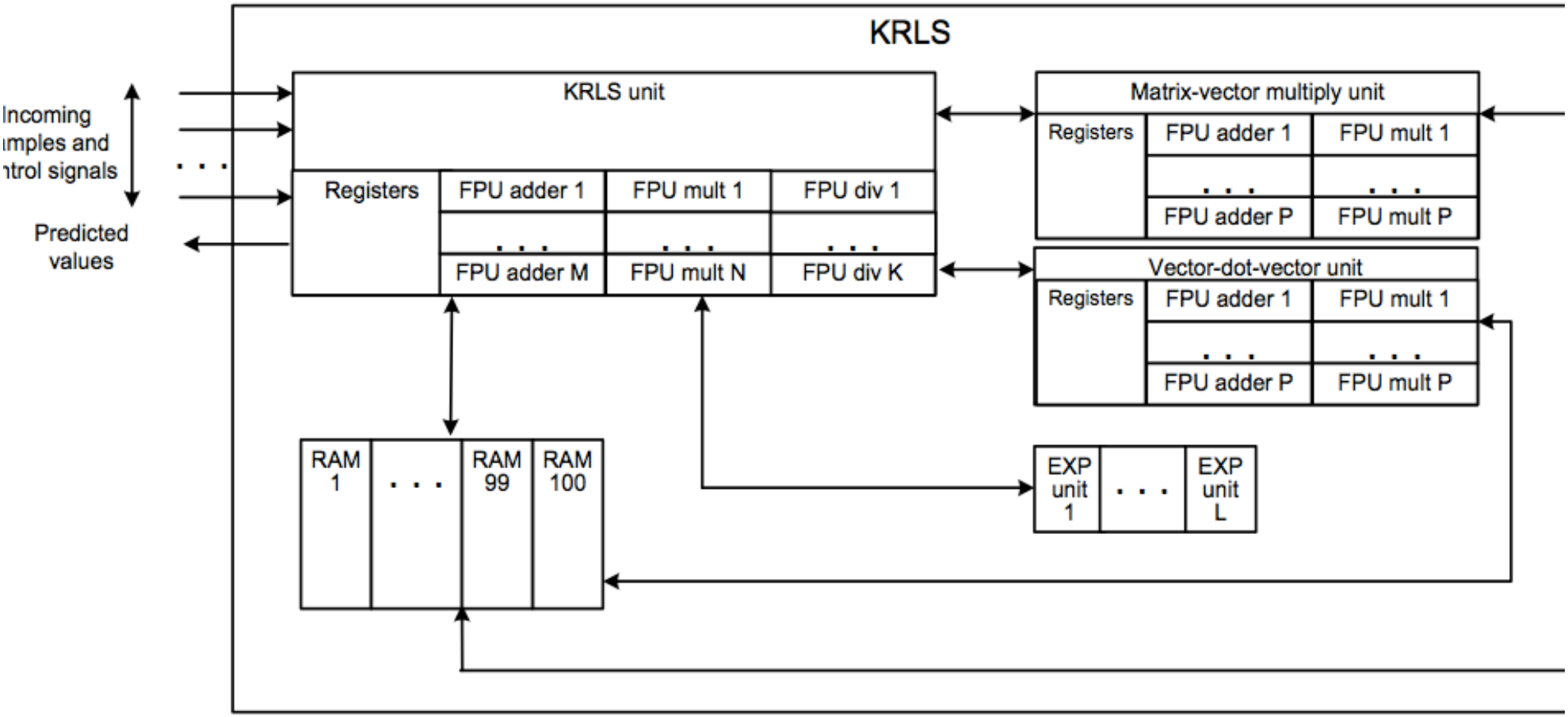
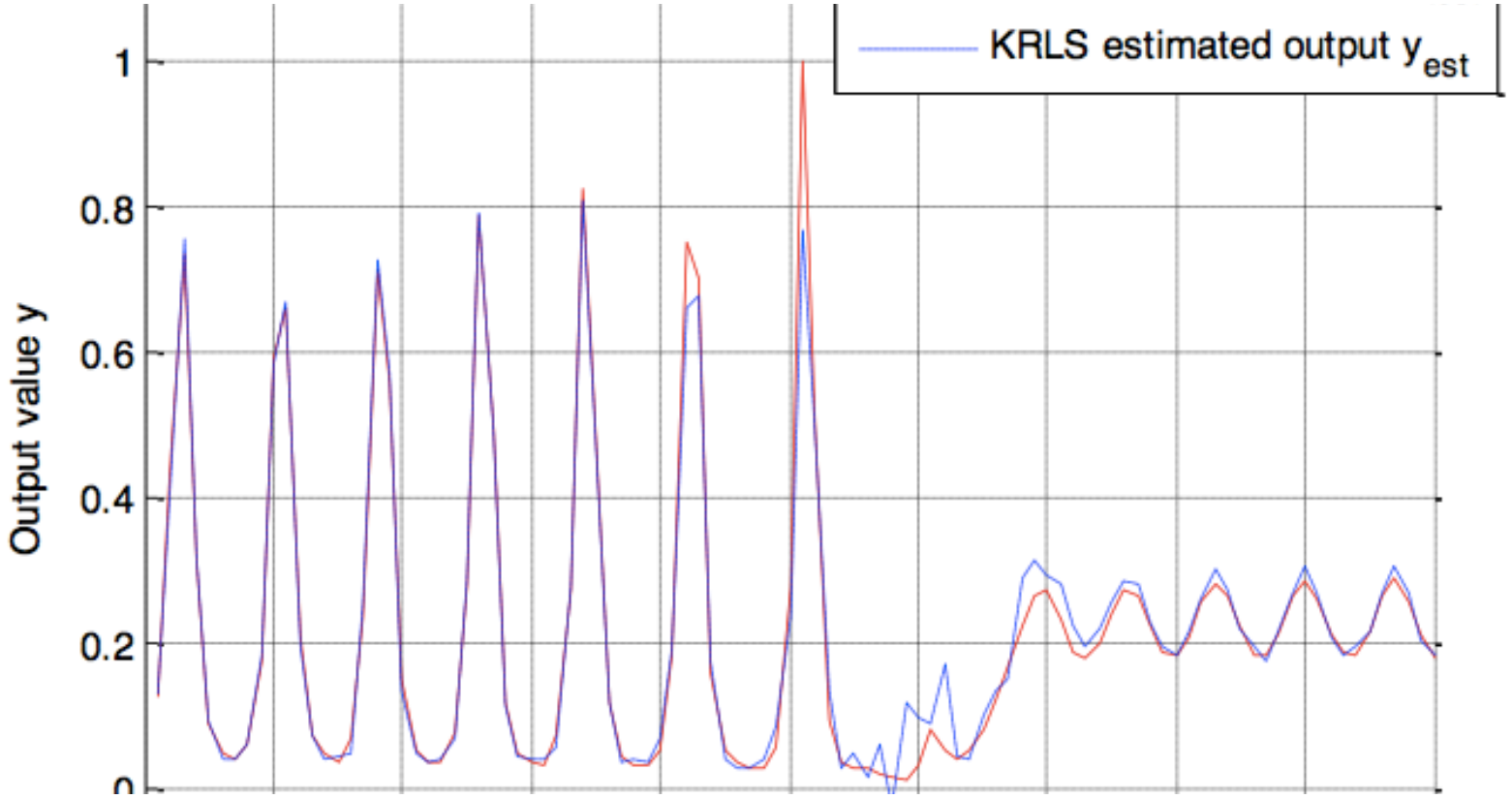


Figure 16: KRLS algorithm hardware top-level design (parallel implementation)

Performance on Sante Fe Competition Laser Dataset



KRLS implementation	Training task average latency (clock cycles per sample)	Prediction task average latency (clock cycles per sample)	Speed up factor (over PC)
C language (software)	264345	22927	1
FPGA parallel	64184	4936	4.1 - 4.6

Table 8: FPGA and software implementations average latencies for training and prediction tasks (laser dataset)

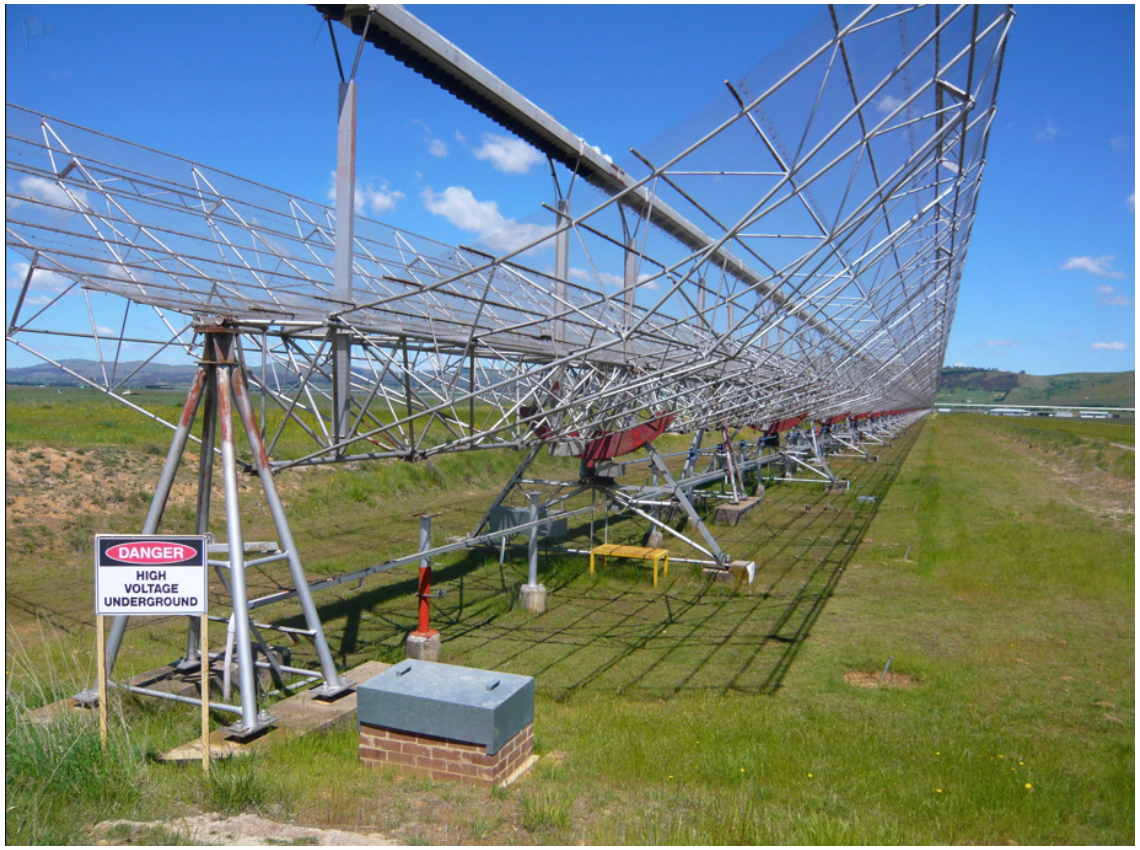
Overview

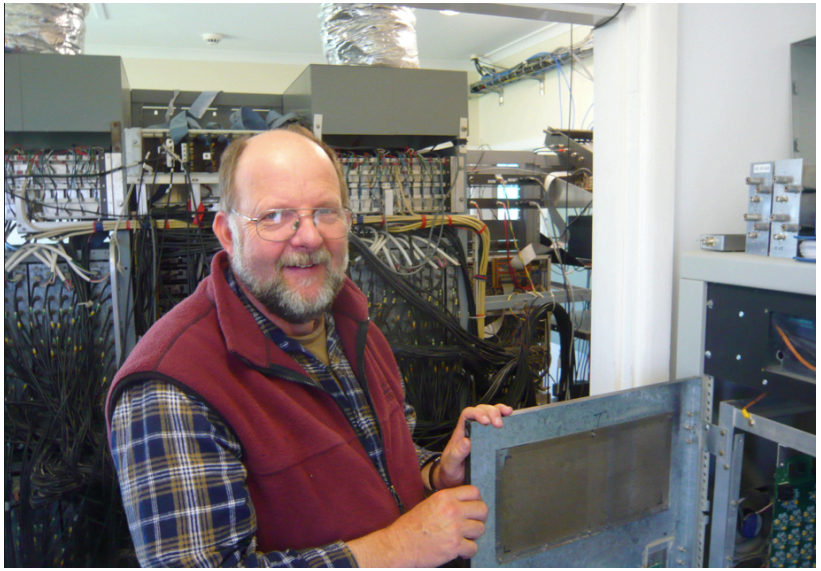
- Hedging of foreign exchange risk
- Spherical Microphone Array
- Rounding Error Analysis
- Online Machine Learning
- **SKAMP**

Molonglo Radio Observatory

Develop and test new technology for the SKA

›





Conclusion

- uPs are the most flexible technology but performance (speed and power) is relatively low
- FPGAs provide
  - Easy interfacing with hardware (tighter coupling than GPUs)
  - Parallelism
  - Have become large enough to implement DSP and ML algorithms
  - Very interesting research area: architectures, tools, applications
- ASICs becoming only be suitable for highest volume, highest performance applications, FPGAs will do the rest
- Many of the highest performance accelerators, particularly for real-time problems, are FPGA-based

THE UNIVERSITY OF SYDNEY

Clustertech

ClusterTech 联科集团（中国）：

We bring you modern computing technologies---

Founded in 2000 to help enterprises in Hong Kong, Greater China and Asia to improve productivity through:

-- **Advanced Software Technologies**: High Performance Computing (HPC), Cluster, Cloud Computing, Grid, GPU, High-speed networks and I/O solutions....

-- **Advanced Computational Technologies**: Business Intelligence, Data Mining and Modeling, Artificial intelligence, Optimization....

THE UNIVERSITY OF SYDNEY

Looking For Talented Employees

联科集团 - ClusterTech

- Cloud Computing
  - Position 1 - Software Engineer (Cloud Computing)
- Cluster Computing
  - Position 2 - HPC Engineer
  - Position 3 - Implementation Engineer
  - Position 4 - FPGA Engineer
- Business Intelligence
  - Position 5 - BI Engineer

* Please refer to appendix for job details

References

- C.K. Cheng, A.B. Kahng, and P.H.W. Leong. Reconfigurable computing. In J.G. Webster, editor, Encyclopedia of Electrical and Electronics Engineering. Wiley, 2007.
- Hauck and DeHon, Reconfigurable Computing: The Theory and Practice of FPGA-Based Computation, Morgan Kauffman, 2007