Author: André Gunst Chris Broekema	Date of issue: 2008-Aug-28 Kind of issue: limited	Scope: Development Doc.id: LOFAR-ASTRON--xxx	
	Status: preliminary Revision nr: 0.6	File:	

Signal-Processing Technologies Assessment

Accepted:			
Role	Name	Date	Signature
		2008-09-02	

Revision	Date	Section	Page(s)	Modification
0.1	2008-09-02	all		Creation
0.2	2008-09-17			Comments Eric Kooistra added
0.3	2008-09-23	1,2, 8		CPU, GPU and exotic hardware rewritten
0.4	2008-12-02			Minor changes
0.5	2008-12-09			Comments Chris Broekema added
0.6	2008-01-15			Comments John Romein added

Table of contents

Introduction

This document describes current trends in hardware technologies that can be used to process data of next-generation radio telescopes. The document discusses the advantages and disadvantages of popular and recently-developed technologies. It covers all signal-processing stages that are required to reduce telescope data, from digitized receiver samples to image.

The processing pipeline starts with simple operations on large amounts of data, but as the data are gradually reduced in size, much more complex algorithms are applied to smaller data volumes. The different processing and I/O requirements of various processing stages justify the use of different technologies. The key parameters that detemine the suitability of a given technology are:

- the processing requirements (in terms of FLOPs or MACs)
- the bandwidth requirements
- the arithmetic complexity, i.e., the ratio between the two aforementioned
- terms
- the required flexibility
- the complexity of an algorithm (e.g., calibration software is more complex
- than a correlator)
- power consumption and cooling constraints
- development time and effort
- production costs
- overall costs

Below, we enumerate the currently available processing technologies.

1 CPU Technology

Systems based on CPUs are easily accessible, since high level languages are available that hide the details and complexity of the chip for the programmer. An expert programmer still has access to these details, with the benefit of analysis and debugging tools available on the chosen hardware and operating system combination. Lots of compilers, libraries and software are available and a programmer can test his/her implementation on common of the shelf hardware.

Moore's law states that the number of transistors which can be placed inexpensively on an integrated circuit will increase exponentially. This results approximately in doubling the density every two years. As a result of the density increase also the clock frequency increases. Due to increasing leakage currents at high clock frequencies leading to excessive power consumption, this trend cannot be maintained. In the next couple of years there will be an increasing trend to on-chip parallelism. Currently, quad-core CPUs are generally available. This trend will continue approximately on par with Moore's law, probably leading to CPUs with hundreds of cores within a decade.

If the trend of multiple cores in a chip continues then the CPU chips will reach part of the territory of the DSP and FPGA chips. One of the differences with FPGAs is the amount of bandwidth which can be fed into the chip and that FPGAs use fixed point precision processing elements. Furthermore, the cores in a CPU include a complete instruction set, while the standard processing blocks in FPGAs are frequently only a Multiply and ACcumulate (MAC) operation.

CPUs consume significantly more power, compared to dedicated hardware. Increasing awareness in the industry has led to green-computing initiatives that aim to reduce power consumption while still maintaining performance. The large market for general purpose CPUs make them relatively cheap.

The amount of IO bandwidth to the chip is primarily used for communication to memory and hence not designed in the first place to accommodate streaming applications.

Programming of general purpose CPUs is generally easier than either DSPs or FPGAs, with manpower easily available. This reduces development time and cost. Additionally CPUs are used in consumer products, which drive the total cost down.

2 GPU Technology

The amount of processing available on graphical cards is significant. This was already acknowledged in 12. At that time the complexity of porting existing algorithms onto graphical accelerators was a tedious process. This has improved in the current generation of programming languages and tools.

The instruction set of GPU chips is much smaller than CPU chips and hence its flexibility to be used for different kind of algorithms is less than for CPU chips. GPUs do however offer significantly more Flops per both Euro and Watt. Interestingly, Intel has acknowledged this and is planning to introduce a hybrid many-core GPU/CPU during the end of 2009 which consists of ~16 X86 compatible, but relatively simple, cores. Such a platform may be eminently suitable for online astronomical applications.

The amount of cores per chip is increasing significantly. The last number reported are 800 cores on one chip running at 750 MHz. Since the amount of cores in CPUs increases as well, the tendency is that both types of technology are converging to one other. One needs to keep in mind however those applications will not scale linearly with the number of cores or CPUs it is mapped onto. Amdahl's law 12 states that the speedup of a program using multiple processors in parallel computing is limited by the time needed for the sequential fraction of the program.

The bandwidth from and to a GPU is often limited by the host machine.

GPUs dissipate a significant amount of power, while the chip itself is relative cheap due to the large game industry. Power consumption per flop is generally better compared to general purpose CPUs.

3 DSP Technology

DSP technology is flexible but not as flexible as CPU technology.

A DSP programmer needs dedicated development software for the DSP processor and is assumed to have detailed knowledge of the underlying chip architecture. Also the IO bandwidth of the DSP chips is much more than is generally offered by the CPU chips. Generally, DSP chips are designed in dedicated boards with their own interfaces and for its own application. So, development costs are higher and manpower may be harder to find.

The DSP market has shrunk significantly due to the growing market of FPGAs.

4 FPGA Technology

FPGA chips are used in streaming processing tasks with a high bandwidth, which require also a large amount of processing power. The strength of a FPGA is that the user can reconfigure it, while the amount of bandwidth and processing power in the chip is huge. Furthermore the cost per chip is also reasonable for a low amount of chips (~ 50k chips) compared with ASICs. The architecture of a FPGA has changed during the past decade from purely logic and Look Up Tables (LUTs) to logic including embedded multipliers and an accumulation functionality. This last trend narrows the gap between FPGAs and ASICs. Also the gap towards CPUs has been narrowed down by accommodating general-purpose CPUs on a FPGA (hard wired or soft wired). Other “ASIC type” building blocks which are integrated on the FPGAs are transceivers including Ethernet MAC and DRAM controllers. This makes the FPGA a true System On Chip (SOC).

An interesting comparison between FPGAs and ASICs is reported in 12 for the 90 nm technology. The area ratio between FPGAs and ASICs was a factor of 40 for logic only and a factor 21 on average logic, memory and DSP functionality. Furthermore the dynamic power dissipation was roughly a factor 10 higher for FPGAs.

These results can be debated because the available ASIC technology lags about three generations behind compared with the FPGA technology 12. Taking that into account the FPGA to ASIC ratio of area is only a factor 4 for logic only and a factor of 2 for logic, memory and DSP functionality. The dynamic power dissipation is for this case a factor 1.5 on average.

Furthermore, the development cycle for a FPGA is longer than the previous technologies discussed. Compared with an ASIC the development time is less.

The FPGA families available vary in processing power, bandwidth, amount of logic and cost.

5 Hardcopy / easypath Technology

The FPGA vendors try to narrow the gap between ASICs and FPGAs even more by offering programs to "freeze" the logic in FPGAs. This is called Hardcopy for Altera devices and in these chips all cells concerning the programming are taken out of the chip. The power consumption in those chips is a factor of two less. Transferring to hardcopy devices can be done quickly and is almost free of risks.

Xilinx calls this easypath and they select rejected chips for applications, where malfunctioning cells are not used. This increases the yield and decreases the price of the chips compared with ASICs.

6 ASSP Technology

ASSP chips are dedicated chips including functionality which is used in many applications/products for standard operations (e.g. transceivers). The disadvantage is that one chip has one certain functionality. Hence, to offer a complete system, multiple chips are necessary surrounding the ASSP chip to make it application specific. The development time is comparable with an ASIC. Since many of those chips are necessary, the NRE costs can be amortized over many chips.

The usage of ASSP technology for most of the key functionalities for next generation radio telescopes is questionable. However for transceivers for example this is an excellent and cheap solution.

7 ASIC Technology

An ASIC chip is normally designed for one single application and hence can be designed for a lot of processing power, high throughput and low power. The main drawback is that the initial development cost of ASIC chips is significant and breaks even at ~ 50k chips. Since an ASIC is designed normally for a single application, there is no flexibility and reconfigurability possible (unless it is designed in). Also the risk involved in ASIC design should be considered, because the NRE can double in worst the case if the design is not right for the first time.

Furthermore the development time of an ASIC is much longer than for the previous discussed technologies. Hence, at the time the system is used the technology is "old" and the same system can be build with much less volume in new technology.

Currently no quick design flow exists wherein a FPGA design can be targeted to an ASIC, although ASIC designs are often (partially) validated on FPGAs.

8 IBM Cell BE Technology

One of the exotic chips discussed are the Cell BE developed jointly by IBM, Sony and Toshiba. These chips are popular these days because of their use in the game console industry. Furthermore, these chips draw the attention of the computing community because of the extreme specifications in terms of processing power and memory bandwidth.

One of the disadvantages of the Cell chips is that the memory bandwidth specified cannot be used directly for streaming data [ref. Shenton from Manchester et. all]. It is unsure what the real-time IO bandwidth will be

for streaming data. Furthermore, the Cell has a limited instruction set compared with a CPU type of system and is not able to run a standard operating system. This adds to the complexity of programming a Cell, and requires a specialized software developer.

9 Supercomputer Technology

Supercomputer processors are optimized for scientific workloads, typically double precision floating point operations, with a high compute to IO ratio. Supercomputers can be divided into three categories:

1. clusters of general purpose PC like computers
    - possibly interconnected using a low latency interconnect
    - possibly equipped with additional processing power in the form of GPU, FPGA or other co-processor boards
2. large, shared-memory systems; these are largely obsolete
3. massively parallel, distributed memory systems

Since the first is already covered and the second is largely obsolete, we will concentrate on the third category. Massively parallel supercomputers are typically designed to their thermal limit. Their CPUs therefore consume less power than contemporary mainstream CPUs. Individually their performance is most likely modest; the large number of processors and associated hardware makes this a supercomputer. Examples of vendors are IBM and Cray.

A supercomputer is, by definition, expensive. It is also, usually, the only way to build an extreme amount of processing power within a budget, be it financial, power or floor-space.

Supercomputer CPUs are halfway between general purpose CPUs and custom designed ASICs or DSPs. Considering the number of CPUs required for SKA, it becomes feasible to design a dedicated CPU, optimized for the expected processing requirements. Such an application specific CPU core would probably combine a 'normal' processor with application optimized or application specific accelerators. Several of these accelerators will fit onto a single core, making it possible to use the same package in different parts of the instrument. Several of these cores, with associated hardware could be combined into a single SoC. Such a CPU would combine the flexibility of a general purpose CPU, with the performance of an FPGA. This does come at a cost, however. Designing and building such a CPU takes time, and the resulting product will lag at least a generation behind contemporary general purpose CPUs.

10 Summary

In Table 1 a summary of the advantages and disadvantages of the discussed technologies are given on each parameter defined.

	Amount of processing	IO bandwidth	Flexibility	Power consumption	Volume	Cost	Development time
CPU	+-	--	++	--	--	++	++
FPU	+-	--	+-	--	--	++	++
DSP	+-	+-	+-	--	--	+-	+-
FPGA	++	++	-	+-	+-	-	-
Hardcopy/Easypath	++	++	--	+	+	-	-
ASSP	++	++	--	++	++	-	--
ASIC	++	++	--	++	++	--	--
CELL	+	+-	+	-	-	--	+-
Supercomputer	+	+-	+	-	-	--	+

Table 1 Summary of chip technologies parameters

11 Glossary

ASIC	Application Specific Integrated Circuit
ASSP	Application Specific Standard Product
COTS	Common Of The Shelve
CPU	Central Processing Unit
DSP	Digital Signal Processor
FLOP	FLOating point Operation
FPGA	Field Programmable Gate Array
FPU	Floating Point Unit
LUT	Look Up Table
OS	Operating System
IO	Input/Output
MAC	Multiply and Accumulate
SOC	System On Chip
SIMD	Single Instruction, Multiple Data
MISD	Multiple Instruction, Single Data
NRE	Non-recurring expenditure

12 References

[1] COTS Correlator Platform, K. v.d. Schaaf, R. Overeem, S. Briggs, The Square Kilometre Array: An Engineering Perspective, editor Peter J. Hall, Springer Netherlands, January 2006.

[2] Gene Amdahl, Validity of the Single Processor Approach to Achieving Large-Scale Computing Capabilities, AFIPS Conference Proceedings, (30), pp. 483-485, 1967.

[3] I. Kuon and J. Rose, Measuring the Gap Between FPGAs and ASICs, IEEE Transactions on Computer-Aided Design of Integrated Circuits, Vol. 26, No. 2, February 2007.

[4] R. Valvekens, System-on-Chip Design using Modern FPGA, Bits & Chips Hardware Conference 2008, Eindhoven.