D2.1.2 – Architectural developments towards exascale

WP2: underpinning and cross-cutting technologies

Project Acronym	CRESTA
Project Title	Collaborative Research Into Exascale Systemware, Tools and Applications
Project Number	287703
Instrument	Collaborative project
Thematic Priority	ICT-2011.9.13 Exascale computing, software and simulation

Due date:	M24
Submission date:	30/09/2013
Project start date:	01/10/2011
Project duration:	36 months
Deliverable lead organization	UEDIN
Version:	1.0
Status	Final
Author(s):	Jeremy Nowell (UEDIN), Michèle Weiland (UEDIN), Alistair Hart (CRAY UK) Jan Westerholm (ABO), Artur Signell (ABO), Berk Hess (KTH), Erik Lindahl (KTH), George Mozdzynski (ECMWF), Stefano Markidis (KTH)
Reviewer(s)	Erik Lindahl (KTH), Jan Westerholm (ABO)

Dissemination level	
<PU/PP/RE/CO>	*PU – Public*

Copyright © CRESTA Consortium Partners 2011

Version History

Version	Date	Comments, Changes, Status	Authors, contributors, reviewers
0.1	29/08/2013	First draft for WP2 and WP6 review	Jeremy Nowell (UEDIN)
0.2	31/08/2013	Version for internal review	Jeremy Nowell (UEDIN)
1.0	20/09/2013	Final version after internal review	Jeremy Nowell (UEDIN)

Copyright © CRESTA Consortium Partners 2011

Table of Contents

Copyright © CRESTA Consortium Partners 2011

Index of Figures

Index of Tables

Copyright © CRESTA Consortium Partners 2011

1 Executive Summary

High Performance Computing (HPC) is a growing market. It is becoming to be seen as vital for a nations scientific and industrial competitiveness; more countries are providing funding for research into HPC, for instance China which has seen a significant growth over the last few years such that the fastest machine in the world is Chinese. The quest to make a supercomputer with Exascale performance requires significant technological advances, particularly given the limited power budget that such a machine will have. Such advances may come from the wider computing market, where the enormous growth in mobile computing is driving research into power efficient technology, or from research funding specifically for HPC.

To understand what an Exascale machine may look like it is informative to look at trends in relevant technology. Underlying trends in both semiconductor and communication technology drive advances across the computing landscape. These lead to advances in system building blocks; processors, memory, interconnect and software. By looking at company roadmaps some trends become clear, firstly the growth in heterogeneous systems involving different types of processor such as a traditional general purpose CPU and GPU. Secondly, the trend towards integration of components in System-on-Chip (SoC) silicon. Thirdly, the growth in licensing intellectual property, such a processor designs, to other manufacturers.

Several factors are important when considering HPC system architecture trends. These include performance, programmability and usability, power usage and efficiency, cost of procurement and cost of ownership. The TOP500 list provides 20 years worth of data to analyse to look at architecture trends. There has been a move towards using commodity components over custom technology, however this has seen raw floating point performance emphasised at the cost of improvements in memory, interconnect and I/O. An example in architecture trends is provided by looking at the development of the Cray XC30 system.

An Exascale machine is only useful if it has applications capable of using it. The CRESTA co-design applications provide an excellent source of information of the impact of architecture trends on application performance and design. Heterogeneous systems are seen as inevitable; however they have to be easier for application developers to exploit. This will be achieved by providing better integration, particularly through a single addressable memory space, and more importantly through the provision of standard, well supported programming models and languages. Highly parallel systems with millions of processors will need a matching high performance interconnect to allow the system to be fully exploited by applications. Although the wider market may provide advances in power efficient processor technology, funding for HPC specific research into interconnect, programming models and application development will be required.

2 Introduction

2.1 Purpose

The purpose of this document is to provide an update of the previous CRESTA deliverable on the same topic, D2.1.1 [1]. The principle of co-design of both Exascale hardware and scientific applications lies at the heart of CRESTA; therefore this document aims to link trends in HPC architectures with the potential impact on the applications.

Key to the development of future Exascale machines are both various underlying development trends, and the HPC marketplace. We therefore give a brief overview of the HPC market as it stands in Section 3. The first of these trends is the development of basic underlying technologies, used throughout the computer industry. Such trends were described in D2.1.1 [1], so here we will only give a short overview in Section 4.1. Next are market trends, both within the entire computing industry, and also those specific to the HPC segment of the market. These are described in Section 4.2 in conjunction with the technologies. Taking a wider view, what might an Exascale machine of the future look like from an architectural viewpoint? Predicting this is very hard, having to take into account the development trends already mentioned, along with HPC specific developments. This is done in Section 5. As a case study we look at the development of the latest Cray machine, the XC30, in Section 6.

Applications are key to achieving Exascale performance on future machines, therefore it is important to try to gauge if and how they might need to adapt to possible architectures. Equally important is to present the requirements of the co-design applications on future machine architectures, so that these might be taken into account in future designs. The impact on some of the CRESTA applications is described in Section 7.

Finally, in Section 8 we draw some conclusions on how we believe HPC system architectures are developing towards the Exascale.

2.2 Glossary of Acronyms

API	Application Programming Interface
CAPI	Coherence Attach Processor Interface
CMOS	Complementary Metal Oxide Semiconductor
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
DAG	Direct Acyclic Graph
DSP	Digital Signal Processor
ECC	Error correcting code (memory)
ECMWF	European Centre for Medium Range Weather Forecasting
EUV	Extreme Ultraviolet
FET	Field Effect Transistor
FPGA	Field-programmable Gate Array
GPU	Graphics Processing Unit
GPGPU	General Purpose Graphics Processing Unit
HPC	High Performance Computing
HSA	Heterogeneous Systems Architecture
hUMA	Heterogeneous Uniform Memory Access
IDC	International Data Corporation
IEEE	Institute of Electrical and Electronics Engineers
IFS	Integrated Forecasting System
I/O	Input/Output
MPI	Message Passing Interface
OpenCL	Open Computing Language
OpenGL	Open Graphics Library
OpenMP	Open Multi-Processing

RDMA	Remote Direct Memory Access
SIMD	Single Instruction Multiple Data
SoC	System on a Chip
TBB	Threading Building Blocks
TSMC	Taiwan Semiconductor Manufacturing Company

3 The High Performance Computing Market

The term “High Performance Computing”, or HPC, is becoming more widely known throughout the technology industry and beyond. More companies are recognising the value of HPC to their business, particularly as part of the product design process. More scientists are becoming familiar with it, whether that be for simulating scientific processes or for data analysis. In short, HPC is starting to be seen as vital for industrial and scientific competitiveness. The consequence of this trend is that the HPC market, as referred to by system manufacturers and market analysts, is becoming deeper, covering a range of systems from the traditional national level supercomputers through to departmental level servers.

Before considering the HPC market in particular, it is worth considering where it sits within the technology industry as a whole. The breakdown used by most analysts is to use three major categories; servers, personal computers and mobile devices. HPC systems then fall into the server category. Some analysts then break down the HPC category further, for example by price. IDC refers to supercomputers as those costing more than $500,000, divisional as $250,000 to $499,000, departmental as $100,000 to $249,000 and workgroup servers those below $100,000. Table 1 shows HPC sales by units and revenue for the last year, as calculated by IDC.

Table 1: HPC Sales by units and revenue. Figures from IDC in [2].

Type	Units Sold (2011)	Units Sold (2012)	Growth (%)	Revenue (2011) (Million USD)	Revenue (2012) (Million USD)	Growth (%)
Supercomputers	2,908	2,397	-17.6	4,370	5,650	29.3
Departmental	3,724	3,650	-2.0	1,237	1,210	-2.2
Divisional	20,625	17,108	-17.1	3,467	2,997	-13.6
Workgroup	84,294	80,692	-4.3	1,226	1,241	1.2
Total HPC	111,551	103,847	-6.9	10,300	11,098	7.7

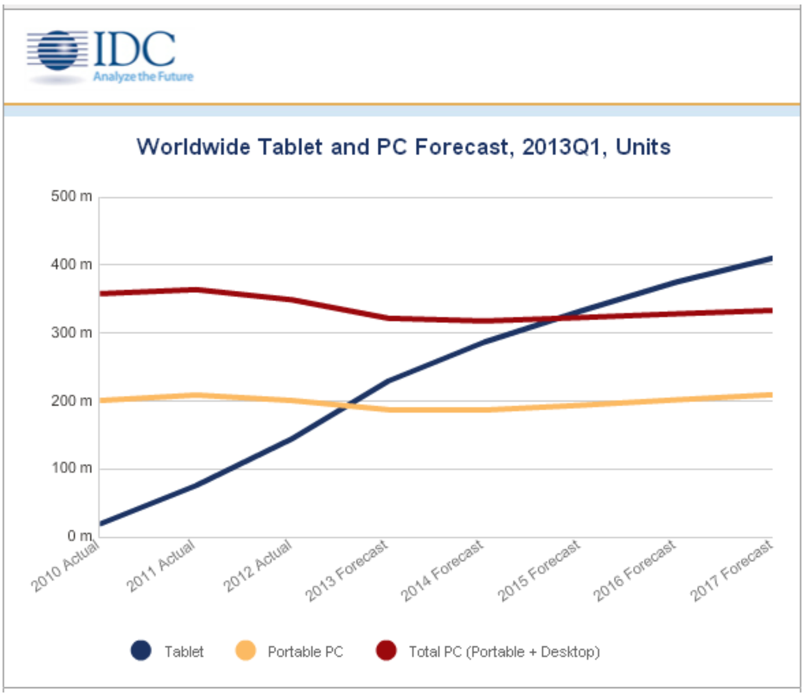


Figure 1: Worldwide Tablet and PC Forecast. Taken from [3].

When looking at the wider technology marketplace it is clear that mobile technology is on a large uptrend; sales of tablets are growing at a rate which will see them overtake personal computers in a few years' time, as shown in Figure 1. IDC expects tablet sales of 229.3 million in 2013 [3]. Meanwhile smartphones are a huge market as well, with that market growing further as they are taken up in emerging economies. When compared to the size of the HPC market in terms of numbers of devices it is clear that the mobile technology market is huge, driving research into low-powered technology. This transformation to a mobile world is driving research forward as consumers are demanding ever more powerful technology with richer, more interactive interfaces, on lighter, more compact devices. Will the drive for Exascale machines, with the widely publicised power budget of 20 MW, be able to capitalise on this mobile technology?

On the other hand, it is also clear from Table 1 that the large, supercomputer segment of the market is providing strong revenue growth. IDC estimates that the HPC market overall will continue to grow at a rate of 6.8%, revenues reaching $15.4 billion by 2017. As HPC take-up becomes more widespread does the scale of these large machines, with millions of processors, mean that the market will be able to sustain either independent or additional research? One answer may be provided by looking at the growth of accelerators in HPC. When GPUs began being programmed by scientific researchers it was using graphics languages such as OpenGL, However Nvidia clearly recognised several years ago that there was enough of a market to justify the development of CUDA to ease programmability, and in fact now sells a range of GPGPUs specifically for technical computing.

It should also be remembered that the largest machines in the world, those at the very top of the TOP500 list [4], are rarely developed by vendors solely to satisfy the gap in the market. They are usually supported by national scientific or defence agencies who wish to gain access to ever larger machines to support their research. This support is then able to be capitalised upon by the vendors who are able to use any new technology developed for these machines in product lines which they can sell in the wider HPC and non-HPC markets. This explains why support is usually provided within national boundaries – governments are willing to fund research if it will provide a boost to their own economies.

Along with national competitiveness it is also worth remembering that the largest systems, especially the system at the top of the TOP500 list, carry with them some amount of national pride. This will certainly be the case for the first Exascale machine, and therefore provides a motivation for some governments to invest in research in the HPC area.

We should therefore look at the international spread of supercomputing today. Such data may be obtained from the TOP500 list. The latest list, from June 2013, may be analysed using the tools available on the TOP500 website to obtain the plots shown in Figure 2, Figure 3 and Figure 4. It is clear from these figures that HPC use is widely spread throughout the world. Although Japan has long been a competitor on the global scene, the recent growth of the market in China has clearly been behind the increase in the share of the market in East Asia. In fact IDC has shown that the Chinese HPC market grew steadily over the last few years, in comparison to others which faltered somewhat during the economic turmoil of the period. This is illustrated by Figure 2, which shows that although North America is still the dominant region, the share of performance in Eastern Asia has grown over recent years. Figure 4 shows that China currently has the second largest share of performance in the TOP500, due largely to the number one machine, Tianhe-2, and the number ten machine, Tianhe-1A. IDC confirms that the Chinese supercomputer segment has grown most heavily of the last few years, and has not been affected at all by the recent recession [2].

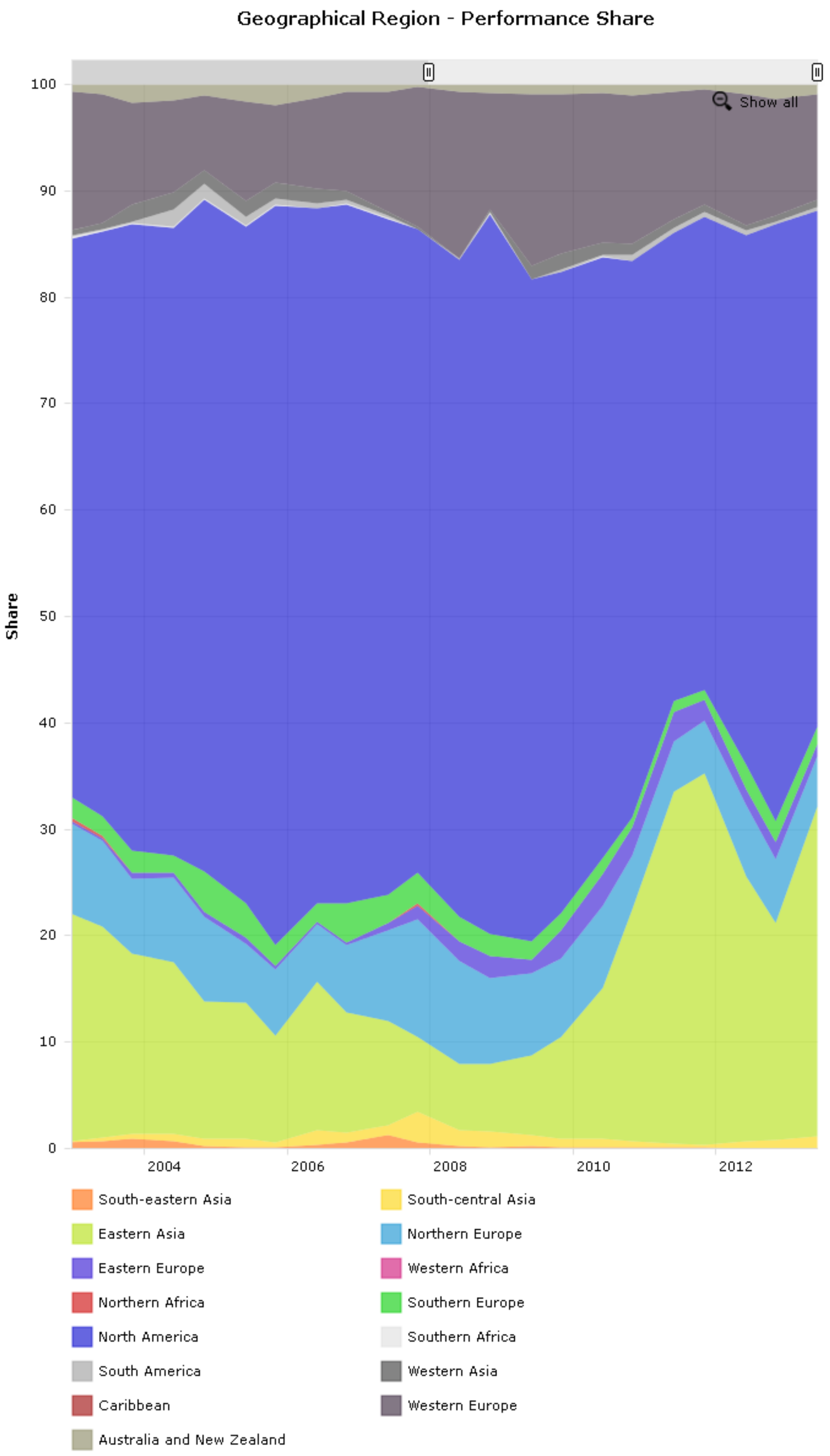
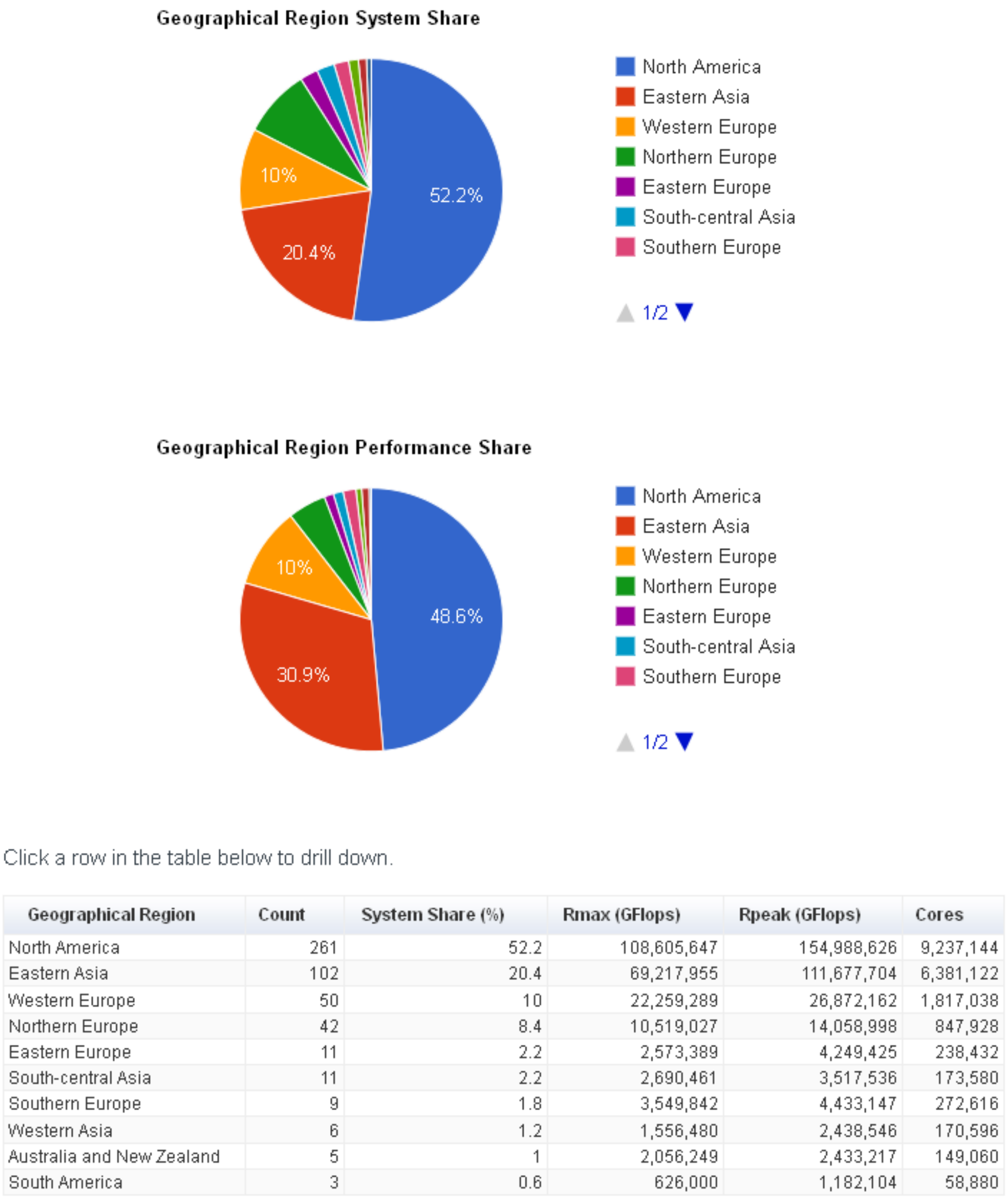


Figure 2: Evolution of the geographical share of the performance in the TOP500 list over the last ten years.

© CRESTA Consortium Partners 2011



Click a row in the table below to drill down.

Geographical Region	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
North America	261	52.2	108,605,647	154,988,626	9,237,144
Eastern Asia	102	20.4	69,217,955	111,677,704	6,381,122
Western Europe	50	10	22,259,289	26,872,162	1,817,038
Northern Europe	42	8.4	10,519,027	14,058,998	847,928
Eastern Europe	11	2.2	2,573,389	4,249,425	238,432
South-central Asia	11	2.2	2,690,461	3,517,536	173,580
Southern Europe	9	1.8	3,549,842	4,433,147	272,616
Western Asia	6	1.2	1,556,480	2,438,546	170,596
Australia and New Zealand	5	1	2,056,249	2,433,217	149,060
South America	3	0.6	626,000	1,182,104	58,880

Figure 3: Geographical share of the June 2013 TOP500 list.

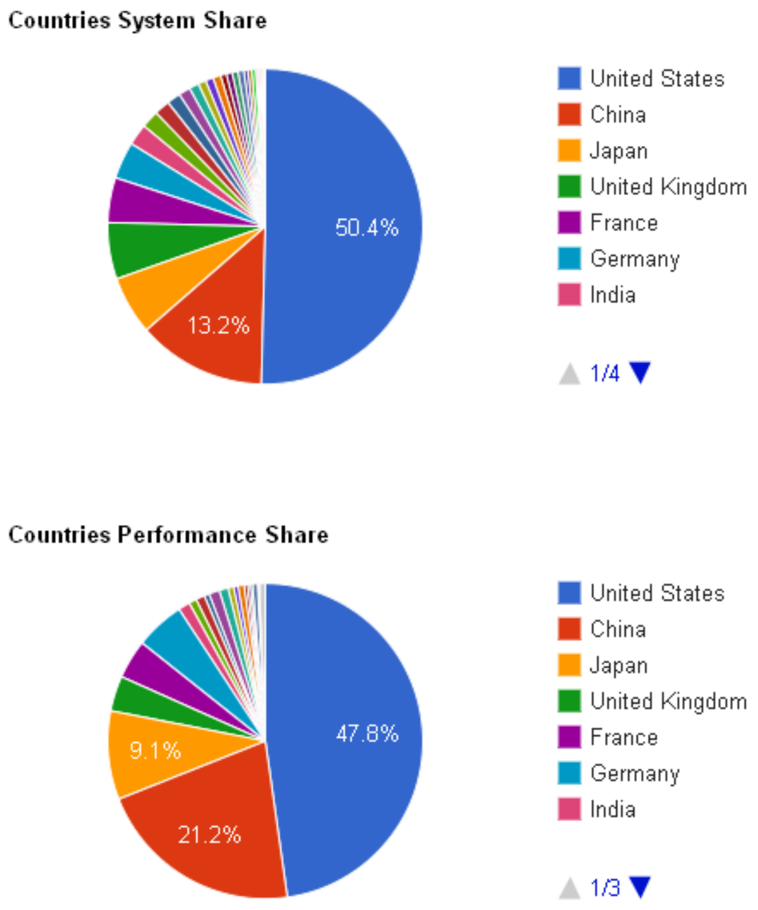


Figure 4: Share of the June 2013 TOP500 list by country.

We can say that progress towards Exascale systems may therefore come from three areas. Firstly, the explosion in mobile technology driving progress in low power technology. This is important given the need for Exascale systems to be provided within a limited power budget. Secondly, the growth in the HPC market, particularly amongst the largest systems. Thirdly, government funded research, whether that be for competitive reasons or for national pride. Progress in technology means it is inevitable that the Exascale will be reached one day, however, external funding sources will necessarily determine the timescale.

4 Current Technology Trends

In this section we look technology trends, both underlying technical trends and market trends – sometimes it is difficult to disentangle the two. We consider basic technologies and then system building blocks.

4.1 Underlying Basic Technologies

4.1.1 Semiconductor Technology

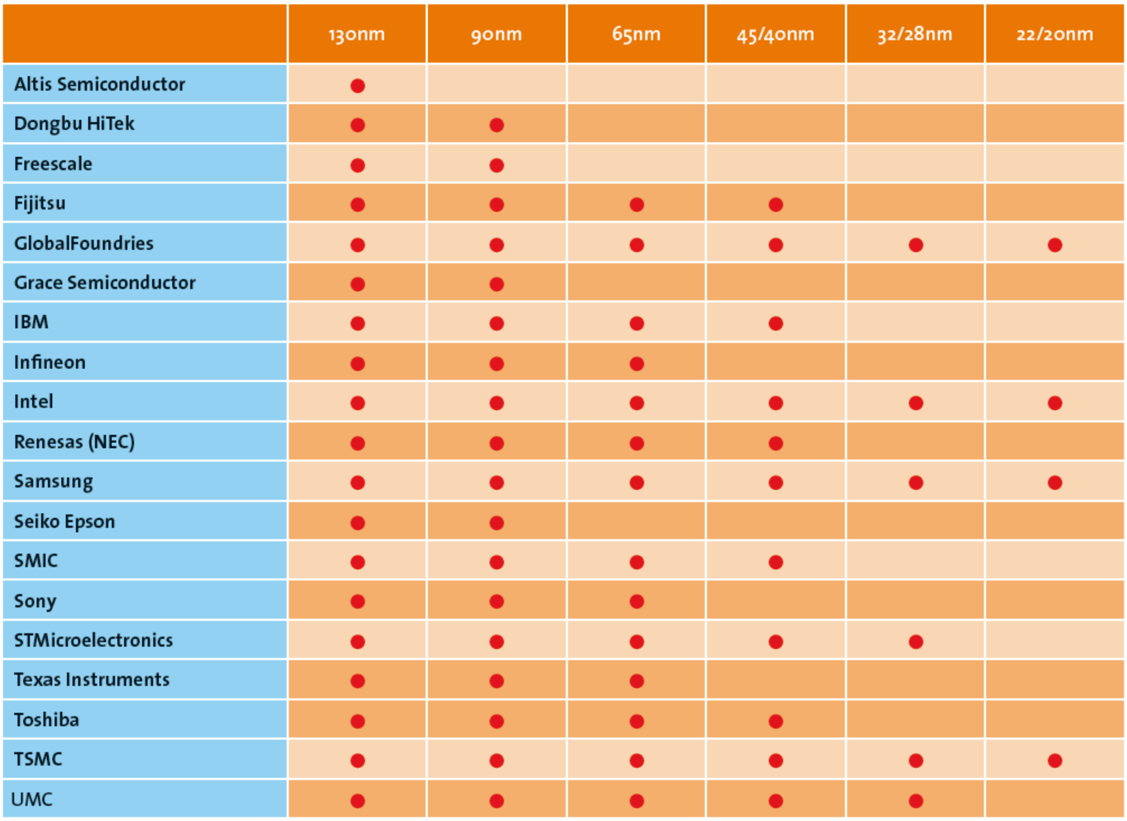
Historically developments in semiconductors have been described in terms of Moore's Law [5], which is commonly formulated as an observation that the number of transistors on an integrated circuit doubles approximately every two years. This has held true for almost fifty years, driven largely by advances in CMOS manufacturing, in particularly photo-lithography techniques which have continually reduced the feature size of integrated circuits. Recently, more advanced processes have been required, such as FinFET or tri-gate transistors, as used by Intel in their "Ivy-Bridge" and "Haswell" processor ranges. These are three-dimensional processes used at scales of 22 nm and below. Future improvements to processes, for example Extreme ultraviolet lithography (EUV), will be required to reduce feature sizes further.

Taiwan Semiconductor Manufacturing Company Limited (TSMC) is one of the major semiconductor manufacturers, counting Nvidia and AMD amongst its clients. Recent presentations of its roadmap [6] have 10 nm manufacturing nodes coming online in 2016. Other innovations presented include 3D chip stacking and the introduction of germanium to replace silicon in order to reduce the feature size even more. Other manufacturers are researching techniques such as carbon nanotubes and other types of semiconductor. Intel have presented their roadmap with aims to go to 5 nm [7].

All indications, therefore, are that Moore's Law will continue to be followed by the chip manufacturers, at least on the timescale of a future Exascale machine around 2018, and probably beyond.

It is more interesting to look at the capabilities of the main semiconductor manufacturers, as shown in Figure 5. This clearly illustrates that a very small number of manufacturers dominate the landscape at small feature-sizes. The main reason for this, as reported in reference [8], is the growing cost of manufacturing as the size decreases and the complexity increases. It becomes harder for companies to design and have manufactured small runs of specialist components as the cost increases. A figure quoted in [9] states that a manufacturer producing 20 nm SoC units with a $20 average selling price needs to ship over 9 million units to break even. Although this may sound like a large figure it is worth remembering that future Exascale machines will likely contain millions of compute cores, therefore it does not seem infeasible that a manufactuer may produce at least a variant of an existing SoC design targetted explicitly at HPC.

The impact for future Exascale machines is that HPC manufacturers are likely to have to rely largely on commodity components designed for major markets, or else spend more on specialist components. An exception to this may be if FPGAs become more widely used, although so far these have proved hard to program, and therefore their takeup has been limited in HPC.



	130nm	90nm	65nm	45/40nm	32/28nm	22/20nm
Altis Semiconductor	●					
Dongbu HiTek	●	●				
Freescale	●	●				
Fijitsu	●	●	●	●		
GlobalFoundries	●	●	●	●	●	●
Grace Semiconductor	●	●				
IBM	●	●	●	●		
Infineon	●	●	●			
Intel	●	●	●	●	●	●
Renesas (NEC)	●	●	●	●		
Samsung	●	●	●	●	●	●
Seiko Epson	●	●				
SMIC	●	●	●	●		
Sony	●	●	●			
STMicroelectronics	●	●	●	●	●	
Texas Instruments	●	●	●			
Toshiba	●	●	●	●		
TSMC	●	●	●	●	●	●
UMC	●	●	●	●	●	

Production capabilities in logic CMOS technology for main semiconductor manufacturers (2011) from HIS iSuppli
Source: IHS iSuppli

Figure 5: CMOS production capabilities. From [8].

4.1.2 Communication

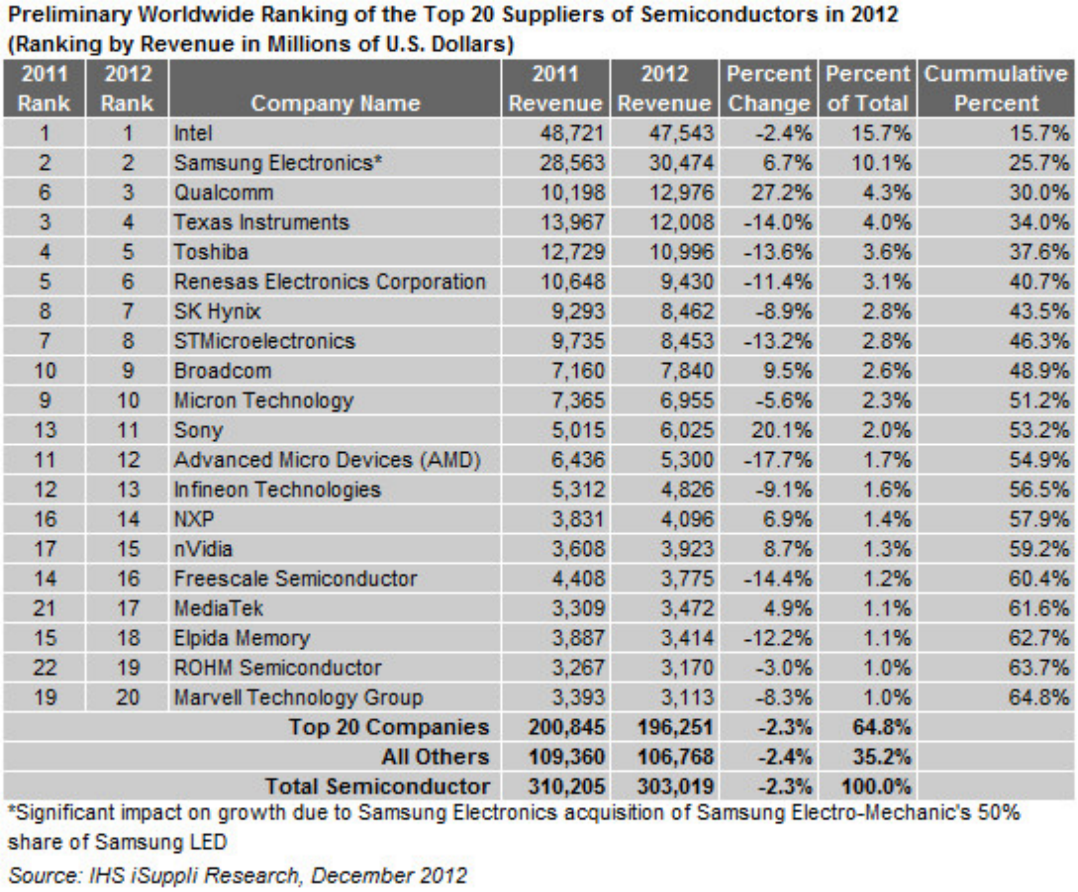
Communication is critical to a supercomputer; the high speed, sometimes custom, interconnect is after all the feature which distinguishes a supercomputer from a collection of servers, allowing processors to communicate with each other to co-operate on a computation. However, even on a more basic level communications are critical to any computer. Data needs to be moved between memory and processor, and even within a processor, before it is communicated to the outside world. However, it now costs much more energy to move data in and out of a processor than it does to perform the computation. Therefore reducing this cost becomes critical to the power efficiency of a device, whether that be a mobile device or a supercomputer.

Possible solutions to this problem come in two forms. Firstly, by lowering the distance between processor and data; and secondly, by reducing the cost of the data movement. The first may be achieved by stacking memory directly on top of processors, making a true SoC. This presents a technical challenge to the semiconductor manufacturers, but is under study. The second may be achieved by moving more towards optical communications rather than electrical ones, in the form of silicon photonics. This may be in the form of connecting system boards, or even chips on an individual circuit board. This is discussed further in D2.1.1.

4.2 System Building Blocks

4.2.1 Processors

Current processor trends are determined by both technical constraints and market demands. On the one hand, the end of Dennard scaling and the subsequent slow down in the growth of clock frequencies has led to the development of multi-core processors, such that they are now commonplace. On the other hand, the demand for low power usage in both mobile devices and the server market has driven progress in power efficient processor design. It is perhaps informative to look at recent developments and future plans from some of the main processor manufacturers.



Preliminary Worldwide Ranking of the Top 20 Suppliers of Semiconductors in 2012
(Ranking by Revenue in Millions of U.S. Dollars)

2011 Rank	2012 Rank	Company Name	2011 Revenue	2012 Revenue	Percent Change	Percent of Total	Cummulative Percent
1	1	Intel	48,721	47,543	-2.4%	15.7%	15.7%
2	2	Samsung Electronics*	28,563	30,474	6.7%	10.1%	25.7%
6	3	Qualcomm	10,198	12,976	27.2%	4.3%	30.0%
3	4	Texas Instruments	13,967	12,008	-14.0%	4.0%	34.0%
4	5	Toshiba	12,729	10,996	-13.6%	3.6%	37.6%
5	6	Renesas Electronics Corporation	10,648	9,430	-11.4%	3.1%	40.7%
8	7	SK Hynix	9,293	8,462	-8.9%	2.8%	43.5%
7	8	STMicroelectronics	9,735	8,453	-13.2%	2.8%	46.3%
10	9	Broadcom	7,160	7,840	9.5%	2.6%	48.9%
9	10	Micron Technology	7,365	6,955	-5.6%	2.3%	51.2%
13	11	Sony	5,015	6,025	20.1%	2.0%	53.2%
11	12	Advanced Micro Devices (AMD)	6,436	5,300	-17.7%	1.7%	54.9%
12	13	Infineon Technologies	5,312	4,826	-9.1%	1.6%	56.5%
16	14	NXP	3,831	4,096	6.9%	1.4%	57.9%
17	15	nVidia	3,608	3,923	8.7%	1.3%	59.2%
14	16	Freescale Semiconductor	4,408	3,775	-14.4%	1.2%	60.4%
21	17	MediaTek	3,309	3,472	4.9%	1.1%	61.6%
15	18	Elpida Memory	3,887	3,414	-12.2%	1.1%	62.7%
22	19	ROHM Semiconductor	3,267	3,170	-3.0%	1.0%	63.7%
19	20	Marvell Technology Group	3,393	3,113	-8.3%	1.0%	64.8%
		Top 20 Companies	**200,845**	**196,251**	**-2.3%**	**64.8%**	
		All Others	**109,360**	**106,768**	**-2.4%**	**35.2%**	
		Total Semiconductor	**310,205**	**303,019**	**-2.3%**	**100.0%**	

*Significant impact on growth due to Samsung Electronics acquisition of Samsung Electro-Mechanic's 50% share of Samsung LED

Source: IHS iSuppli Research, December 2012

Figure 6: Market share of semiconductor manufacturers. Taken from [10].

Figure 6 shows the market share of different semiconductor manufacturers. It is clear from this that Intel dominates the market. Many of the other manufacturers are not involved in the HPC market at all, being either in the mobile or embedded space. Also comparison with Figure 5 shows clearly that several semiconductor suppliers do not actually manufacture their own devices, relying on the foundries to do so.

4.2.1.1 Intel

Intel is the largest supplier of semiconductors in the world by market share. It has dominated the processor market for many years with its x86 based microarchitecture. This forms the basis of different product ranges covering mobile, portable, desktop and server markets. Its development now tends to follow the well-known Intel "Tick-Tock" model [11], in which an improvement (a die shrink) in the manufacturing process technology of a "tick" is introduced to a current microarchitecture design, followed some months later by the introduction of a new microarchitecture using this improved process in a "tock". Typically every year there is either a "tick" or a "tock". This is shown in Figure 7, which illustrates the process over recent years. The current microarchitecture is the recently released "Haswell", available on a 22 nm die. Next year's "tick" should make this available on a 14 nm process.

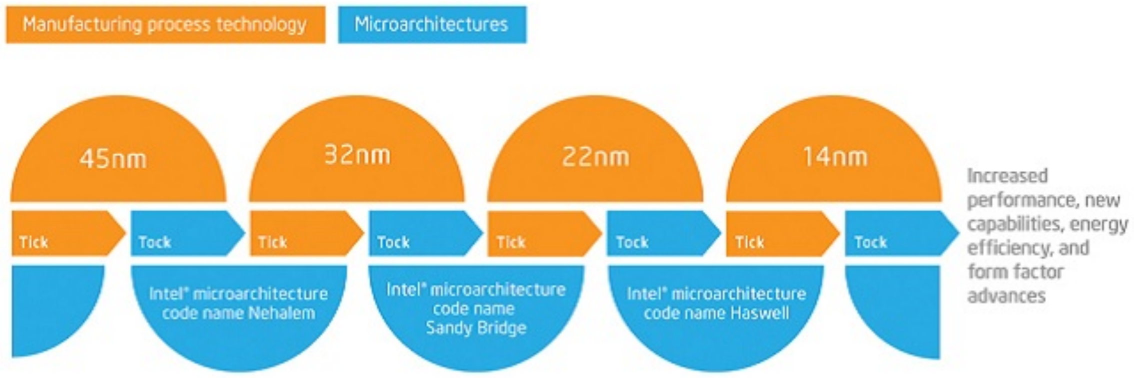

Manufacturing process technology
Microarchitectures
45nm
32nm
22nm
14nm
Tick
Tock
Tick
Tock
Tick
Tock
Tick
Tock
Intel® microarchitecture code name Nehalem
Intel® microarchitecture code name Sandy Bridge
Intel® microarchitecture code name Haswell
Increased performance, new capabilities, energy efficiency, and form factor advances

Figure 7: The Intel tick-tock model, taken from [11].

The current Intel product ranges include Xeon processors for the server/workstation market, and Core processors for the consumer desktop/laptop market, both based on evolutions of the Core microarchitecture. These are either based on the "Ivy Bridge" microarchitecture which was produced as the result of the shrink of the "Sandy Bridge" architecture to a 22 nm process using FinFET transistors, or the new "Haswell" architecture. Processors within the ranges differ in both their operating frequency, and in the features on offer, such as the number of cores they contain, whether they support Hyperthreading, the amount of cache memory, support for Turbo Boost and their power consumption. For a detailed overview of Haswell see [12].

Some desktop models contain on-die graphics processors in the form of Intel HD or Iris Graphics. Unlike standalone GPUs, use of these graphics cores for technical computing seems to have been limited, however Intel are now promoting their programmability using OpenCL [13], in order to increase their usage in this market. Presently the performance of the integrated graphics also seems to lag that of dedicated GPU cards.

Intel also has the Atom product line of low-power processors aimed at the portable and mobile computing market. Uptake of these seems to be limited compared to ARM processors, however there are recent indications that development of these is being accelerated, with a move to a 22 nm process due later in 2013, but then a rapid move to 14 nm in 2014 [14],[15]. Although initially designed for the mobile market it is also likely that the Atom ranges will see use in low-powered dense servers.

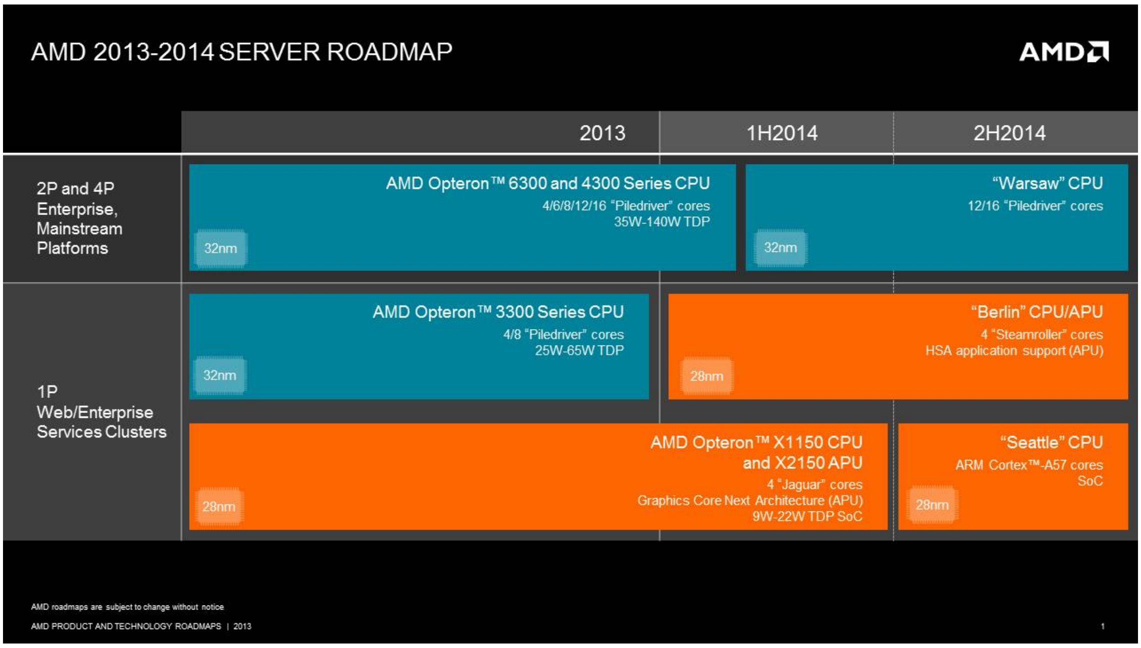
A recent Intel event entitled "Reimagine the Datacenter" [16][18] gave some interesting insights into Intel's future directions. As well as continuing with the low-powered Atom range Intel are also focussed on lowering the power consumption of their Xeon server range. The 14nm shrink of the "Haswell" architecture, known as "Broadwell", will be made available as a System on a Chip (SoC), incorporating I/O and network controllers and accelerators on the same die [19]. The Xeon E3 Haswell is available with a consumption as low as 13 W, therefore it is expected that the Broadwell equivalent will be even lower powered. This can be compared with recent processors, which have a power consumption several times this figure.

Although available only in prototype for some time, a fairly recent product launch from Intel is the Xeon Phi co-processor. This consists of up to 61 x86 cores on a single die, with a high-speed interconnect between them. The cores are derived from an old P5 architecture, but have been augmented with 512-bit vector units amongst other additions. It is able to act as an off-load engine for highly parallel computational tasks in tandem with a traditional Xeon processor, or alternatively to run an executable program itself on a reasonably fully featured version of the Linux operation system. It is utilised in the fastest machine on the TOP500 list, Tianhe-2.

4.2.1.2 AMD

AMD are, for all practical purposes, the only competitor to Intel in the traditional x86 processor market. They also produce the Radeon series of GPUs for the desktop and workstation PC market. The market dominance of Intel has often caused AMD to try and differentiate their products in particular markets, rather than compete directly. For instance, the current generation of processors for the server market comprise x86 compatible integer units, which are paired together and share a floating point unit. It has also placed a lot of focus on its accelerated processing units (APUs). These integrate traditional x86 cores with graphics cores. AMD have many years of experience in the GPU area, having acquired the graphics company ATI, and the graphics performance of the AMD processors is currently recognised to be better than that of the Intel equivalent, although still lagging in performance compared with standalone GPUs. Given AMD's expertise in this area the performance gap may well narrow rapidly in the coming years.

AMD recently presented an update to its server strategy and roadmap [21]. This highlights several interesting developments, with AMD targeting total cost of ownership of servers through low power consumption. The roadmap is shown in Figure 8.



AMD 2013-2014 SERVER ROADMAP

2013 | 1H2014 | 2H2014

2P and 4P Enterprise, Mainstream Platforms

AMD Opteron™ 6300 and 4300 Series CPU
4/6/8/12/16 "Piledriver" cores
35W-140W TDP
32nm

"Warsaw" CPU
12/16 "Piledriver" cores
32nm

1P Web/Enterprise Services Clusters

AMD Opteron™ 3300 Series CPU
4/8 "Piledriver" cores
25W-65W TDP
32nm

"Berlin" CPU/APU
4 "Steamroller" cores
HSA application support (APU)
28nm

AMD Opteron™ X1150 CPU and X2150 APU
4 "Jaguar" cores
Graphics Core Next Architecture (APU)
9W-22W TDP SoC
28nm

"Seattle" CPU
ARM Cortex™-A57 cores
SoC
28nm

AMD roadmaps are subject to change without notice

AMD PRODUCT AND TECHNOLOGY ROADMAPS | 2013

Figure 8: AMD Server Roadmap. Taken from [21].

The first chip of interest is the "Berlin" APU. It is the first server processor based on AMD's Heterogeneous System Architecture (HSA), bringing together four standard x86 compatible "Steamroller" cores with a GPU. HSA is interesting to the HPC market as it simplifies programmability of the GPU. AMD is calling the memory architecture heterogeneous Uniform Memory Access (hUMA), as shown in Figure 9 and Figure 10. The CPU and GPU are now able to address the same memory space, removing the need to communicate data between CPU and GPU. Furthermore, HSA provides the same cache-coherency and virtual memory mapping between CPUs and GPUs. As well as easing programmability this also allows for better GPU program management and isolation – up to now there has been no memory protection on GPUs allowing different executables to access each other's data. See reference [22] for more details.

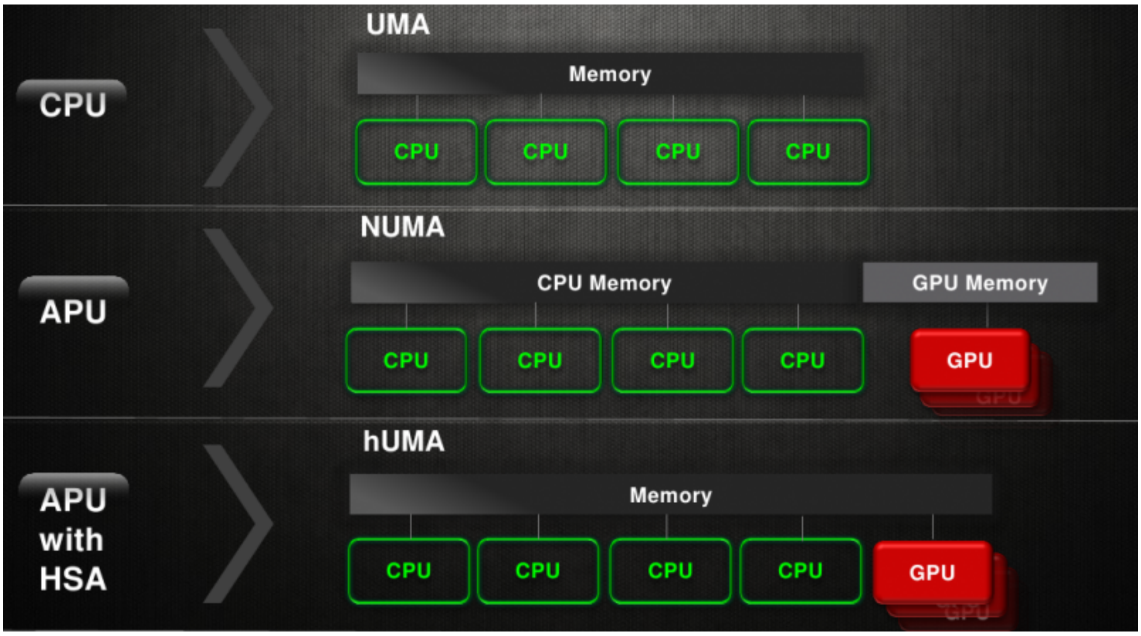
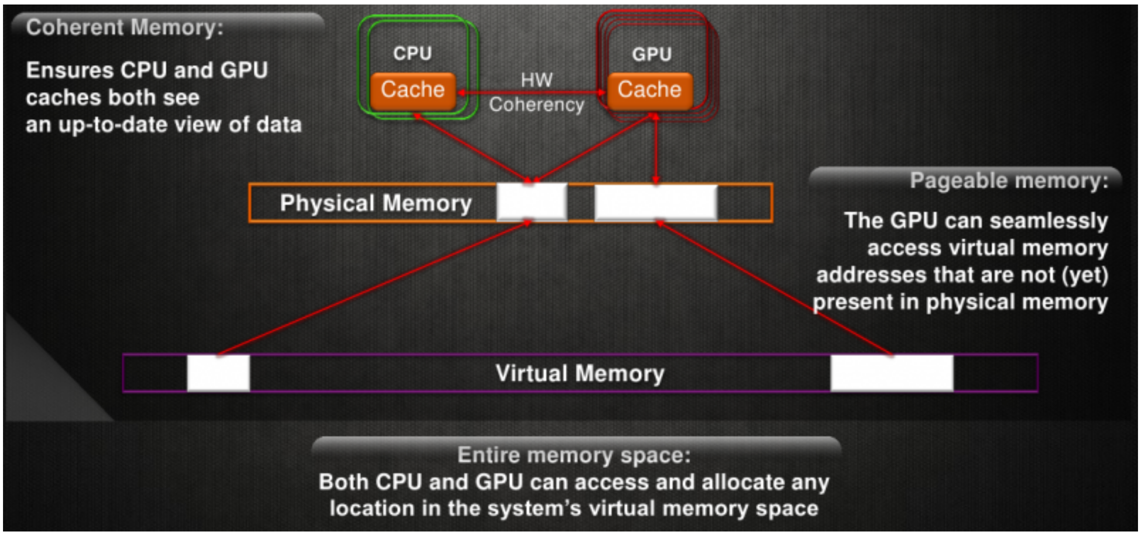


Figure 9: The evolution of AMD's memory architecture. Taken from [22].



Coherent Memory:
Ensures CPU and GPU caches both see an up-to-date view of data

Pageable memory:
The GPU can seamlessly access virtual memory addresses that are not (yet) present in physical memory

Entire memory space:
Both CPU and GPU can access and allocate any location in the system's virtual memory space

Figure 10: Memory layout on AMD's HSA processors. Taken from [22].

Also of interest on AMD's roadmap is the "Seattle". This will be a low powered server SoC, replacing the Opteron X-Series. However, instead of using AMD's own compute core it will use 64 bit ARM Cortex-A57 cores, integrating these with networking components [21]. See reference [23] for a more detailed discussion of this roadmap.

4.2.1.3 IBM

IBM's processors are based around the Power Architecture. Firstly it has the Power range of processors, which have been around since the early 1990s. They have been used in IBM's range of servers, generally in large shared memory machines. These have been linked together to form HPC machines using a variety of different interconnects over the years. The current processor is the Power7+, although the Power8 processor was recently announced by IBM at the Hot Chips conference [24], with a release date in 2014. This will be a based on a 22 nm process, and each core is capable of running eight simultaneous threads. Each chip will have an integrated PCI-Express 3.0 controller. IBM has also designed a new transport layer called the Coherence Attach Processor Interface (CAPI) to operate over the PCI-Express 3.0 bus. This layer will allow accelerators such as GPUs or FPGAs plugged into the bus to access main memory. This is a very interesting development for the HPC market.

The other IBM processor used in HPC is of course the PowerPC A2 as used in the Blue Gene/Q. It has 18 cores, of which 16 are used for compute. One is used for operating system services whilst one is a spare which is normally shut down.

4.2.1.4 ARM

The UK company ARM is not a processor manufacturer, but rather a processor designer. It licenses its technology to manufacturers to use in their products. Historically it has specialised in processors with low power consumption, and so ARM processors have been used in mobile phones, tablets and handheld devices. Microprocessor companies manufacturing ARM processors include AMD, Nvidia, Qualcomm and Samsung.

The recent interest in developing low-powered servers for data centre use has prompted ARM to develop a range of 64-bit processors [25]. These have been picked up by a range of manufacturers, typically for use in a SoC. Users include HP, Dell and Calxeda, all of whom typically integrate processor, memory controller, I/O and network controllers on the same silicon.

4.2.1.5 Nvidia

Nvidia's main business is developing high performance GPUs, particularly for the computer gaming market in the form of the various GeForce ranges. However it recognised that its products were being used by some for technical computing, and put some effort into developing general purpose GPUs (GPGPUs). Since then it has developed this range, along with the CUDA language to ease their programmability. They began to be taken up more widely when the GPGPUs were produced with ECC memory and double-precision IEEE compliant arithmetic. The current Kepler K20X model is capable of 1.31 Tflops double precision performance utilising 2688 CUDA cores [26].

Nvidia also produce the Tegra SoC for mobile devices [27]. This integrates an ARM CPU, Nvidia GPU, memory controller and I/O in one package. It is currently targeted at mobile devices; however it raises the prospect of future devices being used in HPC.

4.2.1.6 Processor Licensing

The practice of licensing microprocessor designs to other manufacturers has been commonplace for some time. ARM has operated using this model very successfully for quite some time, mainly in the mobile and embedded market. However, as described above, interest in their designs has increased in other areas, mainly due to their low power consumption, but also to provide a processor where a company might not have an existing design. The obvious example of this is Nvidia, who are combining ARM cores with their own GPUs on the Tegra device. It is also interesting to note AMD licensing ARM technology for its own low powered servers, as noted above.

The trend towards SoC can only lead to an increase in this operating model. Companies with existing technology in a particular area, for example interconnect, will likely end up licensing this technology to another company to include on their processor. The alternative to this model is to buy a technology outright; however this is often too costly to contemplate, except for the largest companies. Such an example is Intel, who have recently acquired both the Cray interconnect intellectual property and QLogic's Infiniband program.

However, an interesting development in this area is the entrance of IBM, who have recently announced their intention to license their Power technology. They are doing this through the formation of the OpenPOWER consortium, initially with Google, Mellanox, Nvidia and Tyan [29]. Although IBM have collaborated in the past to license the PowerPC processor technology, as used in the Sony Playstation Cell processor, this is the first time that they will have made available their core Power processor technology. Such a prospect is intriguing for HPC – the suitability of ARM processors for HPC workloads is yet to be proven, whereas IBM's Power processors have a long history of being used in HPC servers. Given the initial partners it is easy to imagine a product integrating Power compute cores, Nvidia GPUs and Mellanox interconnect technology for example; an interesting prospect indeed.

4.2.2 Memory

When thinking about HPC machines it is easy to overlook memory technologies – it is usually the CPU that gains all the headlines. An overview of current and possible future directions is given in D2.1.1 [1], so that will not be repeated here. Instead we discuss possible issues affecting HPC.

Some HPC applications are already limited by memory bandwidth. Unfortunately the increase in DRAM bandwidth has not kept pace with the increase in FLOPS, particularly when going to multi-core processors, when the bandwidth must be shared between cores. Either increased bandwidth has to be provided in the future by using new technologies, or some applications will need to be completely rewritten. Better communications between processors and memory need to be provided to enable an increased bandwidth, and one would hope that the trend towards SoC will eventually include memory, perhaps in a stacked design, although it is not clear how close this is to fruition.

The growth in accelerators has led to memory locality issues. Programmers now have to worry about moving data between different memory locations depending upon which part of the system needs to operate on it, the main CPU or an accelerator. However, technologies discussed above such as AMD's hUMA and IBM's CAPI seem designed to overcome this problem. Therefore one hopes it is a short-lived trend and future heterogeneous architectures will provide uniform memory access. Indeed companies such as Convey Computer are looking to provide global virtual addressable memory across heterogeneous architectures.

4.2.3 Interconnect

The interconnect in a supercomputer is critical, and has been lagging behind CPU technology in the same way that memory has. In order to achieve Exascale performance from real applications it is important that this is addressed. The trend towards SoC shown by all manufacturers can only be a good thing; there are signs that, particularly for low-powered data centre chips, the fabric controllers are beginning to be integrated with the CPU. As previously mention silicon photonics are important here.

An interesting development looks to be on the cards from Intel. They look ready to announce a new optical connector called MXC [31][32], based on silicon photonics and new fibre technology. This will allow optical signals to go greater than 300 metres at 25 Gb/s, with a headline peak transfer rate of 1.6 Tb/s. To put this in context, the top-end 12x EDR InfiniBand link provides 300 Gb/s of throughput.

4.2.4 Software

The current trend towards heterogeneous systems, i.e. traditional CPUs linked with accelerators of some kind, is clearly a challenge for application developers. Currently developers have to worry about managing memory on each device, explicitly moving data between main memory and device memory. Although those looking to exploit the latest technology will always learn the techniques required, it provides a significant barrier to non-expert developers such as scientists who are just attempting to write simulations or analyse their data. It is therefore crucial that the move towards heterogeneous computing is supported by appropriate software tools.

The most obvious example of software supporting adoption of new hardware is in the GPU area. When GPUs started to be used by a few determined individuals for technical computing they were programmed using the shader APIs meant for graphics programming. However, once their usefulness was demonstrated, Nvidia produced the CUDA language extensions for C, which eased their programmability and therefore rapidly accelerated the uptake of GPUs in technical computing and HPC. CUDA is now at version 5.5 and in this latest version supports programming of ARM devices [30]. CUDA extensions were also added by the Portland Group (PGI) to their compilers, allowing GPUs to be also utilised by Fortran developers. Recently Nvidia purchased PGI [33], which shows that they see long-term value in the HPC market for their products.

However, CUDA is not the solution for everyone. While useful for exploiting the full performance of GPUs, it still requires developers to explicitly consider the architecture and the movement of data between host memory and device memory. Recent developments such as OpenACC [34] and version 4.0 of OpenMP [35] are aimed at making programmability easier, although developers still have to decide themselves which loops or code fragments are suitable to offload to accelerators. Currently OpenMP 4.0 is not yet implemented, and although OpenACC is supported by the CAPS, Cray and PGI compilers, it is only really available for Nvidia GPUs.

To support its hardware Intel has a well-integrated development suite. To support the recently released Haswell range of processors with inbuilt graphics, Intel announced the Intel SDK for OpenCL Applications 2013, which allows developers to target both the CPU and GPU components of Haswell. Intel also has their Threading Building Blocks (Intel TBB) [37] for task parallelism, and Cilk+ extensions [38] for multicore and vector processing. It will be interesting to see if any of these gain traction in the HPC market, but this would seem to depend on whether Intel processors come to dominate HPC completely in the future.

Another interesting development is the formation of the Heterogeneous System Architecture (HSA) Foundation, a not-for-profit consortium of SoC designers and vendors, software companies and academia, many from the mobile and embedded space. Partners include AMD, ARM, Qualcomm and Samsung amongst many others, although it is interesting to note that neither Intel nor Nvidia are members. The aim is to make it much easier to program heterogeneous parallel devices including CPUs, GPUs, DSPs and other accelerators. A good overview of an HSA presentation is available at [40]. Here it is explained that a key feature of HSA will be to “move the compute rather than the data”. This will be achieved through unified memory addressing so that memory can be allocated on one processor and then a pointer passed to another processor for execution on that data. Although currently available as language libraries the goal is to push HSA as low as possible so that it becomes, for example, a part of the Java virtual machine, and it knows which data to send to which processor. Another aim is to ease programmability. An example shown in [40] shows an increase in performance slightly lower to that gained using OpenCL in C, but with much less code complexity. HSA therefore looks to be a development well worth following, but again whether it gains any traction in the HPC world will remain to be seen.

5 System Architecture Trends

Making predictions about future system architectures will always be difficult. A large number of (high-level) factors influence the technology trends that will determine HPC system architectures. It is however possible to draw some conclusions on trends: in the following paragraphs, we will discuss the factors that we believe are most influential in dictating trends in HPC system architectures and then align those with an analysis of historical data. Finally, we will try to make predictions for future systems based on our experience of past developments.

5.1 Defining Influences

The following factors are key influences on the architectures of HPC systems:

- performance;
- programmability and usability;
- power usage and efficiency;
- cost of procurement;
- cost of ownership.

The main factor that influences HPC system architectures is that of performance. HPC systems are, by definition, designed to deliver high-end computational power. However, performance alone is by no means sufficient; the most powerful HPC system is worthless if it cannot be exploited, making programmability and usability an equally important factor. An example of high performance, but low (mainstream) programmability is that of FPGAs. They are capable of delivering the high performance required by HPC applications, however they are notoriously difficult to program and use. The issue of poor programmability is slowly being addressed by FPGA vendors, who are increasingly trying to develop sophisticated compilers to support widely used programming models (such as OpenCL) on FPGAs. Nvidia's success in bringing GPUs to the HPC market is only partially due to their potential performance benefits. The fact that Nvidia also developed the CUDA programming model, allowing developers to write programs for GPUs relatively easily without resorting to programming shaders directly, was much more instrumental in the success of GPUs for HPC and technical computing.

Cost of procurement is also a non-negligible factor that influences the types of system architectures. Funding for HPC has not increased in proportion to the performance that is expected of modern systems; we now want much more performance per dollar spent than we did even 5 years ago. The result of this is that commodity components, which due to economies of scale are much cheaper than custom components, have increased in popularity significantly. The use of off-the-shelf products allows the cost of procurement of high-end HPC systems to remain at realistic levels. The ultimate effect is that the HPC market depends on technology that is designed and developed for the consumer market.

The cost of ownership on the other hand keeps increasing at a steady level. The main factors for this are the power and cooling requirements of top-end systems. The current number 1 system in the TOP500 list, Tianhe-2, consumes a total of 24 MWatt, a quarter of which is spent on cooling. It is clear that the current trend in power consumption of HPC systems is not sustainable. Future HPC systems will need to be vastly more power efficient, not only to keep the cost of ownership at a manageable level, but also because there is a limit in the amount of power that can be delivered to a HPC installation.

5.2 Trends in the TOP500 list

The first version of the TOP500 list [4] was released in June 1993, providing us now with 20 years' worth of data on the most powerful HPC systems. The list not only records system performance (in flop/s), but also additional information such as, amongst others, architecture, interconnect, processors and operating systems. By analysing this information, it is possible to see clear changes in the trends in HPC system architectures over time.

5.2.1 Processors

The first processors with more than one core per socket emerged in the TOP500 list in 2001. A big change then happened in 2006-2007, when the share of systems that had single-core processors went from >70% in June 2006 to just under 15% in June 2007. From that point onwards, multi-core processors have been ubiquitous in HPC systems and dual-core sockets were quickly replaced by sockets with four, eight or even more cores. Up until 2006 performance gains could made by increasing a processor's clock rate and thus the number of floating point operations per second it could perform. Physical limitations (such as heat dissipation) however meant that this was an avenue that could not be pursued forever. Instead, processor manufacturers opted for an alternative: increasing the number of processing cores per socket and enabling parallelism at the processor level.

5.2.2 Accelerators

Not long after multi-core CPUs, accelerator technologies also started emerging in the TOP500. The first systems with GPGPU cards were listed in 2010. The rationale behind introducing accelerators is similar to that for multi-core CPUs, i.e. it is a straightforward method to add more performance to a machine. Another benefit that GPUs bring is the amount of performance they can deliver per Watt; although GPUs used in the HPC environment are not low power per se, their flops-per-Watt ratio make them a relatively energy efficient solution.

It is clear from the June 2013 TOP500 list that accelerators (both Nvidia GPUs and Intel Xeon Phi co-processors) have made a significant impact in a very short period of time. Although only 11% of systems in the list use accelerators, these systems account for over 33% of the performance of the list – largely because they are present in some of the biggest systems. This is illustrated in Figure 11.

The big question is how long it will be before the main obstacle to good performance with accelerators (i.e. the data transfer between host and device via the PCI Express bus) is a problem of the past? It is already possible to run code completely independently on a Xeon Phi without relying on a host processor, and Nvidia GPU Direct RDMA [41] enables RDMA transfers across an Infiniband network between GPUs, bypassing host memory. Similarly, AMD's hUMA technology enables a shared virtual address space between GPU and CPU.

What is clear from the TOP500 list is the massive parallelism already present in the largest systems, mostly due to the use of accelerators. Core counts of several hundred thousand are seen, and Tianhe-2, at the top of the list, has a total of 3,120,000 cores, comprised of Intel IvyBridge processors and Xeon Phi co-processors. By any stretch of the imagination this number of compute cores presents a massive programming challenge to attempt to utilise them efficiently.

5.2.3 System Architectures

In the early days of HPC, supercomputers were largely purpose-built specialist systems constructed from high-end components. They were expensive to design and build and because of this they were not affordable by small organisations such as university departments. Then, during the mid-to-late 1990s, the concept of building small HPC systems from commodity components emerged, suddenly opening the world of (small-scale) supercomputing to a much larger market. This concept of the cluster has evolved over the past 20 years and can now range from small Beowulf-type set-ups all the way to high-end HPC systems (which still rely on mostly commodity hardware) at the top of the TOP500 list. From not being represented at all in the TOP500 list in 1994, clusters represented ~35% of the performance share of all systems in 2003. This has increased to 60% in the latest release of the list in June 2013.

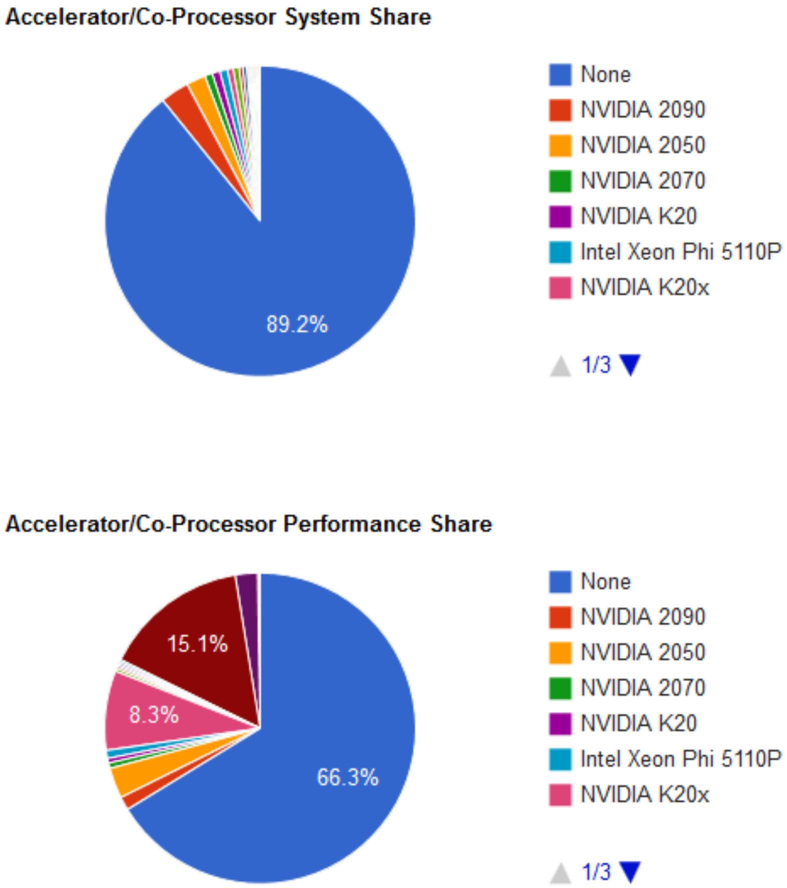


Figure 11: Market share of accelerators in June 2013 TOP500 list, taken from [4].

5.2.4 Interconnect

The interconnect landscape has changed dramatically over the past 10 to 15 years. In 1999, Infiniband and Gigabit Ethernet came to the market and within a few years these two technologies held more than 50% of the performance share of the TOP500. Recently, the dominance of Infiniband has reduced and custom interconnects (such as can be found in the IBM BG/Q) take an increasingly large performance share. The popularity of solutions such as Gigabit Ethernet and Infiniband can be aligned with the rise (and ethos) of the cluster: they are comparatively cheap interconnect technologies which make HPC accessible to a wide audience.

5.2.5 Conclusions from TOP500

The previous paragraphs illustrate one point: a large amount of HPC (though of course not all, the IBM BG/Q being a notable exception) currently relies on commodity off-the-shelf hardware, which is used to build fairly standard clusters that are augmented with accelerators for increased computational power. The evolution of the HPC system over the past 20 years has followed Moore’s law, however we have reached the point we are at today partially by moving away from specialist hardware. This has made HPC more affordable and accessible to a wide audience of scientists, and increased competition. However at the same time it has focussed too much on increasing flop/s performance, leaving factors such as memory, interconnect and I/O behind. Only the very top end – the so-called Tier-0 systems – are now likely to use purpose-built components.

5.3 More Recent Factors

In addition to the TOP500 list, which focuses on the best Linpack performance that HPC systems can deliver, two alternative classification systems have recently emerged: the Green500 list [42] (since November 2007) and the Graph500 [43] list (since November 2010). These lists are analogous to the TOP500 list, but they concentrate on energy-efficiency in the former and data intensive applications in the latter. The Green500 list classifies systems by performance per Watt; it is derived from the results that are submitted for the TOP500. Figure 12 shows how, since the start of the Green500 list less than six years ago, energy efficiency has increased by an order of magnitude. This is testimony to the fact that power usage and energy efficiency are now recognised as real challenges for HPC, especially with a view to the Exascale. However, even using the most energy efficient system today (which can achieve 3208.22 Mflops per Watt), an Exaflop calculation would still require more than 300 MWatts of power.

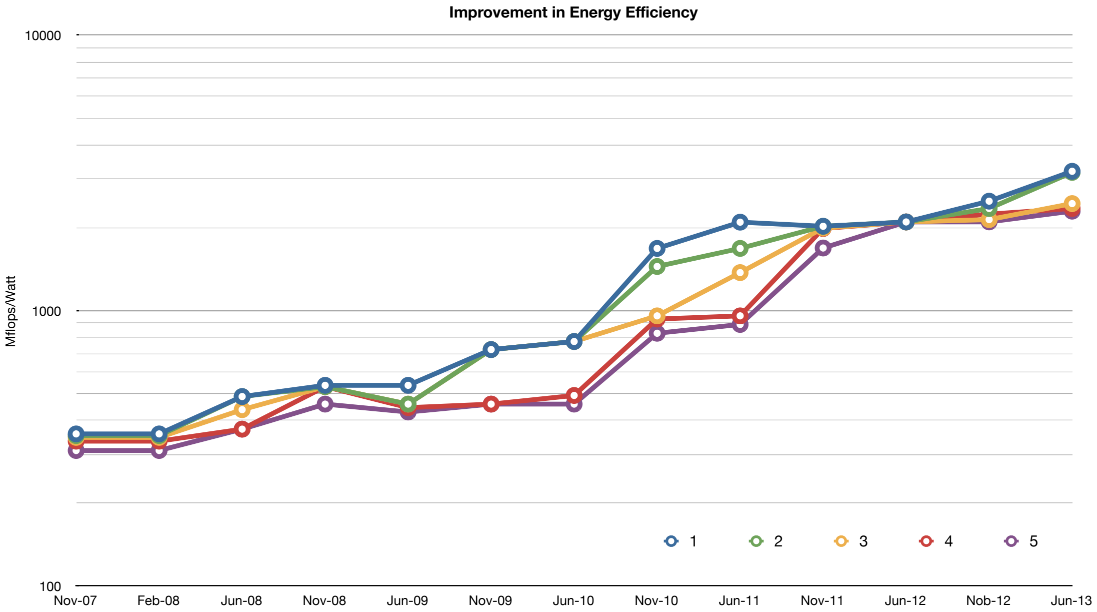


Figure 12: Evolution of the energy efficiency of the top 5 systems from the Green500 list, since November 2007.

The Graph500 list is interested in the ability of HPC system to perform data intensive tasks and it assesses this ability through graph-related benchmarks and search/optimisation kernels. The metric that is used to rank the system is “billions of edges traversed per second”. Since the start of the list in November 2010, this number for the top placed system has gone from 7.08 to 15363. The emergence of the Graph500 list shows that the HPC community is aware of the specialist hardware requirements of this particular HPC application area – the list is an excellent way of tracking evolution and progress.

5.4 Future HPC Systems

Having looked at the influencing factors for, and recent evolution of, HPC system architectures in the previous paragraphs, we can draw a few conclusions and make predictions about future architectures:

- Heterogeneous systems are very likely here to stay. They established themselves very quickly on the HPC scene (with GPUs and Xeon Phis) and have brought both improved performance and (where the parallelism can be exploited) power usage. Future systems may become even more heterogeneous and offer several different “cores” per chip; it is not inconceivable that small conventional CPUs may live along side massively parallel chips and possibly even FPGAs, which could be custom-configured to perform specific repetitive calculations (e.g. sparse matrix-vector multiplications). It is clear that programming models would need to support such systems and allow application developers to target the hardware without the need for in-depth knowledge on how to get be best performance out of each different component. Pragma-based models such as OpenACC, or even domain-specific languages with backend code generators could be the long-term solution.

- A lot of commodity components are used in the TOP500 systems, however it is also clear that the very top-of-the-range systems tend to use custom components. This is a reassuring observation, which shows that in order to get the very best performance together with energy efficiency, commodity components that have been derived from consumer products simply will not be a solution long-term. In order to reach the Exascale, HPC vendors must invest in technology that is specific to the market they are targeting. Using gaming, mobile and Cloud technologies as points of inspiration is of course acceptable, but the supercomputing community should not rely on those markets driving the development for HPC.
- Accelerators brought more flop/s, but other factors such as memory, interconnect and system software, have a lot of catching up to do. The emergence of accelerators has maybe allowed vendors to slightly take their eyes off those issues, but they need to be addressed very soon. Future systems will need to have lower-latency interconnects and faster memory.
- None of the lists that were analysed above mention I/O, however this will be one of the great bottlenecks for Exascale systems and one of the areas where a revolution is due. Future systems will need to have much improved I/O, both through high-performance file systems and an Exascale I/O software stack. HPC applications will of course have to play their part by offloading I/O to separate nodes in a way that is the least intrusive to the computation core of their applications.
- The Green500 has shown a lot of progress in energy efficiency over the past few years, but there still is a very long way to go. The energy efficiency of future HPC systems needs to be increased by an order of magnitude so that the cost of ownership of an Exascale system is realistic and the environmental impact of such as system is kept to as low as possible.

6 Case Study: Technology Trends and the Cray XC30 Supercomputer

The Cray XC30 supercomputer was launched in November 2012 at the annual Supercomputing conference in the USA and represents a major leap in the Cray HPC roadmap. As a completely new engineering product it combines a number of new technologies that were not used in previous generations of Cray supercomputers. The Cray XC30 supercomputer is Cray's first system based on next-generation Intel Xeon E5 processors. Its architecture enables support for a multiple processor line-up of Intel and other x86 processors as well as emerging coprocessors and accelerators. It also introduces the Aries interconnect, using a new system interconnect topology called "Dragonfly". This innovative topology provides scalability in system size and global network bandwidth. Cray continues the evolution of the Cray Linux Environment with the Cray XC30 system, providing a software stack optimized for performance at scale of real-world HPC applications. Cray's programming environment combines flexibility and high productivity features to facilitate effective performance tuning and easy porting. The Cray XC30 system also features increased processor density with high efficiency cooling and power solutions.

The Cray XC30 is the first of Cray's Cascade series systems and the design solution has been influenced by a number of technology, engineering and market trends.

6.1 Processor Trends

When the Cray-1 was launched in the mid-1970s, it was based on a custom-designed CPU. Whilst these are principally remembered for the huge vector performance, this was also the fastest scalar processor to date. Almost all applications ran faster on this CPU than on previous systems, even before the rewards of vectorising the codes. Vector processors continued to dominate the HPC market until the mid-1990s, but absolute performance gains were increasingly only seen for well-vectorised applications that took specific advantage of the hardware. Such processors were no longer the fastest for any kind of application. At the same time, the mass market for commodity servers and, to a lesser extent, PCs drove microprocessor speeds higher than those of the more specialist vector processors, and also pushed the price down sufficiently to warrant the design of massively parallel supercomputers. The performance gains across all processor designs also pushed developers to include increasingly complicated physics and chemistry models in their applications. This application complexity made vectorisation increasingly difficult and accelerated the move towards supercomputers based on commodity microprocessors. This trend continues today, with the Cray XC30 being based on the latest Intel Xeon Ivybridge and Sandybridge E5 server processors.

Vector processors have not entirely disappeared; in many ways the current GPUs can be seen as an extension of these and GPUs are increasingly prevalent in HPC systems, including Cray XK7 systems (such as the Titan installation at Oak Ridge National Laboratory) and future Cray Cascade-class systems, which include Nvidia Tesla series GPUs. What has largely disappeared is the commercial viability of developing such custom, HPC-specific vector processors "in-house". Whilst the Nvidia Tesla series GPUs are specialised for HPC computation, their development is driven by the mass-market GPU models sold by the company, which benefit from the same technology. Likewise, development of the Intel Xeon Phi processors (also available in future Cray Cascade-class systems) is tied to the bigger market for Intel CPUs.

In certain market areas there is still room for HPC vendors to develop their own processors optimised for specialist problems. The Threadstorm processors used in the Cray XMT system and the Cray uRiKA appliance are one example, with the custom design supporting up to 128 threads per processor and a large, globally-addressable main memory. Even here, however, the design benefits from sharing much of its support technology (including blades, interconnect and cabinets) with the larger-volume Cray HPC systems.

6.2 Interconnect Trends

Unlike processors, there is still some scope for vendors to design HPC-specific networks. The cost of the network is significant (especially from a cost-of-goods perspective), however, and market forces do influence the design. The Cray XC30 uses the new Cray Aries interconnect to link the nodes. This combines a custom Application-Specific Integrated Circuit (ASIC) to implement the Network Interconnects (NICs, one per node) and router (one per blade, containing 4 nodes). Avoiding external switches improves the scalability of the network, whilst also controlling system cost.

The Cray Aries network[44] uses a novel "Dragonfly" topology[45] to link the nodes. Previous generations of Cray interconnects (SeaStar and Gemini) required the nodes to be linked in a three-dimensional toroidal mesh. Moving away from this allows improved scalability of applications, particularly on busy systems or when the application is using a significant fraction of the system's nodes. The Dragonfly design is a two-level network. Within a two-cabinet "group", the 96 Aries NICs are linked with electrical cables in an all-to-all pattern. The groups are then linked together using longer optical cables. In this way, any two nodes can be linked in 5 network link hops, only one of which is optical.

The number of optical cables needed depends on the number of cabinet groups in the system, but can also be adapted to customer needs. For a system with 8 groups (16 cabinets), 952 optical cables are used for a full network, giving 20 TB/s bisection bandwidth. To reduce costs, fewer optical cables can be used, with a linear reduction in the bisection bandwidth. Further details on the topology can be found in [44].

The Cray Aries network also continues the trend towards NIC designs (which started with Gemini) more suited to the multi-core CPUs used in modern HPC systems. The SeaStar interconnect used in the Cray XT series was designed (initially by Sandia National Laboratory) with the emphasis on the prevalent two-sided communication models (principally MPI). It contained an MPI offload engine, running on a dedicated PowerPC CPU. Message matching is, however, complicated and the speed of the PowerPC processor limited the peak uncoalesced message rate to around 250,000 per second. As the clockspeed and core-count of the nodes' CPUs increased, this limited scalability and performance. The succeeding Gemini and Aries NICs moved away from MPI offload to focus on hardware support for the Remote Direct Memory Access (RDMA) operations needed in single-sided PGAS communication models. MPI message matching moved onto the CPU, avoiding the possibility of the node/NIC CPU performance gap seen with SeaStar. This massively increased the peak uncoalesced message rate to around 10 million per second with the Gemini NIC. The current Aries NIC improved the hardware support for PGAS operations. The focus on RDMA operations in Gemini and Aries continues the trend seen in the much earlier Cray T3E architecture.

Interconnect message latency dropped significantly between the Cray SeaStar series interconnects and the (in-place upgradeable) Cray Gemini interconnect, reducing from 3-4 μs to around 1.5 μs. This was largely due to removing the by-then slow PowerPC CPU from the NIC. The latency has improved slightly with Aries, to around 1.2 μs. As hardware limits are approached (principally the PCIe protocol and the host interface), it is unlikely that MPI message latency can be reduced much below 1 μs without much tighter integration of the CPU and NIC. Lighter-weight single-sided PGAS programming models have an advantage here; the latency of an RDMA put operation is already much less, at around 700 ns for the Aries interconnect.

Network bandwidth, however, is markedly increasing, especially with the introduction of optical cabling between Cray XC30 cabinet groups. The SeaStar NIC had an injection bandwidth of 2.3 Gb/s and a 6:1 ratio of router to injection bandwidths. On a toroidal mesh machine, this allowed messages to make 6 hops before network suffered traffic congestion (assuming traffic were evenly distributed). With Gemini, the injection bandwidth was increased to 6 Gb/s, but the router/injection excess was only 3:1. This made application performance more sensitive to job placement on the toroidal mesh (as is being studied in WP3 of CRESTA and noted in [46] and [47]). Aries increases the ratio once more to 5:1 but, crucially, the Dragonfly network means a (minimally-routed) message can get between any two nodes in only 5 hops. The network is thus very balanced and application performance is much less sensitive to job placement.

Global bandwidth is important in applications, as it reduces the penalty for non-local communications. A relevant metric is the ratio of global to injection bandwidth. A very distributed data analytics application might require a ratio of 1 to 1.25, but for most nearest-neighbour domain decomposition codes the ratio is nearer 0.25.

Once more, economics influences the network design; the optical cabling is expensive, so customers have the flexibility to choose how many cables to install (balancing cost against bandwidth needs for their application mix). The Cray XC30 also represents a shift towards more generic data buses, using PCIe to connect the CPUs to the NICs rather than the AMD HyperTransport protocol used in the Cray Rainier and Baker class systems (the XT, XE and XK architectures). This allows a greater flexibility in node design, making it easier to connect in a variety of accelerators, for instance. This is another aspect of the general trend away from bespoke or vendor-specific solutions towards more flexible, industry-standard approaches.

6.3 System Integration and Software

An HPC system requires a lot more than a CPU and an interconnect to work. A key component is the hardware that integrates the system together. The trend has been towards more densely-packaged systems. The original HECToR installation comprised 80 cabinets of Cray XT4. This was upgraded to 20 cabinets of Cray XT5 (which was upgraded in place to Cray XE6). The follow-on Archer system is likely to be around half the number of cabinets of Cray XC30. At each step, there was a significant jump in the performance of the system, despite the smaller footprint. This is only possible through advances in system infrastructure and particularly cooling. The Cray XT5 EcoPhlex cabinets moved away from vertical ambient air cooling towards using liquid-cooled chillers above each cabinet. The Cray XC30 uses one set of transverse blowers per two-cabinet group, with regular water-cooled chillers ensuring the constant air temperature along the row of cabinets. This is better both from an energy-efficiency (and thus economic) point of view as well as for performance. Modern CPUs have the facility to internally boost their clock speed when environmental conditions allow. Maintaining a constant low temperature inside the cabinet increases the proportion of time the CPUs can run in a boosted state, and gives measurable gains in application performance.

As mentioned previously, the emergence (or perhaps re-emergence) of accelerators in HPC has created a need for flexibility in system design to allow customers more control when configuring their systems. This trend was already evident in the mid-2000s in Cray Rainier class systems, when the Cray X2 vector processing blades could be substituted into Cray XT5 cabinets. Similarly, GPU-accelerated XK blades were interchangeable with CPU-based XE blades in the later Baker class systems. The Cray Cascade design takes this further, with each blade holding two daughter-cards. These can then be swapped to not only allow accelerators to be added to the system (in the near future, Nvidia GPUs or Intel Xeon Phi coprocessors), but also to allow a more flexible CPU upgrade path that is less constrained by socket design. This is also part of the trend towards HPC systems being incrementally upgradable systems, rather than being fixed in architecture. The HECToR installation was almost completely replaced when moving from the original Cray XT4 system to the final Cray XE6, however this was done as three separate replacements of the cabinets, CPUs and interconnect.

The Cray XC30 continues to use the Cray Linux Environment, which runs a stripped-down Linux kernel on the compute nodes. Omitting unnecessary OS services leads to far few interrupts on the nodes. This significantly reduces "jitter"; fluctuations in application performance resulting from these interrupts that ultimately affect the performance and scalability of applications. Additional performance gains come from tuning the CPU BIOS. Cray also develops its own MPI libraries, adding Aries-specific code to the widely used MPICH2 implementation. The Fortran coarray (CAF), UPC and SHMEM implementations likewise take advantage of Aries-specific features.

Again, a clear trend here is towards more Open Source software and away from proprietary packages. The Cray XT3 (and earlier Red Storm) systems used the Catamount OS, but for the past few generations of systems a modified Linux has been developed. The earlier Cray T3E did not have a vendor-supplied MPI (this was written by EPCC) but Cray did supply a PVM communication model, whereas Cray now adapts an industry-standard MPI version. This trend is likely to continue; fewer supercomputing centres are single-vendor "closed shops" and users increasingly demand application portability. It is also expensive for vendors to develop and maintain entire custom software stacks.

6.4 Summary

A number of trends have led to the current design of the Cray XC30. A clear trend in both processors and system software has been to move from vendor-specific, custom-designed solutions towards use of more standard components. Not only does this respond to the customer-driven trend for application portability, but it also satisfies the requirement for competitive pricing. This move has also driven some of the system design, such as the Cray XC30 using the more-flexible PCIe data bus and blade daughter-boards to allow more flexibility and upgradability in the system configuration.

Interconnects are currently still an area where vendors can justify bespoke developments like the Cray Aries interconnect. The trend towards performance (and price) flexibility is also evident here; in addition to the long-standing choices of processor and memory speeds, the Aries network offers an additional variation in the amount of optical cabling.

After a number of high-profile purchases of interconnect technologies by major processor vendors, however, it is not yet clear how interconnects will develop over the next five to ten years. A tighter on-die integration of NIC and CPU would overcome the current PCIe (or similar) hardware limits on the message latency, but this may take a while to fully develop.

In recent years, energy consumption has become a driving force in HPC design, perhaps for the first time. This already influences some of the system integration design choices and it is also likely to have an effect on the node architecture. The use of accelerators is one part of this, but with a number of competing approaches at present, it will be the market that ultimately decides how future architectures beyond the Cray XC30 will look.

7 Application Impact

A key feature of CRESTA is the presence of the six co-design scientific applications, which guide the systemware developments, and also take advantage of developments in the rest of the project. Therefore for the purposes of this report they have a very important role to play. They can provide important feedback on how current and future architectures may impact application performance and future developments. They can also provide guidance on how architectures may be best developed for the benefit of the applications. We present here the results of discussions with the applications on this subject.

7.1 Elmfire

7.1.1 Introduction

Elmfire is a particle-in-cell code for the simulation of plasma in a toroidal geometry as used within fusion reactors. It is a Fortran code which is approximately twenty years old and is under continuous development with respect to both the physics used in the simulation and to improve its performance. Traditionally the performance work has been to target the current generation of machine architectures. In this respect it is similar to many scientific codes, with the parallelisation being added as an afterthought rather than being designed with parallel performance in mind. The scientific owners of the code are also reluctant to make major changes to the code due to the cost involved and the difficulty in verifying the correctness of such changes.

7.1.2 Impact of Current Architectures

As processor counts have increased then Elmfire has not been able to scale accordingly, mainly due to its memory consumption. This is largely due to the collection and distribution of electric charge data. Work is ongoing to address this in CRESTA. A global electric field is required as it turns the traditional O(N^2) problem of particle-particle interactions into an O(N) algorithm. This algorithm propagates each particle forward in time and space according to the global electric field and a given magnetic field, before the charge contribution of each particle towards the global field is collected.

Due to the requirement to calculate the global field, collective operations are required at every timestep. This provides the main bottleneck, as the forward propagation part can be completely parallelised. Again work is ongoing to address this within CRESTA.

7.1.3 Impact of Future Architectures

The challenges of future machines are expected to be covered by further developments of the improvements described above. As accelerators become more widespread then they may easily be used by the particle propagation part of the simulation, due to its parallel nature, however the difficulty will then be in the collective communications and ensuring that this does not become a major bottleneck. The major impact of heterogeneous architectures is on the developer, providing a programmability challenge, especially for developers who are foremost scientists rather than software engineers.

7.1.4 Requirements on Future Architectures

For Elmfire ideally a future machine would have lots of cores with plenty of memory per core and a fast network for the collective operations required. For the forward propagation step Elmfire could make use of a SIMD machine with an extremely large vector length, of the order of 1000 double precision numbers.

As the developments will target the current machine, whatever that may be, most of the requirements are for support for the developer. This includes better monitoring and profiling so that a developer can get an easy overview of what is happening on the machine, covering processor utilisation, memory consumption and communication patterns and usage. Ideally utilising accelerators would be done by compiler support rather than requiring specific developer effort. Lastly, the owners and users of the code need it to run on a wide range of available machines, therefore they require any performance developments to use well-supported, standardised developments.

7.2 GROMACS

7.2.1 Introduction

GROMACS is a molecular dynamics package, primarily designed for simulating biomolecular molecules such as proteins, lipids and nucleic acids. Throughout its history the GROMACS developers have focussed on making the most of the computer hardware available to them, therefore they are well placed to comment on architectural trends and challenges.

Changes to the GROMACS code may be described as transformative, that is working around limitations in the hardware available to them to gain an increase in performance, or incremental increases in performance working to make best use of that hardware. When considering improvements, these should also be done to increase real application performance for the overall benefit of the scientific results, rather than to produce "artificial" scaling curves.

An example of a transformative approach is that of ensemble calculations to increase simulation throughput, and provide answers to information that may only be obtained through the statistical analysis of many systems, such as the free energy. This has become necessary due to the rapid growth in the number of compute cores not being matched by the growth in communication bandwidth and latency between the cores. Thus it becomes infeasible to simulate a realistic system on tens of thousands of cores as each core has fewer and fewer floating point calculations to perform.

On the other hand, the incremental improvement approach to GROMACS is demonstrated by the fact that the application code, at almost two million lines, contains a lot of highly-optimised code for specific machine architectures. Examples include fully utilising SIMD support of modern x86 processors, and CUDA-based acceleration on Nvidia GPUs, both of which have been added within the CRESTA project.

7.2.2 Impact of Current Architectures

As an example of the impact that changing architectures can have on applications it is informative to look at the impact of GPUs on GROMACS design. This has involved an effort, sometimes painful, running over the last two years, to try and obtain GPU accelerated code which runs faster than the highly-optimised CPU code of previous versions of GROMACS. However, this effort has led to a code restructuring which should be of benefit to future heterogeneous architectures, whatever they may be.

It has involved moving as much of the core force calculations of GROMACS to the GPU, while also trying to optimise data movement. Rather than looking in the traditional way of explicitly overlapping computations with communications, it has considered both of these as different operations on the data, with dependencies between the different operations to be performed - either computation or communication. In parallel molecular dynamics simulations it is necessary to communicate particle co-ordinates between processes. While these are being transferred it is possible to start computations on local co-ordinates, but as soon as remote information is received then it becomes necessary to interrupt this local computation to work on the remote co-ordinates so that the results can be communicated back.

The major concern in this work has been to cope with the communications latency, particularly to the GPU. However, the code has been restructured as a result to cope with heterogeneous architectures, by considering two different types of compute cores. These are latency optimised cores (e.g. the typical x86 derived processors), and throughput optimised cores (e.g. on a GPU). The goal is to offload computations onto the throughput optimised cores.

7.2.3 Impact of Future Architectures

GROMACS appears to be well placed to cope with future architectures, through a combination of the ensemble approach described above, along with the code-restructuring for heterogeneous machines. The move towards fast memory local to cores at the expense of global shared memory is beneficial, as GROMACS attempts to always keep as much as the computation as possible within cache anyway.

Due to the simplicity of the core molecular-dynamics algorithm, and the simple dependency tree, it is often simpler to adapt this directly to new architectures and programming models rather than use some sort of framework. An exception to this may be in the efficient execution of tasks, which may take advantage of a framework such as the Intel Thread Building Blocks, rather than relying on a self-written task scheduler. Work to look at task parallelism is ongoing in CRESTA.

If I/O improvements fail to keep pace with the increase in the number of cores then this may present a barrier to ensemble calculations. Each of these produce data which needs to be written to disk for later analysis. Solutions may be required which only store a subset of data produced by each separate simulation, or which parallelise the post-run analysis step.

7.2.4 Requirements on Future Architectures

Requirements on future architectures may be summarised by the need for low (ideally zero!) latency between compute cores. This is needed far more than bandwidth. There is also a desire to remove the complicated data transfers currently required between CPU and GPU. Both of these may be solved by the move towards integration of a more powerful CPU core with a multi-core GPU on the same silicon die, together with access to the same shared memory. The major concern is that the CPU core still needs to be quite powerful to cope with the serial parts of large scientific applications such as GROMACS, where it is still difficult to exploit any parallelism. Although these may be a small fraction of the overall code, if they are run on an underpowered CPU then they will significantly hinder any parallel speedup.

7.3 IFS

7.3.1 Introduction

The Integrated Forecast System (IFS), developed at ECMWF, forms the basis for all the data assimilation and forecasting activities there. It is a multi-million line Fortran code base developed over many years. Its use for operational weather forecasting means that it has to be stable, with a well-defined roadmap for future development to increase model resolution. Along with the need to produce forecasts to a defined schedule, ECMWF also has a limited power budget for computers, perhaps limiting future power to somewhere between 5-10 MW. This has to be divided between two machines which are run independently for operational reasons.

7.3.2 Impact of Current Architectures

For the petascale era IFS is parallelised using MPI to run thousands of tasks, with the addition of OpenMP to run between 8 and 16 threads per task. Within CRESTA IFS has been extended to use Fortran 2008 coarrays to overlap computation and communication whilst also reducing the volume of halo data communicated between tasks.

Although ECMWF have been aware of the trend towards GPGPU and accelerator technology in large machines little research on their use for IFS has been done so far. One reason for this is the complexity and size of the IFS code base - it is a major undertaking to consider moving to a relatively immature technology, especially when such technology is still challenging to program. In addition, the usefulness of GPGPUs to IFS is limited by the need to transfer data between the CPU and GPU over the PCIe bus, along with the limited amount of memory available considering the number of threads which may run concurrently on a GPGPU.

7.3.3 Impact of Future Architectures

As previously described, future Exascale machines will consist of millions of cores. Taking into account the fact that communications latency has a lower limit, it becomes infeasible to envisage a global communication between millions of tasks - this would just take too much time. Therefore it is necessary to move to a model of keeping the number of tasks to no more than O(10,000) but with 100's or 1,000's of threads per task. It may also require changing the model to consider a 3D parallelisation scheme where a 2D scheme is used today.

There is therefore a recognition that future IFS developments need to take place to enable it to run on GPGPU like systems in order to make use of the many threads available per task. In general this will require major code restructuring and the exposing of much greater parallelism within the IFS code. Current ideas include:

- using co-models (for instance radiation, wave, land-surface) which run in parallel with the atmospheric model,
- using directed acyclic graph (DAG) technology to execute tasks whilst being more sympathetic to jitter and allowing dynamic load balancing,
- developing a new I/O scheme which may require the use of dedicated, larger memory nodes,
- developing a new solver and discretisation method which uses only local communications.

All these developments are major in terms of both time and effort required, and must ensure that the code remains both portable and maintainable for use for production forecasting.

However, the move towards GPGPU-like technologies may also present opportunities for IFS in terms of new algorithms. Such an example is the possibility to run radiation computations on the same grid as the atmospheric grid, and at every time step, rather than running them on a much coarser grid only every model-hour.

7.3.4 Requirements on Future Architectures

Before beginning serious development of the above ideas it is critical that the limitations of current GPGPU technology, as described above, are addressed. The GPGPU-like compute cores need to be integrated on the same die as a small number of conventional CPU cores. They should have access to a single addressable memory. The communication cost between CPU and GPU should also be much reduced by this development.

Such technology must be easy to program. This is required due to the size of the IFS codebase. In addition, due to the need to ensure the code remains maintainable, any development needs to be done using standards that will be well supported. This is also required by the ECMWF procurement system, which needs IFS benchmarks to run on a range of vendors' systems in order to ensure competitiveness. At the moment it is unclear whether this standard will be OpenACC or a development of OpenMP.

7.4 Nek5000

7.4.1 Introduction

Nek5000 is a computational fluid dynamics solver based on the spectral element method. It consists of about 100,000 lines of code, largely written in Fortran77 with some about 10% written in C, being based on code developed during the 1980s. It's design is therefore largely based on machine architectures of that time. The computations are based on large numbers of multiplications of small matrices.

7.4.2 Impact of Current Architectures

As machines have grown bigger, containing more and more cores, then collective communications have become important for the scalability of Nek5000. This has led to alternative implementations of collective communications being developed within CRESTA, as described in [48]. These use non-blocking communications, optimised for latency. The improvements are visible in both overall runtime and scalability of Nek5000. These collectives will continue to be optimised, perhaps making use of MPI 3.0 features. Nek5000 is able to perform its own benchmarks to determine the best collective to use for a particular system.

At the same time, work is ongoing within CRESTA to offload the computationally intensive parts of Nek5000 to GPGPUs. This work has been performed using OpenACC directives, requiring a few hundred lines of code (less than one percent of the overall code). Since Nek5000 is a legacy code it is felt that OpenACC has been essential to this effort - it would have been much more difficult using CUDA.

7.4.3 Impact of Future Architectures

As machines become larger and more complex then it will become increasingly important to use techniques such as auto-tuning to get the best out of Nek5000. This is because the potential parameter spaces will become larger and larger. As well as such things as compiler flags this will apply to the selection of the best parameters for offloading the matrix-matrix multiplications to accelerators, and selecting the best collective communication algorithm to use.

7.4.4 Requirements on Future Architectures

Any future machine for Nek5000 should have a network which is able to cope with the collective communications described above. This should be low latency, and be balanced to match the floating point performance of the machine. As the majority proportion of the computational cost of Nek5000 is taken up by small matrix-matrix multiplications it would be ideal if the floating point architecture was suited to this. Perhaps this could be provided by FPGAs?

To support such a machine should be a good software stack. This would support autotuning of parameters, in the compiler if at all possible. The compiler should also support communications by supporting single-sided communications rather than relying on libraries to provide this. For legacy codes such as Nek5000 it is important that new programming models are easy to implement, as OpenACC has proved to have been rather than using CUDA.

7.5 Summary of Application Impact

From discussions with the application owners, as recorded above, several points stand out. The applications stand ready to try and make use of bigger and faster supercomputers, but only as long as they are practical to program. The massive increase in parallelism requires an interconnect architecture with both the latency and bandwidth to support communication between the vast number of processors. Memory architectures need to be as simple as possible, ideally with a single shared address space between the main CPU and any accelerator that is connected to it. The tools and programming models need to support the developer, ideally through standardised language features so that the applications are maintainable and portable across machines.

8 Conclusions

There is no doubt that reaching the Exascale is a major challenge for all involved – machine vendors, component manufacturers, software providers and particularly application developers. Technology continues to progress, as it always has done, so that an Exascale system is inevitable. However, such a machine needs to be built to consume a "reasonable" amount of power, but above all be useable and programmable.

It seems likely that some components for an Exascale machine will leverage the industry wide quest for power efficiency coming from the mobile computing market; however this seems largely dominated by improvements in processor technology. In tandem there must be improvements in both memory performance and interconnect performance to deliver a true supercomputer. It also seems likely that despite the trend towards using commodity components for HPC machines, recent developments in processor licensing and customisation will allow for component variants targeted specifically at the HPC market, especially as this market grows.

Although the market, both in a wider sense and specifically for HPC, may provide a driver for some of the technological development required it seems likely that government funding will still be needed to drive forward an Exascale machine. Such funding is driven by both national pride and the desire to increase national competitiveness in both scientific research and industry. The spread of supercomputing throughout the world can only help this drive.

It seems likely that heterogeneity is here to stay, integrating CPU and accelerator in one way or another. Whether this be a GPU, co-processor, FPGA or some other device is hard to predict. However it is clear that heterogeneous systems need to be programmable to support application developers. The trend to increasing integration between CPU and accelerator in a SoC device will help this, as will the trend towards single addressable memory spaces. Indeed, some of the CRESTA applications see this as critical to their take up of accelerators.

The other main hardware feature demanded by the applications include a high-performance interconnect with low latency and high bandwidth; it is useless to have millions of compute cores without being able to efficiently communicate between them.

Lastly, it is critical that future Exascale machines are supported by an appropriate software ecosystem. They should be programmable using standard, portable techniques, with the heterogeneity being taken care of at a low level, preferably by the compiler or system run-time. Systems should also provide useful monitoring so that application developers can see easily what is happening in terms of processor, memory and network utilisation.

The quest for Exascale promises to provide a challenging but exciting few years for all involved!

9 References

[1] Architectural developments towards exascale, Project Deliverable D2.1.1.

[2] Earl Joseph, Steve Conway, Chirag Dekate, “HPC Trends, The Emerging Market For High Performance Data Analysis (HPDA), and the HPC ROI Model”, talk at ISC13, June 2013, http://www.slideshare.net/insideHPC/idc-hpc-trends-june-2013, last accessed 20/08/2013.

[3] “IDC Forecasts Worldwide Tablet Shipments to Surpass Portable PC Shipments in 2013, Total PC Shipments in 2015”, http://www.idc.com/getdoc.jsp?containerId=prUS24129713, last accessed 20/08/2013.

[4] TOP500 list, http://www.top500.org/statistics/list/, last accessed 9/08/2013.

[5] G. E. Moore, “Cramming more components onto integrated circuits,” Electronics, vol. 38, no. 8, 19 April 1965.

[6] Theo Valich, “TSMC Future Revealed: 450mm Wafers, 16/10/7nm FinFET, CoWoS & More”, 3 June 2013. http://www.brightsideofnews.com/news/2013/6/3/tsmc-future-revealed-450mm-wafers2c-16107nm-finfet2c-cowos--more.aspx, last accessed 9/8/2013.

[7] Stewart Mitchell, “Intel reveals roadmap to 5nm process”, http://www.pcpro.co.uk/news/374626/intel-reveals-roadmap-to-5nm-process, last accessed 20/08/2013.

[8] M. Duranton, D. Black-Schaffer, K. De Bosschere, J. Maebe, “The HiPEAC Vision for Advanced Computing in Horizon 2020”, March 2013.

[9] “SoC Silicon and Software Design Cost Analysis: Costs for Higher Complexity Continue to Rise”, http://www.semico.com/content/soc-silicon-and-software-design-cost-analysis-costs-higher-complexity-continue-rise, last accessed 20/08/2013.

[10] Dale Ford, “Qualcomm Rides Wireless Wave to Take Third Place in Global Semiconductor Market in 2012”, http://www.isuppli.com/Semiconductor-Value-Chain/News/Pages/Qualcomm-Rides-Wireless-Wave-to-Take-Third-Place-in-Global-Semiconductor-Market-in-2012.aspx, last accessed 20/08/2013.

[11] “Intel Tick Tock Model”, http://www.intel.com/content/www/us/en/silicon-innovations/intel-tick-tock-model-general.html, last accessed 20/08/2013

[12] Tony Smith “Inside Intel's Haswell: What do 1.4 BEELLION transistors get you?”, http://www.theregister.co.uk/2013/06/03/feature_inside_haswell_intel_4g_core/, 3rd June 2013.

[13] “Unlock Performance with OpenCL* Programmability on 4th Generation Intel® Core™ Processors with Intel® Iris™ Graphics and Intel® HD Graphics family”, http://software.intel.com/sites/billboard/article/opencl-programmability-4th-generation-intel-core-processors, June 4th 2013.

[14] Rik Myslewski, “Intel to put pedal to metal in 14nm Atom upgrade”, http://www.theregister.co.uk/2013/08/19/intel_to_put_pedal_to_metal_in_14nm_atom_upgrade/, 19th August 2013.

[15] Tiernan Ray, “Intel May Speed Introduction of Better Phone, Tablet Chips”, http://online.barrons.com/article/SB50001424052748704148304579008830525485044.html#articleTabs_article%3D1, 17th August 2013.

[16] “Reimagine the Datacenter”, http://newsroom.intel.com/docs/DOC-4116, 22nd July 2013

[17] Damon Poeter, “

[18] Intel Seeks to 'Re-Architect' the Data Center”, http://www.pcmag.com/article2/0,2817,2422097,00.asp, July 22 2013.

[19] Diane Bryant, “Re-Imagining the Datacenter”, http://download.intel.com/newsroom/kits/datacenter/pdfs/Re-Imagining_the_Datacenter.pdf, July 22nd 2013.

[20] Rajeeb Hazra, “High Performance Computing: The Essential Tool for a Knowledge Economy”, http://download.intel.com/newsroom/kits/datacenter/pdfs/High-Performance-Computing.pdf, July 22nd 2013.

[21] “AMD Unveils Server Strategy and Roadmap”, http://www.amd.com/us/press-releases/Pages/amd-unveils-2013june18.aspx, 18th June 2013.

[22] William Wong, “Unified CPU/GPU Memory Architecture Raises The Performance Bar”, http://electronicdesign.com/microcontrollers/unified-cpugpu-memory-architecture-raises-performance-bar, May 3rd 2013.

[23] Micheal J. Miller “AMD Pivots to ARM on Servers”, http://forwardthinking.pcmag.com/architecture/312794-amd-pivots-to-arm-on-servers, 19th June 2013.

[24] Timothy Prickett Morgan “You won't find this in your phone: A 4GHz 12-core Power8 for badass boxes”, http://www.theregister.co.uk/2013/08/27/ibm_power8_server_chip/, 27th August 2013.

[25] “ARM Launches Cortex-A50 Series, the World’s Most Energy-Efficient 64-bit Processors”, http://www.arm.com/about/newsroom/arm-launches-cortex-a50-series-the-worlds-most-energy-efficient-64-bit-processors.php, 30th October 2012.

[26] “TESLA GPU ACCELERATORS FOR SERVERS”, http://www.nvidia.com/object/tesla-servers.html, last accessed 20/08/2013.

[27] “Introducing NVIDIA® Tegra® 4, The World's Fastest Mobile Processor”, http://www.nvidia.com/object/tegra-4-processor.html, last accessed 20/08/2013.

[28] “Kayla DevKit - ARM Development Platform for CUDA® and OpenGL”, https://developer.nvidia.com/content/kayla-platform, last accessed 20/08/2013.

[29] “Google, IBM, Mellanox, NVIDIA, Tyan Announce Development Group for Data Centers”, http://www-03.ibm.com/press/us/en/pressrelease/41684.wss, 6th August 2013.

[30] Mark Harris, “CUDA for ARM Platforms is Now Available”, https://developer.nvidia.com/content/cuda-arm-platforms-now-available, June 18th 2013.

[31] “CLDS010 - MXC – The Next Generation Optical Connector” https://intel.activeevents.com/sf13/connect/sessionDetail.ww?SESSION_ID=1110, last accessed 20/08/2013.

[32] Gareth Halfacree, “Intel teases 1.6Tb/s optical interconnect tech”, http://www.bit-tech.net/news/hardware/2013/08/15/intel-mxc/1, 15th August 2013.

[33] Ian Buck, “NVIDIA Pushes Further Into High Performance Computing With Portland Group Acquisition”, http://blogs.nvidia.com/blog/2013/07/29/portland/, July 29th 2013.

[34] OpenACC, http://www.openacc-standard.org/, last accessed 20/08/2013.

[35] “OpenMP 4.0 Specifications Released”, http://openmp.org/wp/2013/07/openmp-40/, July 23rd 2013.

[36] “Unlock Performance with OpenCL* Programmability on 4th Generation Intel® Core™ Processors with Intel® Iris™ Graphics and Intel® HD Graphics family”, http://software.intel.com/sites/billboard/article/opencl-programmability-4th-generation-intel-core-processors, June 4th 2013.

[37] Intel Threading Building Blocks, https://www.threadingbuildingblocks.org/, last accessed 20/08/2013.

[38] Intel Cilk Plus, http://software.intel.com/en-us/intel-cilk-plus, last accessed 20/08/2013.

[39] The HSA Foundation, http://hsafoundation.com, last accessed 20/08/2013.

[40] Rik Myslewski, “The 'third era' of app development will be fast, simple, and compact”, http://www.theregister.co.uk/2013/08/25/heterogeneous_system_architecture_deep_dive/, 25th August 2013.

[41] NVIDIA GPUDirect, https://developer.nvidia.com/gpudirect, last accessed 9/8/2013.

[42] Green500 list. http://www.green500.org, last accessed 16/08/2013.

[43] Graph500 list. http://www.graph500.org, last accessed 16/08/2013.

[44] B. Alverson, E. Froese, L. Kaplan, D. Roweth, "Cray XC Series Network", http://www.cray.com/Assets/PDF/products/xc/CrayXC30Networking.pdf

[45] J.Kim, W.J. Dally, S.Scott and D.Abts. “Technology-Driven, Highly-Scalable Dragonfly Topology” In Proc. of the International Symposium on Computer Architecture (ISCA), pages 77-88, 2008. DOI:10.1109/ISCA.2008.19

[46] R.F. Barrett, C.T. Vaughan, S.D. Hammond and D. Roweth, "Application Explorations for Future Interconnects", Workshop on Large-Scale Parallel Processing, at the International Parallel and Distributed Processing Symposium: IPDPS 2013, Boston, MA, 2013.

[47] R.F. Barrett, C.T. Vaughan, S.D. Hammond and D. Roweth, "Reducing the bulk in the bulk synchronous parallel model", 27th IEEE International Parallel & Distributed Processing Symposium (Boston, May 2013).

[48] Michael Schliephake, Erwin Laure, "Towards Improving the Communication Performance of CRESTA's Co-Design Application NEK5000," sc companion, pp.669-674, 2012 SC Companion: High Performance Computing, Networking Storage and Analysis, 2012