D2.1.1 – 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:	M6
Submission date:	31/03/2012
Project start date:	01/10/2011
Project duration:	36 months
Deliverable lead organization	UEDIN
Version:	1.0
Status	Final
Author(s):	Stephen Booth (UEDIN)
Reviewer(s)	David Henty (UEDIN), Erik Lindahl (KTH)

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

Version History

Version	Date	Comments, Changes, Status	Authors, contributors, reviewers
0.1	23/01/2012	First version of the deliverable	Stephen Booth (EPCC)
0.2	22/02/2012	Draft for internal review	Stephen Booth (EPCC)
1.0	28/03/2012	Implements comments from internal review	Stephen Booth (EPCC)

Table of Contents

Index of Figures

Index of Tables

1 Executive Summary

Technological evolution can be thought of as a combination of Incremental and disruptive changes. Predicting future evolution is hard but can be attempted by a combination of extrapolation of the incremental trends and identification of potentially disruptive technologies.

Evolution of hardware architectures to the Exascale is likely to be dominated by power consumption.

Power considerations will limit clock speed so Exascale performance will only be achievable via an increase in parallelism rather than by any significant increase in the speed of individual operations. This is a key concern given the difficulty that many current applications have achieving good parallel scaling on current Petascale systems.

Memory performance is a key factor in determining the performance of applications on current system. Though there are promising developments in memory technology that might go some way towards addressing the memory-wall, memory will continue to be one of the key system parameters at the Exascale. Memory is expected to contribute an increased fraction of the total power costs and so the ratio of memory capacity to computational capability is expected to be much less than in current systems.

The electrical communications between different parts of a node are expected to be a more significant fraction of the overall time and energy costs. As a result node architectures are expected to become more hierarchical and memory access times within a node are expected to become significantly more non-uniform so applications will not only need to exhibit a high degree of parallelism but also a high degree of locality to make good use of these systems.

Inter node communications will have to utilise optical technology to achieve acceptable performance within a reasonable power budget.

2 Introduction

We can divide technological evolution into two types. The first of these is incremental evolution corresponding to the progressive improvement of an underlying approach or technology. The second is disruptive evolution where a new approach or technology comes into play.

In practice incremental development of the underlying technologies seems to frequently result in progress occurring as a geometric progression. Each new generation of technology takes roughly the same period of time to develop and delivers roughly the same percentage improvement at each generation. This continues until some underlying absolute limit is encountered, or an alternative technology takes over. Undesirable developments resulting from design trade-offs, such as an increased power consumption, also take place over the same product generations and can similarly follow geometric growth until some limit of acceptability is reached. Incremental developments can therefore be extrapolated to future dates with some degree of accuracy. However care does need to be taken to identify any technical or economic limits and any disruptive technology changes that might invalidate these predictions.

While it is possible to identify potentially disruptive technological developments from research literature and from preliminary deployments of the technology in niche areas, evaluating their potential impact is much harder.

The geometric growth “laws” are largely self-fulfilling prophecies as they are used to set the expectations and product road-maps for the entire industry. Nevertheless they are extremely robust. A single technological area might undergo a disruptive evolution allowing development at a faster rate, but the lack of equivalent progress in the other technologies needed to produce a product will reduce the impact of this development. In addition as this change will have been unexpected, it will take a comparatively long time for product road-maps to be updated to take advantage of a disruptive change. On the other hand if a single technology hits an underlying limit, additional funding will frequently be channelled to the “problem” area in an attempt to force a disruptive change that will keep the overall industry trend on track.

The evolutions of hardware and software designs are also much more difficult to predict. Though undoubtedly difficult and expensive, design work tends to be much more agile in nature than manufacturing processes. New approaches to software or hardware design can be researched and brought into production much more easily than changes to manufacturing processes. In addition, incremental improvements are more frequently one-off changes rather than general approaches that can be repeated at each product generation. As a result design trends will tend to be shorter lived and harder to predict.

The purpose of this document is to attempt an extrapolation of current computer technology to the Exascale.

Section 3 will attempt to identify current technological trends that can be extrapolated into the future. Section 4 will attempt to identify limiting factors that might invalidate a straightforward extrapolation of current trends. Section 5 will look for potentially disruptive technologies. Section 6 will look at the impact that these changes might make on Hardware and Software at the Exascale.

2.1 Glossary of Acronyms

DRAM	Dynamic Random Access Memory
SRAM	Static Random Access Memory
ILP	Instruction Level Parallelism
SIMD	Single Instruction Multiple Data
HPC	High Performance Computing
GPGPU	General Purpose Graphic Processing Unit

SOC	System On a Chip
RDMA	Remote Direct Memory Access
PGAS	Partitioned Global Address Space
WDM	Wave Division Multiplexing
SMP	Symmetric Multi-Processing
NUMA	Non-Uniform Memory Access
IC	Integrated Circuit

3 Current and Historical Technology Trends

3.1 Logic Trends

The most significant historical technological trend affecting High Performance Computing has been the incremental evolution of micro-electronics characterised by Moore's Law. The original formulation of Moore's Law [1] was an observation of a geometric progression reducing the cost per transistor and increasing the optimal size (again in terms of cost per transistor) of integrated circuits.

The optimal size of an integrated circuit depends on the transistor density and the incidence of defects in the underlying wafer. If the device is too large then the fraction of devices made that coincide with a wafer defect is also high, reducing overall yields and increasing costs. On the other hand large devices are desirable because packaging and manufacturing cost are reduced by increasing *integration*; that is using a smaller number of larger devices. The increase in optimal size is driven by both a reduction in the size of circuit features (the *line-width*) and by an increase in die area. This increase in optimal size seems to apply to various different types of device built using photo-lithography techniques including micro-processors and solid-state-memories.

The more commonly talked about formulation of Moore`s law is in terms of the evolution of the performance of a microprocessor. This is a higher level concept; though obviously related to the number of transistors in the microprocessor, performance also depends on many other factors, in particular the clock frequency. The physical dimensions, operating voltage, frequency and power consumption of a micro-processor are all closely related. The last 30 years of development have been largely dominated by CMOS logic and the expectations of incremental evolution of this technology have largely been set by the principal of constant-field scaling [2]. A new process technology is introduced approximately every 3 years with the feature size reduced by a factor $\kappa \approx 1.4$. The principle of constant-field scaling resulted from the observation that various desirable properties of the transistors can be preserved across this shrink while retaining the same basic transistor design by scaling the operating parameters in the following way:

Table 1 constant field scaling

Parameter	Scaling factor
Device dimensions (including oxide thickness)	$1/\kappa$
Doping concentration	K
Voltage	$1/\kappa$
Current	$1/\kappa$
Capacitance	$1/\kappa$
Delay time	$1/\kappa$
Dynamic Power dissipation/gate	$1/\kappa^2$
Dynamic Power density	1

The reduction in delay time allows the possibility of increasing the operating frequency of the microprocessor. Until recently industry practice has been to reduce the voltage more slowly than the above factors and to increase the clock frequency at a higher rate. Unfortunately this also resulted in a dramatic increase in power density which has

proved to be unsustainable so operating frequencies have stopped increasing or only undergone modest increases in recent microprocessor generations with performance increases being carried almost entirely by increased parallelism within the device. Though this has allowed system performance to continue to scale over processor generations it has become harder to achieve this scaling at the application level as the burden is on the application and library developer to take advantage of this device parallelism. Previously application would see large benefits from an increase in clock rate even if no modification were made to the code at all.

While the progressive shrinkage of integrated circuits has been the dominant trend in the development of computer performance for the last 50 years and it is largely this trend that has set our expectations for the availability of Exascale systems. It is impossible that this trend will continue forever. For example gate oxide layers in current 65nm transistors are already only a handful of atoms thick so future process generations will not be able to simply scale this dimension. The death of Moore's law has been predicted many times in the past, and avoided due to disruptive technological innovations such as the introduction of strained silicon or High-K metal gates. It is worrying to note that disruptive changes of this kind seem to be needed quite frequently (every couple of process generations). So while Moore's law is not dead it currently seems to be being sustained by industry expectations and investment rather than a simple incremental technological evolution.

Currently a similar process revolution seems to be required to address problems associated with an increasing fraction of the power budget being taken up by static power (power consumed by transistors in a steady state) rather than dynamic power (power consumed by transistors switching state) [3]. To address this issue Intel have introduced completely new transistor geometries for use in their 22nm process which will be used in their "Ivy Bridge" processors [4]. Though other techniques are also being investigated to address this issue Intel currently appears to have a lead in this area.

3.2 Memory Trends

Another significant historical trend has been the growing mismatch between processor and memory performance. Microprocessors and DRAM memory are both Integrated-Circuit devices manufactured using similar photo-lithographic processes. Both are growing at geometric rates. Since 2000 DRAM density has been doubling approximately every 3 years where processor logic is doubling approximately every 2 years. However the process technologies used to efficiently produce DRAM and processor logic are sufficiently dissimilar that it is not cost effective to manufacture both kinds of device on the same silicon wafer. A DRAM cell is relatively simple consisting of a single transistor and a capacitor and can be manufactured in relatively small number of fabrication steps. Processor logic is much more complex and requires more fabrication steps. Though it is possible to manufacture a DRAM cell on the same silicon as complicated processor logic the additional fabrication steps (needed by the processor logic) effectively increase the cost per cell of the DRAM. Historically on-chip memory has been manufactured as SRAM rather than DRAM. SRAM is intrinsically more complex than DRAM consisting of multiple transistors but has higher performance helping to compensate for the higher manufacturing costs. However SRAM also has significantly higher energy consumption so on-chip memory is now increasingly being manufactured as embedded DRAM to reduce power consumption, most noticeably in designs from IBM.

Unfortunately memory performance as seen by the processor has been evolving at a much slower rate than memory capacity. This is the "memory-wall" which has become a major performance problem for modern microprocessors. This is in part due to the memory interfaces used to connect these two critical components. Even though off-chip memory is significantly cheaper microprocessors do contain relatively modest amounts of the less cost effective on-chip memory, typically organised as caches.

As these caches are integrated into the same chip as the processor, their performance can scale much closer to that of the processor. The impact of these trends depends on

the data access requirements of the application. On the one hand the growing size of DRAMs and on-chip caches benefit some applications by allowing them to run entirely from a higher level of the memory hierarchy. For example some databases can be fully resident in DRAM rather than on disk and some HPC applications may be able to fit their key working data set into on-chip cache rather than external DRAM. On the other hand applications with larger working data-sets will see only modest benefits due to the impact of the slow increase in effective DRAM performance.

3.3 Communication Trends

The different rate of growth of memory interface speeds is a special case of a more general problem. All communications between different integrated circuits (and communications between different parts of the same IC) are subject to different scaling constraints from IC logic and therefore unsurprisingly their performance evolves at a different rate. Within an IC a straightforward process shrink (as outlined in Table 1 constant field scaling) will result in the resistance of a communication line increasing by a factor of κ and the line response time remaining unchanged. Therefore the relative impact of the communication lines on processor performance and power dissipation will increase with processor generations. This problem is made worse by the growth in size and complexity of processors and the two dimensional layout of processor components. This problem is even worse when connections between different ICs are considered, though size scaling has occurred at the circuit board level the rate of this change is much slower than that of logic.

The performance and evolution of off-chip communication technologies is closely linked to the packaging technology. The traditional method of making electrical connections to ICs by thermo-sonically bonding fine metal wires to metal pads manufactured around the edge of the IC. This process has much less scope for size scaling than the photo-lithographic processes used to make the ICs. These wires are connected to package pins that in turn make connections to the circuit board. The number of connections is limited by the number of bonding pads that can be fitted around the edge of the IC and can only increase slowly as die sizes increase. This has driven most off-chip communication technologies to undergo a disruptive change moving from the use of parallel bit-line interfaces to the use of high-speed serial connections. This makes more efficient use of the limited number of pins as well as being more power efficient. Even though the transistors in the transceivers need to operate at a high frequency (consuming relatively large amounts of power) the overall energy per bit sent is reduced. Unfortunately the same factors that make it uneconomic to manufacture DRAM on the same dies as the processor makes it uneconomic to manufacture high-speed serial transceivers on the same die as DRAM so memory interfaces have continued to use parallel bit-lines and memory interface performance has in fact been evolving at a much slower rate than other networking technologies. High-speed serial memory interfaces have been developed most noticeably the RAMBUS [5] interfaces. However these have typically been implemented using additional devices that convert the high-speed serial interface to a more conventional memory interface implemented by a commodity DRAM device. This limited the performance gains to the extent that the additional costs have prevented these interfaces from becoming a mainstream technology.

One recent trend in packaging technology has been to combine multiple dies in a single package or multi-chip-module, essentially making the connections directly between separate dies, eliminating the package pins and printed circuit board traces that would have been needed if the individual dies had been packaged separately. This radically reduces the wire-length and capacitance of the connection allowing increased speed and improved energy efficiency. Many high-end microprocessors consist of multiple dies in a single package and some mobile electronic devices like the iPad use this technique to stack the memory chips on top of the processor.

Wire-bonding technology is starting to be replaced by flip-chip techniques where the die is mounted face-down and small bumps of solder are used to directly connect the

communication pads to the underlying substrate. This technology has the advantage that the communication pads are no longer limited to the edge of the chip. This technology is particularly common in mobile devices where power considerations and physical size are particularly important.

3.4 Processor Design Trends

The obvious trends in current processor design are being driven by the underlying technology trends outlined above.

The geometric increase in available transistors and the slowed growth in clock-speeds mean that in order to meet the market expectation for a geometric increase in computing power the processor designs need to exploit increased parallel processing. Previous super-scalar processor designs focused on extracting instruction-level parallelism (ILP) from the instruction stream. This had the advantage that from the programmer's perspective the programming model remained essentially unchanged. There now seems to be little of this kind of parallelism left to exploit resulting in the recent trend towards multi-core processor designs. To exploit multiple cores effectively the programmer or the compiler (or some combination of both) need to extract some higher level of parallelism from the problem and generate explicit parallel instruction streams. This hard problem will only become more difficult as technology progresses as the geometric growth in transistor numbers will result in a geometric growth in the number of cores in a processor.

The core count per node is usually increased further by combining multiple dies in a single package to form a composite processor and combining multiple processors in a node. This trend towards "fatter" nodes is driven by a number of factors. Partly this is due to HPC nodes being built out of the same parts used to build shared memory servers where the core count relates directly to the performance of the system. Also the total number of available IO pins, and hence the number of DRAM interfaces per node can be increased. Though the average memory bandwidth per core does not increase this does allow higher bandwidths per core to be achieved when only a subset of the cores are in use. In addition manufacturing cost savings can be made on components that are only required per-node.

There is much competition between processor designs to squeeze maximum performance out of the available gates. Modern processor cores have mostly reverted to relatively simple designs in order to fit as many cores into the available area as possible. Many include additional SIMD instruction sets to exploit ILP. SIMD instruction sets are particularly desirable as they can provide very high energy efficiency per flop. This is one of the major driving forces behind the development of GPGPU/streaming designs. Even more conventional processor designs are significantly expanding their use of SIMD instructions.

Another identifiable design trend is the introduction of inhomogeneous designs where instead of being a simple replication of a single core design the processor is made up from a selection of different hardware components targeted at different types of operation. These may be special purpose functional units, such as cryptographic hardware, that are shared by multiple cores or specialised core designs targeted at particular purposes. For HPC use this typically takes the form of many lightweight cores optimised for high floating point throughput sometimes coupled to a private high-performance memory. The Intel MIC architecture [6] is one example though being fully x86 compatible the "lightweight" cores are only moderately lightweight. The GPGPU accelerated systems can also be thought of as an example of this style of inhomogeneous design but at the other extreme. Current GPGPUs are entirely separate devices requiring the host processor to explicitly copy data to and from the GPU however the trend seems to be towards a tighter integration between GPU and CPU in the future, allowing the GPU some access to the main processor memory as well as its private memory spaces. Though these architectures differ greatly in the details the high level picture is very similar. Very high floating-point capacity is available, but only for computations that can be written to have a high degree of

parallelism and locality in their data access patterns. For applications that contain operations that do not fit into this category the performance of these operations dominate and the floating-point capabilities of the processor become far less relevant.

An alternative way of looking at these trends would be as a trend towards deeper hierarchies of interconnect/memories between compute elements. The original generation of MPP systems used single-core processors, each with its own memory system and interconnected by a single network layer. These evolved into clustered SMP nodes where interconnect between cores was divided into a local shared-memory interconnect and a long distance network. These in turn evolved into clustered NUMA introducing more layers into the memory hierarchy. The recent addition of GPGPU accelerators can be thought of as the addition of several more layers to this.

Many designs have experimented with using private memory spaces (each core has a region of memory that only it can access). Examples include the Cell processor and most types of GPU. This can be seen as an alternative to conventional memory caches. Normal cache coherency protocols require large amounts of inter-communication between the caches to ensure they remain consistent. This becomes progressively more expensive as the number of caches increases. The energy cost of these communications in particular will become a significant issue. Private memory spaces do not require any of this overhead. Instead of data re-use being automatically recognised by the hardware; the compiler or the programmer needs to recognise potential data re-use and make an explicit copy of the data to the private memory space.

Processors have diversified into wide families of designs targeted at different markets such as embedded/portable/desktop/server. There is also a clear trend towards design integration: functionality that was previously implemented in support chip-sets such as memory controllers have been integrated with the main processor. In the embedded computing market space (where power efficiency is critical) this has resulted in System-On-a-Chip (SOC) designs where the entire computer (other than memory) is implemented on a single chip.

3.5 Interconnect Trends

Almost all current HPC systems are manufactured using standard commercial microprocessors. This has been an economic necessity because of the relatively small size of the HPC market sector compared to the global market for microprocessors. This means that interconnect, which is a critical component for HPC workloads, needs to be implemented as a separate component rather than integrated as part of the CPU.

There has been a recent convergence between the technology used for networking within data-centres and HPC interconnects. In the November 2011 Top500 list 45% of the systems used Gigabit Ethernet and 42% used Infiniband. Most current networking technologies (including many of the custom HPC networks) are using very similar high-speed-serial networking technology and will ultimately be limited by the performance of the PCI-e interface or the node memory system. As a result many of these technologies have similar performance for bandwidth limited data transfers and it is difficult to justify anything other than commodity networking components for general workloads. At the limits of processor scaling communication latencies can become a significant factor in HPC application performance. Unfortunately communication latency is not similarly important to mainstream data-centre workloads reducing the incentive to develop low latency commodity networks. Custom HPC networking technology may make attempt to make greater efforts to address communication latency but cannot afford to be significantly more expensive than the more mainstream technologies. In addition the trend towards SoC designs may end up with commodity networking being an integrated part of standard processors making custom HPC networks uneconomic. Communication latencies (as measured from MPI, so including software overheads in the MPI library and operating system) have been relatively unchanged over recent

years so there is little evidence that incremental developments will reduce them much. However the relatively large disparity between the size of the unavoidable propagation latency and the actual latency measured in applications suggest that lower latency communication might be technically possible. Because the number of messages handled by a compute node typically remains constant or increases as a program is scaled to higher node-counts some disruptive improvement in message latencies may be required in order to produce usable Exascale systems.

Longer range and high capacity interconnects tend to use optical rather than electrical signalling. Optical signalling has the advantages of very long range, very low energy losses and very high bandwidth. Unfortunately the costs (both manufacturing and power) of the optical transceivers that convert between electrical and optical signalling mean that electrical signalling remains the dominant technology over shorter length scales. There is a clear trend that the distance above which it is preferable to switch to optical communications is becoming shorter over time. Currently this is roughly at the level of cross machine-room or between racks.

4 Limiting Factors

4.1 Power and Energy

The major limiting factor dominating the discussion of future machines seems to that of energy consumption. The current biggest systems in the Top-500 already consume several MW of power. Extrapolations to the Exascale in the 2018 time-frame generally assume that there will be no market for these systems unless power consumption can be kept down to the 20MW level [7].

By keeping clock frequencies relatively constant and scaling down operating voltage with feature size it might be possible to keep the dynamic power per chip (or at least the power per chip area) roughly constant over time though the static power costs due to leakage current are becoming more significant [3]. The significance of the energy consumption of the memory and interconnect are also expected to increase over time. A recent analysis suggested that the power efficiency of computation is following a similar curve to Moore's law [8]. However this concentrated on desktop and mobile computing platforms so at the very least does not include networking energy costs.

A number of manufactures are actively investigating near threshold logic designs where the operating voltage is reduced to an absolute minimum. Because dynamic power is proportional to the voltage squared, this results in extremely low power designs. However clock speeds are also scaled back with the voltage so parallelism will need to be increased to compensate.

DRAM memory stores data as a charge in a microscopic capacitor manufactured inside the DRAM cell. Each time the cell is accessed, energy is required to preserve the value in the cell. In addition, as charge leaks from the capacitor over time each cell needs to be "refreshed" regularly which also consumes energy. As cell sizes become smaller it becomes harder for these capacitors to retain charge so refresh rates and hence power consumption will increase. As a result of this, all other factors being equal, the proportion of the available energy budget taken by DRAM will also increase.

4.2 Reliability

Machine reliability is also seen as a major challenge for future machines. The rate of failure of a complex system is proportional to the number of components in the system and the rate of failure of the individual components. As machine parallelism increases the component count goes up. In addition as the components are shrunk in size each bit of information is represented by smaller amounts of energy, requiring a smaller external disturbance to corrupt the data increasing the likelihood of an error occurring. Failures are also particularly prone to occur in the connectors that link macroscopic components (IO-pins, cables, board connectors) often due to mechanical stresses. For a fixed machine performance technological trends towards higher levels of integration will reduce this class of failure. This will not occur with the first generation of Exascale systems as these are projected to have a similar (or slightly increased) number of macroscopic components as current top-end systems.

Machine reliability is already a significant issue even for current technology. Errors occur continually in modern machines. The majority of these are detected and corrected automatically. A smaller fraction of more serious errors can be detected but not corrected resulting in a failed job that needs to be re-run. There remains a small but non-zero chance of undetected errors that can only be detected by application level consistency checks or when the results are analysed.

As there does not seem to be any way of preventing errors occurring so machines and software need to be designed to handle these errors. Hardware designs can incorporate greater redundancy in order to detect and correct a greater fraction of these errors. This will in turn require a larger fraction of the hardware and energy budget and introduce additional gates into critical circuits which will inevitably reduce the effective capability of the system to some extent. The increased complexity of

resilient circuit design is also a problem though this should be addressable by improvements in the chip-design tools.

Application codes could also take a larger role in fault handling, increasing the number of application level consistency checks to help detect errors and, where possible, using fault-tolerant algorithms to recover from those errors. The hardware and system software only needs to provide an acceptable level of reliability for the least fault-tolerant of its target applications, which might be a consideration when co-designing applications and machines.

5 Potentially Disruptive Technologies

5.1 Chip Stacking

One of the most exciting potentially disruptive technologies is the emerging ability to directly integrate multiple integrated circuits using vertical chip stacks. Unlike bonding-wire technologies the inter-chip connections are spread across the contact area of the chips allowing a much higher density of interconnection between the devices. More importantly in future generations of the technology the number of inter-chip connection might be expected to scale more closely in line with the number of transistors in the device.

The simplest of these technologies is a variant of flip-chip mounting where two ICs can be mounted face-to-face using small balls of solder to connect metal pads on their surfaces. However more advanced techniques use Through-Silicon-Vias (TSVs), allowing multiple ICs to be stacked vertically to form a 3 dimensional device. A TSV is a vertical electrical connection passing from one face of the die to the other. These are usually manufactured by constructing a deep metal filled pit, then “thinning” the IC until this metal feature is exposed on the rear face. Various commercial products already use TSVs to mount ICs on a Silicon interposer. These are large silicon chips used instead of circuit boards. Because interposers are passive (they have no transistors constructed on them) they can be economically manufactured at a larger size, but they provide a much higher degree of inter-connectivity than conventional circuit boards. For example the Xilinx Virtex-7 FPGA uses this approach to build larger FPGAs than would otherwise be possible with current technology [9].

The possibility to stack multiple levels of IC to form true 3D circuits has been an active research topic for several years. The traditional 2D layout of ICs has become a problem as chip size and complexity increases, separating the component parts by greater distances. In addition, larger numbers of metal layers need to be manufactured within the IC to increase the opportunity for wires to cross each other. Both these problems could be addressed by chip-stacking. This approach seems to be particularly promising for building memory systems. Memory systems typically use stepped designs where a basic cell design is repeated a large number of times in a geometric pattern. This approach can be naturally translated into 3 dimensions with a 3 dimensional memory system built out of multiple layers of the same chip design. Current chip design tools do not provide good support for chip stack designs which seems to be holding back the development of more complex designs for the moment. Chip stacking also allows different types of device (such as DRAM cells and high-speed logic) to be manufactured using different processes and then combined into a single chip-stack to form a composite device [10]. There are two (not necessarily incompatible) architectural approaches that could be used with this kind of memory. The first is to stack the DRAM chips onto a Rambus-like memory interface chip to create a stand-alone memory device accessed through high-speed serial or optical interfaces. Alternatively the chips could be stacked directly on top of the processor configured either as caches or as directly attached memory. The high degree of connectivity provide by the TSVs would allow the memory to be partitioned into large numbers of independent banks. These could be used to give each core dedicated access to part of the stacked memory or interleaved to give higher throughput to a shared memory space. However the amount of memory that could be stacked directly on top of a processor is likely to be limited by thermal constraints.

This technology now seems close to commercialisation. It has been available in niche markets for some time from specialist companies such as Tezzaron [11], more recently the Hybrid Memory Cube consortium [12] was formed. This is a consortium including Altera, IBM, Micron, Open-Silicon Inc., Samsung and Xilinx that are aiming to develop and standardise a new class of memory device by combining high speed logic dies with a stack of TSV bonded memory dies. These will be stand-alone memory devices accessed through high-speed serial links. IBM and Micron have announced the

decision to begin production and are claiming x15 speed increase and 70% energy reduction for this technology [13].

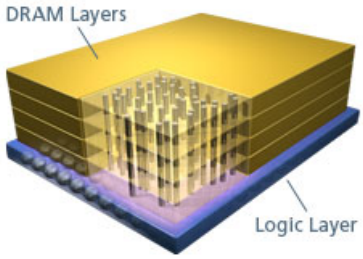

Figure 1 Stacked die memory

It is not unreasonable to expect that experience developed in 3D memory devices of this type will eventually be used to stack memories directly on top of the processor die. A number of proof-of-concept devices along these lines were presented at ISSCC 2012.

5.2 New Memory Technologies

A great many different memory technologies have been proposed in recent decades. So far none of these have significantly eroded the position of DRAM as the dominant memory technology. The most dramatic recent trend has been the rise in importance of non-volatile flash memory, however so far the relatively low speed of flash memory and restrictions on the number of read/write cycles a device can support have meant that the impact of flash memory has primarily been as a replacement for disks rather than for DRAM.

However the energy costs of DRAM are seen as a major potential problem for future large systems. A non-volatile memory technology with reasonable cost/performance and capable of supporting a high re-write rate could significantly change this position, either by replacing DRAM entirely or by increasing the performance of virtual memory, allowing a reduction in the size of DRAM memory systems. Several new technologies show promise in this area though they would be expected to become significant as a storage device before becoming a serious replacement for DRAM:

5.2.1 Phase Change Memory

This a type of non-volatile memory that stores data in reversible crystalline/amorphous phase changes of a material. This technology is commercially available for niche products, for example from Micron [14]. It supports many more update cycles than conventional flash memory and has read access times close to those of DRAM, though the write times are much more. Current devices can store much less data per device than current DRAM devices though this gap is expected to narrow at the next generation [15].

5.2.2 Memristors/Resistive-RAM

Memristors are a type of electrical circuit where the resistance of the circuit depends on the “history” of the voltage applied to the circuit. These can obviously be used to construct non-volatile memories. Recent research has concentrated around titanium-dioxide memristors. A joint project to commercialise the technology on the 2013 timescale has been announced by Hynx and HP [16]

5.3 Silicon Photonics

The increasing energy costs of communication are driving the uptake of optical interconnects. Compared to electrical signalling optical interconnects have significant advantages. They support high speeds, are very energy efficient and are robust against cross-talk and interference. Optical technologies already dominate long range communication and are increasingly being used down to the inter-rack level. A single optical fibre is also capable of carrying multiple independent signals by utilising multiple

light frequencies, a technology called Wave-Division-Multiplexing. Recent research indicates that this trend will continue with optical connections eventually being made directly to the processor. Research into Nano-photonics (for example at IBM [17]) has demonstrated the ability to build all the necessary optical components (emitters, detectors, wave-guides, WDM transceivers and switches) directly on the same silicon with CMOS logic, with minimal additions to the manufacturing process. Alternatively the optical transceivers could be manufactured using different processes and combined with the other components as part of a chip-stack. This opens the possibility of optical connections becoming more common at shorter length scales replacing the currently ubiquitous electronic high-speed serial connections. Though fully optical switching is possible, and can be implemented using silicon photonics, this is circuit switching rather than packet switching. Even for long range networks where optical networking is the default technology WDM and optical switching is generally used to set up long lived virtual circuits and general traffic routing takes place electronically. Even relatively static virtual circuits could still be useful in an HPC context to provide fully optical short-cuts between distant parts of the network, reducing the number of routers that need to be traversed. This effectively builds a more richly connected virtual communication topology embedded in a simpler (and therefore cheaper) topology of physical connections. In hierarchical networks WDM can be used to increase the bandwidth of high-level links in the hierarchy.

6 Possible Impact on Exascale Machine Architectures and Software

Energy use seems to be the dominant design constraint on future Exascale systems. Not only the energy needed to power the system but also the energy needed to cool it. These energy constraints should have the greatest impact on the design of the memory system and the communication network as these technologies are expected to have poorer energy scaling over time.

It seems highly desirable that the nodes will be manufactured as single compact integrated devices in order to minimise the communication costs between different parts of the node. This would be similar to current SOC devices except that chip-stacking could be used to extend the designs to three dimensions, increasing the density further as well as allowing a mix of different chip processes. Mainstream processors aimed at the cloud/data-centre markets are subject to the same technological trends as HPC systems and may evolve in the same way.

Though the internal parallelism within a node will probably continue to increase with Moore's law, however the cost (both time and energy) of communicating between different parts of the node is expected to increase. This may well result in a very non-uniform memory hierarchy within a node. In addition processor architectures may provide fast local memory that is only accessible from the associated core. To make effective use of such a system operating systems, compilers, run-times and application codes may have to ensure locality at a much smaller scale than the nodes. Application codes need to be written to only make infrequent use of global data structures and to copy data to thread-private variables rather than make multiple accesses to the same data. Operating-systems, compilers and run-times then would need to cooperate to ensure that these thread-private variables are located in the most efficient parts of the memory system relative the core executing the thread.

Even when taking into account possible developments in memory technology the energy efficiency of the memory system is expected to scale less well than other parts of the system. It is therefore reasonable to assume that main memory capacity will not be increased to the same extent as the computational capability though some authors assume that it will be necessary to introduce an additional memory layer using a slower non-volatile memory technology to preserve the 10-flops per byte ratio seen in current systems.

It is very unlikely that Exascale systems will be able to support global cache-coherency. As mentioned previously cache coherency protocols require large amounts of additional communication. However a single global address space (with un-cached access to remote locations) might still be possible which would facilitate the use of PGAS programming models. Even then it would not be desirable to simply map the remote memory into the memory space of the local processor. The round-trip latency needed to access the remote location would result in very high access latencies if remote data was accessed directly with conventional load/store instructions, as well as resulting in single word transactions that would make inefficient use of the network. Instead better performance would be achieved by providing a separate RDMA facility to copy larger blocks of data between local and remote memory.

Processor clock frequencies will not be significantly increased and may very well decrease to keep energy consumption within acceptable limits. This in turn implies that Exascale performance can only be achieved via a significant increase in parallelism. Low clock speeds will exacerbate the Amdahl law limit on application scaling. Replicated work like subroutine calls and software overheads in the operating system threading and message passing library take a time proportional to the clock speed and contribute towards the sequential fraction of the program.

Optical networking will have to become the dominant networking technology over longer scales. Eventually this will have to extend down to the connections between individual nodes though this may not occur in the first generation of Exascale systems.

7 References

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics,* vol. 38, no. 8, 19 April 1965.

[2] "The Impact of Dennard's Scaling Theory," *IEEE SSCS News,* vol. 12, no. 1, 2007.

[3] N. S. Kim, D. Blaauw, T. Mudge, K. Flautner, J. Hu, M. J. Irwin, M. Kandemir and V. Narayanan, "Leakage Current: Moore`s Law Meets Static Power," *Computer,* December 2003.

[4] "Intel 22nm 3-D Tri-Gate Transistor Technology," [Online]. Available: http://newsroom.intel.com/docs/DOC-2032.

[5] "Rambus web site," [Online]. Available: http://www.rambus.com.

[6] "Intel Many Integrated Core Architecture," [Online]. Available: http://www.intel.com/content/www/us/en/architecture-and-technology/many-integrated-core/intel-many-integrated-core-architecture.html.

[7] J. Dongarra, "Exascale Computing Panel," in *ISC2010*, 2010.

[8] J. G. Koomey, S. Berard, M. Sanchez and H. Wong, "Implications of Historical Trends in the Electrical Efficiency of Computing," *IEEE Annals of the History of Computing,* vol. 11, pp. 1058-6180, 2011.

[9] "Xilinx Stacked Silicon Interconnect Technology," [Online]. Available: http://www.xilinx.com/technology/roadmap/ssi-technology.htm.

[10] G. H. Loh, "3D-Stacked Memory Architectures for Multi-Core Processors," in *International Symposium on Computer Architechture proceedings*, 2008.

[11] "Tezzaron semiconductor website," [Online]. Available: http://www.tezzaron.com.

[12] "The Hybrid Memory Cube consortium," [Online]. Available: http://www.hybridmemorycube.org.

[13] "HMC, Micron Technology Inc.," [Online]. Available: http://www.micron.com/innovations/hmc.html.

[14] "Micron Parallel PCM," [Online]. Available: http://www.micron.com/products/pcm/parallel_pcm.html.

[15] G. Atwood, "The Evolution of Phase Change Memory," 26 July 2010. [Online]. Available: http://www.micron.com/get-document/?documentId=5539.

[16] "HP Collaborates with Hynix to Bring the Memristor to Market in Next-generation Memory," [Online]. Available:

http://www.hp.com/hpinfo/newsroom/press/2010/100831c.html.

[17] Y. Vlasov, “CMOS nanophotonics for Exascale,” 1 December 2010. [Online]. Available: http://www.research.ibm.com/photonics/publications/SEMICON_Tokyo_12_1_2010.pdf.