COMMUNITY RESEARCH REPORT

# High Performance for All

## Responding to the needs of compute-intensive workloads

Jon Collins, Tony Lock and Dale Vile, March 2010

*Most medium and large organisations need to run 'compute-intensive' applications of some form. While High Performance Computing (HPC) is not new, it has traditionally been seen as a specialist area – is it now geared up to meet more mainstream requirements?*

## EXECUTIVE SUMMARY

### Compute-intensive application workloads are not industry-specific

Today's computer systems are more powerful than ever. But also on the increase, are the needs of medium and large businesses to run highly demanding workloads that make maximum use of available computer power. Understandably, larger organisations have more requirements than smaller organisations, and such workloads are more prevalent in certain verticals such as financial services, telecoms and research. However the need is evident across the board.

### Not all compute-intensive needs are currently being met

More often than not, such demanding workloads are being run in batch mode rather than interactively, which cannot be ideal: smaller organisations (with sub-5,000 employees) in particular tell us that their compute-intensive needs are not being met. Hurdles to solving this problem are not only to do with finding sufficient time and resources, but also involve both existing applications and current infrastructure, suggesting legacy issues are holding organisations back.

### The gap is closing between specialist HPC and more mainstream, compute-intensive IT

While traditional HPC may have been about building custom compute platforms for specialised applications, today's HPC is not as isolated as many might think. Specialists no longer see HPC as a separate domain; in addition, HPC is increasingly reliant on commodity equipment and software. While the gap with mainstream computing may be closing, the journey is not over yet, as HPC systems still require considerable customisation compared to general-purpose machines.

### The HPC community has much to give in terms of skills and experience

Lessons learned in HPC environments are equally applicable in delivering infrastructure to support more general compute-intensive workloads – for example, architecture and design skills around networking and communications, power, cooling and so on. Indeed, the HPC community is better placed than most to identify candidate workloads that could benefit from the HPC treatment – candidates which might not be evident to those who are not HPC-savvy.

### Meanwhile however, the evolution of HPC itself needs to accelerate

While demand for compute-intensive platforms may be high, the traditional supply chain for HPC is not changing that fast. It may be that developments in other areas of IT, such as adoption of virtualisation and cloud-based hosting models, may increase momentum in this area. In particular it is generally agreed that automation and configuration tools are lacking – though this will inevitably change as such models become more widely used.

*This report is based on the findings of a research study completed in November 2009 in which feedback was gathered from 254 predominantly IT professionals with direct or indirect experience of high end server computing environments. The report was sponsored by Microsoft, though the study was designed, executed, analysed and interpreted on a completely independent basis by Freeform Dynamics.*

## Who wouldn't want high performance?

Delivering IT has always been an act of balancing the cost of resources with the levels of service delivered: the more performance you want, the more you will have to pay. Even though processor power has increased by many orders of magnitude over the decades, so have both user demands and the size and scale of workloads to be run. As a result, decision making around IT server infrastructure has not changed all that much.

Specifically, when it comes to buying computer systems for higher performance requirements, the following criteria generally have an influence:

- **Infrastructure continues to evolve** – today's higher-specification hardware platforms supporting virtualisation enable organisations to do more with less, and support the ultimate goal of making best use of available resources [1].
- **The highest performance is not necessary for everything** – indeed, cost pressures, power and cooling, and more general sustainability requirements drive the need to scale the amount of IT according to actual need.
- **However, some applications are inherently high-demand.** There will always be a place for certain applications that by their nature need to 'go large' on IT processing resources.
- **There remains a "High Performance Computing" (HPC) domain which deals with the highest end workloads.** These are traditionally used to supply specialist business requirements.

Before we move on, it is worth expanding on what is meant by HPC. This area is generally understood to have evolved out of 'supercomputing', which has traditionally referred to building and operating extremely powerful, highly customised computers capable of meeting the modelling, simulation and graphics-intensive processing needs found in sectors such as academia, financial services, research and media.

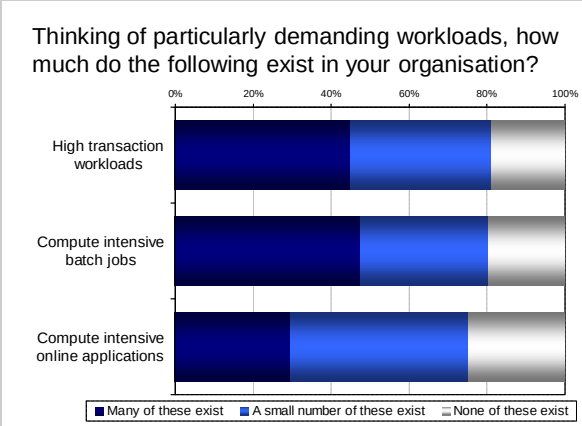
HPC has traditionally been considered as a black art, the domain of specialists at the very leading edge of IT. While it is quite niche relative to more general purpose computing, HPC is a highly competitive area – there exists, for example, a top 500 list of the most powerful computers in the world. As this report shows however, HPC is itself commoditising. The theoretical gap between mainstream high-performance requirements and those which qualify as HPC is closing, in a way that can bring more cost-effective capabilities to both sides.

As the two sides become closer, there is much that can be learned from the HPC community to help define and deliver IT systems for mainstream high-performance use. As documented in Appendix A, this report is based on an online research study which by its nature was self-selecting – it was up to respondents to decide whether to participate. Inevitably therefore, the sample contains a higher proportion of people with an interest in the area of compute-intensive IT in general, and HPC in particular. In this context, the nature of the sample acts to our advantage, as respondents are inherently going to have more experience than if the sample was taken from a random list.

As a result, the aim of this report is to show not only how the gap is closing, but also to highlight key areas of learning from the HPC community that can be built upon as technology continues to commoditise and become more broadly relevant.

## The prevalence of compute-intensive workloads

Taking the mainstream perspective first, what merits being qualified as "high performance"? From the business point of view, the answer is, "anything that needs to be". In the research we wanted to distinguish between the high-transaction workloads (for example banking or payment processing systems) which have been relatively well served in their own industries, and focus on other types of compute-intensive workload, either running in a batch or online/interactive mode (Figure 1).

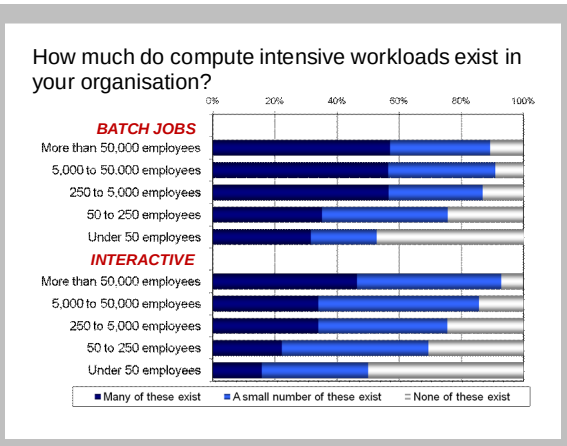


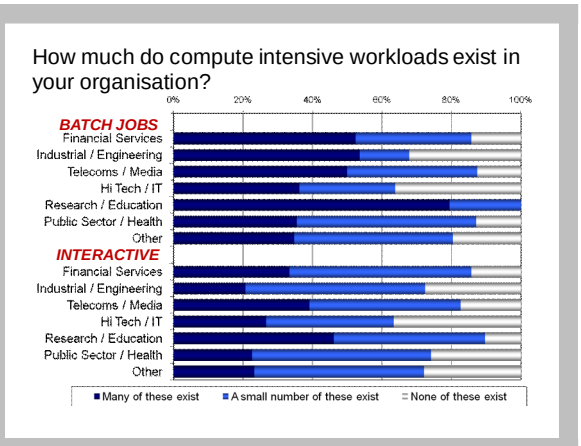

*A more significant proportion of compute-intensive workloads are running in batch mode, rather than interactively.*

*Figure 1*

Drilling into the research sample, it is clear that in many organisations, general compute intensive requirements sit alongside transaction processing needs when it comes to high-end computing. While many such requirements can be met in an interactive manner, a larger number are still running in batch mode, that is, a job is sent for processing with the results being returned up to several hours later. We shall return to this point later in the report.

It is not surprising to see compute intensive workloads being more prevalent in larger organisations, nor indeed in research and educational establishments. Nevertheless a clear need exists across all but the smallest of companies, and in all sectors (Figures 2 and 3).

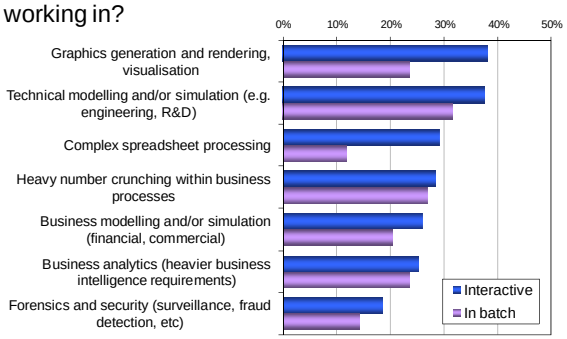



*Compute-intensive workloads are more prevalent in larger organisations, and in certain verticals.*

*Figures 2/3*

We can break this view down somewhat further, in terms of the kinds of workloads involved. Graphics visualisation and simulation are well represented, but so too are heavy spreadsheet processing and number crunching, business modelling and analytics, in both batch and interactive modes (Figure 4). While we would expect the former categories as they are more traditionally associated with HPC, it is interesting to see how highly complex spreadsheet processing and heavy number crunching also figure.




*A variety of traditional and more business facing workloads are represented.*

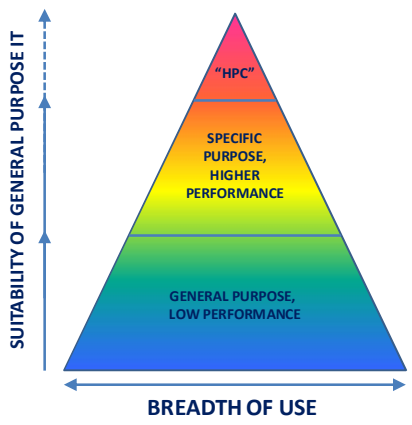
*Figure 4*

But why does higher performance matter in these (latter) scenarios? The number one reason is of course "time", but this answer isn't sufficient in itself. The timeliness of completing a compute-intensive task feeds such things as:

- Productivity, in terms of the amount of work an individual can complete to achieve his/her own goals.
- Value, to determine whether the job is worth doing at all – if it is going to take too long, it may well miss the window for the answers to be useful.
- Responsiveness, for example to support the required cycle times for a given business process, or more generally to ensure business service levels are maintained.
- Risk, for example if delays leave the organisation in some way vulnerable.

A clear example of risk in this context was given to us by a CIO at a payroll company, who told us how his monthly payment processing window was about eight hours: if anything went wrong in this period, his customers would suffer. "The smaller I can get that time, the less risk I have," he told us.

Of course there is no such thing as an exclusively compute-intensive application: a spreadsheet can be used just as well for low-performance data analysis as for large-scale number crunching. In reality, a spectrum of performance requirements exists for applications across the board. At the lower end of the spectrum for example, we have programs that have low requirements in terms of latency, throughput and so on. These tend to be the more common applications in use (Figure 5).




*General purpose IT resources become less appropriate for higher performance processing needs.*

*Figure 5*

Towards the higher end of the spectrum are more specific workloads with higher performance requirements. These applications will be less prevalent either because they are used by a smaller number of people, or because they are used less often. Of course a poorly written application can always benefit from additional RAM or processor power. But even well-written applications can hit performance thresholds which can render them unsuitable or inefficient in terms of time, cost, value or risk.

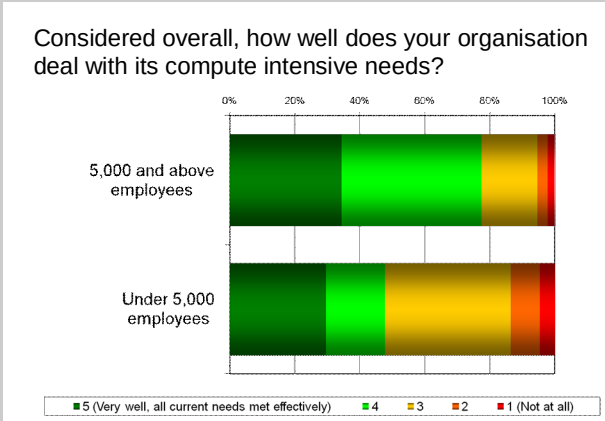
Right at the top end of the performance scale, are those workloads traditionally considered as applicable to HPC. We shall look at the specifics of HPC in a moment, but for now let us consider two points illustrated by the figure:

- first, that applications in more general use are becoming increasingly demanding, as are their users;
- and second, that commoditisation pressures are driving down costs of both general purpose and HPC infrastructure.

Thinking back to the business perspective, the lines between the different categories are largely theoretical – suffice to say that business needs can be met more effectively, if more general compute intensive needs can be better accommodated by previously unattainable computer systems. Let's look at why it is important to do so.

## The scale of the problem

So, how much of a challenge is this? To be clear, this is not a "sky falling down" situation. However respondents did make it plain that that their compute-intensive needs could be met more efficiently than they are currently (Figure 6). Larger organisations are not in such bad shape – over three quarters think they are adequately covered. However this figure drops substantially, to under half for organisations with fewer than 5,000 employees.



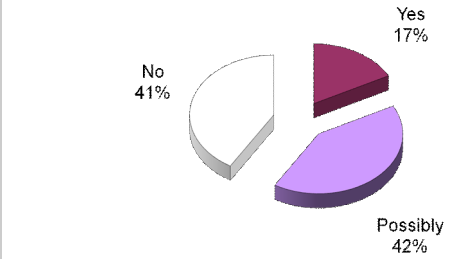

*While over three quarters of larger companies feel their needs are met, under half of organisations under 5000 employees think the same.*

*Figure 6*

In companies of all sizes, one area where improvements could be made is to close the gap between batch and online processing. While not all jobs need to be run interactively, we know from other research [2] that some of the areas already highlighted in Figure 4 – notably around analytics and business modelling – are more desirable as interactive applications, rather than the batch way in which they are frequently delivered.

From this research sample, we can see a good percentage of respondents – 17% – believe that certain jobs would benefit from being migrated to execute on an interactive basis (Figure 7). This would be particularly the case for workloads supporting ongoing business operations, for example to avoid process interruptions and delays while waiting for batch activity to complete. Note that self-

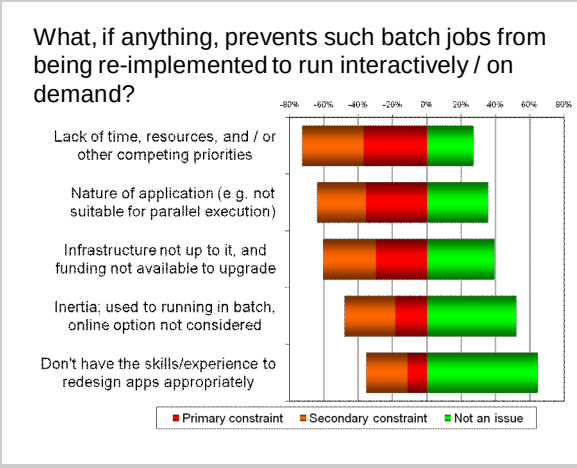
selection is probably acting against us – from previous research [2] we know that if we had asked the business users waiting for the results of such jobs, the number would likely be higher.




*17% of respondents felt that specific batch jobs would benefit from being run interactively. And a further 42% saw the possibility.*

*Figure 7*

If we put aside applications that are simply not suitable for running interactively, the challenges when considering the move from batch to online comes down to the old staples, "time and money" (Figure 8). For anyone who has been working in IT for a while, this is reminiscent of when transaction processing was largely a batch-based affair – it is interesting to reflect on how this has changed over the years to being almost exclusively interactive, as business demands have increased.




*Time and resources are the number one criterion, followed by nature of application.*

*Figure 8*

Second in the list is the nature of the application, followed by lack of adequate infrastructure – which suggests that existing applications and infrastructure may be holding things back. It is interesting to note that skills/experience are seen as far less of an issue, which is a good segue into HPC itself, and what those working in this area can bring to the party.

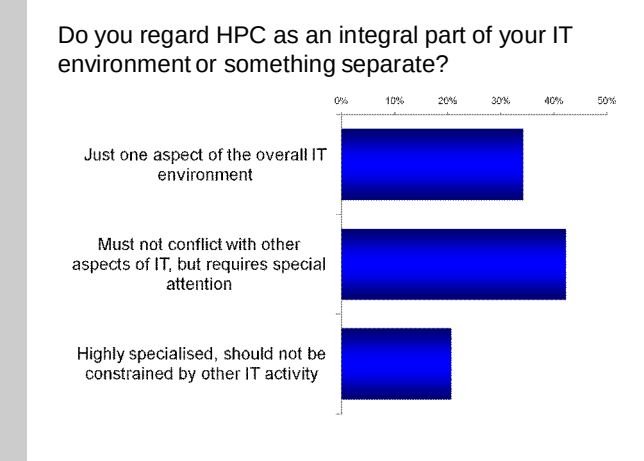
## Bringing HPC into the mix

So, what of High Performance Computing? HPC is often considered to be a highly specialist area, the domain of supercomputers that cost the equivalent of the national income of a small country to buy and operate, or indeed highly complex Linux and UNIX clusters that are lovingly put together and nurtured by talented technicians who dedicate their lives to performance and tuning.

However, the view from the survey sample is that HPC is perhaps more 'mainstream' than some quarters would like to make out. To clarify perceptions, we asked the following question:

> **When considering HPC, as things stand today, do you regard it as an integral part of your overall IT environment or something distinct that stands separately from it?**

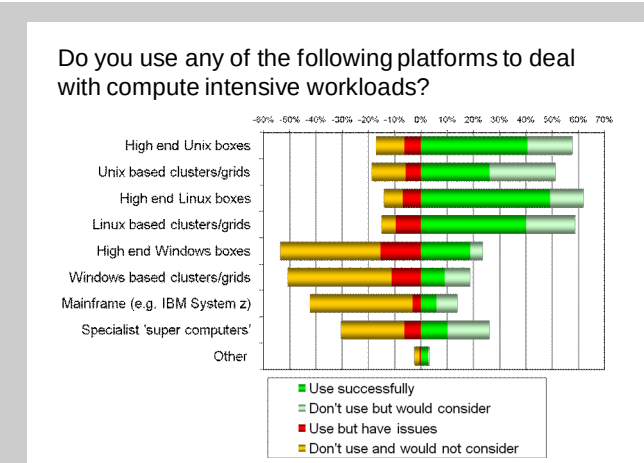
As can be seen from Figure 9, a full third of respondents saw HPC as just one aspect of the overall IT environment. Indeed, the majority view from participants was that HPC should work in harmony with the overall business IT landscape and operations, even if it needs specialist attention.




*Only 21% of respondents saw HPC as a highly specialised area that needed to operate in isolation from the rest of IT.*

*Figure 9*

Further corroboration of this view can be gained when we look at the platforms organisations are using to run 'high performance' or 'compute intensive' workloads. From an operating system perspective, proprietary UNIX is still playing its part but Linux is now at the number one position in the list (Figure 10). Meanwhile it is interesting to see that, despite being a relatively new entrant to the space, Microsoft is today more widely favoured than 'specialist super computers'. It is also clear that the mainframe is not widely utilised to handle these workloads reflecting its usage to deliver more mainstream, transaction-oriented business applications.

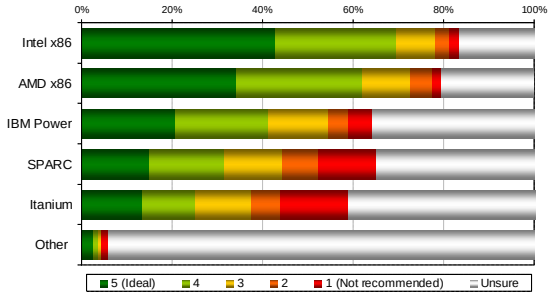



*Linux is now the number one operating system platform for HPC, followed by UNIX.*

*Figure 10*

It is equally clear that commercially available chip sets based on the x86 architecture now dominate the HPC market, displacing specialist suppliers. Intel and AMD x86 offerings have clearly taken

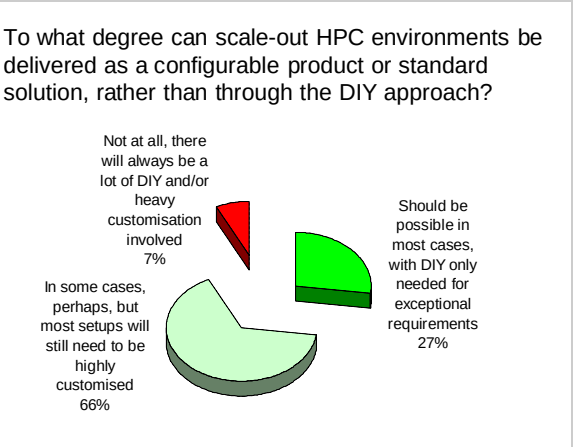
over the bulk of the market in terms of widespread perceptions of their suitability to handle such workloads (Figure 11).




*The majority view is that x86-based chipsets are most suitable/desirable for HPC use.*

*Figure 11*

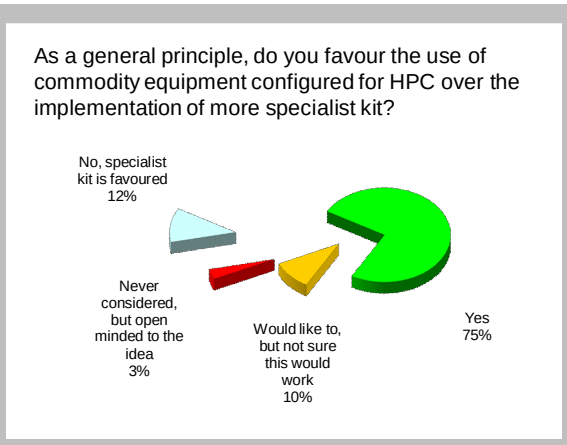
So momentum does indeed exist to bring capabilities traditionally associated with HPC more into the mainstream. From Figure 12 for example, it is encouraging that over a quarter of participants think a more configurable standard solution approach should be applicable for HPC workloads in most cases. Let's be clear however – two thirds of respondents believe that although it may be possible in some cases, most setups will still need to be highly customised.

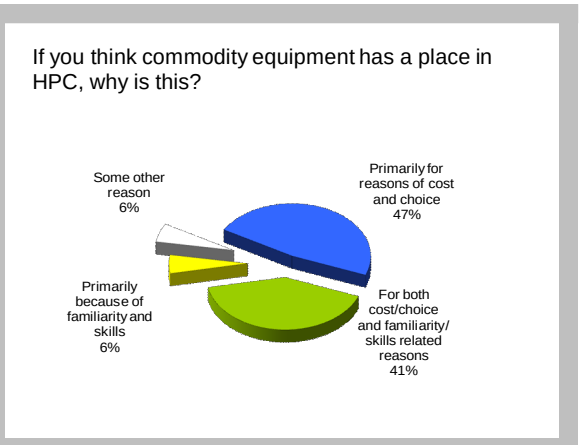



*While standards-based approaches are broadly applicable in a quarter of cases, the majority view is that most setups will require customisation.*

*Figure 12*

On the surface this appears to be a conflict; but in reality it is recognition from the respondent base that there will continue to be highly specialised workloads, which in turn require highly specialised treatment. Meanwhile however, the drive towards use of commodity platforms for HPC is staunchly supported by around three quarters of respondents (Figure 13). Of the remaining respondents, some 13 percent are either open to the idea of using commodity equipment or would like to do so but would need to be convinced that it would work for them. Only one organisation in eight actively favours using specialist equipment in their HPC operations.




*A significant proportion of respondents favour commodity kit over specialist kit for HPC, primarily for reasons of cost and choice.*

*Figures 13/14*

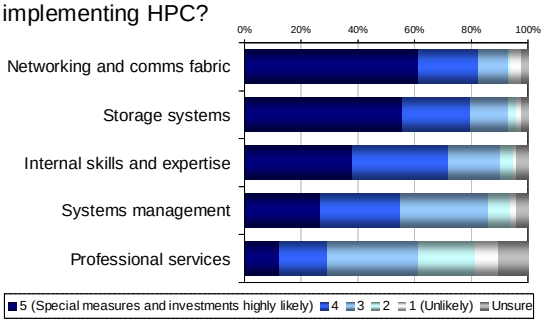
As illustrated in Figure 14, the major factors favouring the use of commodity equipment in HPC solutions are cost and choice, though familiarity and skills are also significant. On this latter point, it is important to remember that 'familiarity' can be a career choice for IT professionals, who will be more keen to develop skills that are transferrable, rather than gaining expertise in a proprietary platform, only for it to become obsolete.

## Taking things forward

Factors such as those mentioned above are directly driving the continued commoditisation of both software and hardware. This will enable more workloads to be considered as appropriate for the HPC environment, or indeed – for HPC capabilities to be seen as more appropriate for general compute intensive tasks.

This raises the question however – just how likely are those who are not HPC-savvy to spot such opportunities? People may not today recognise a need for HPC type systems simply because they are unaware of what HPC can achieve. For this reason as much as any, it is important that the more general domain of compute-intensive IT benefits from the experience and knowledge built within the HPC community.

What other lessons can be learned? Thinking back to the 'specialist HPC' audience, it can be taken as read that HPC skills sets in terms of building, configuring and operating high-performance computer systems will continue to be important. But architectural design aspects such as networking and storage appear higher on the list (Figure 15).



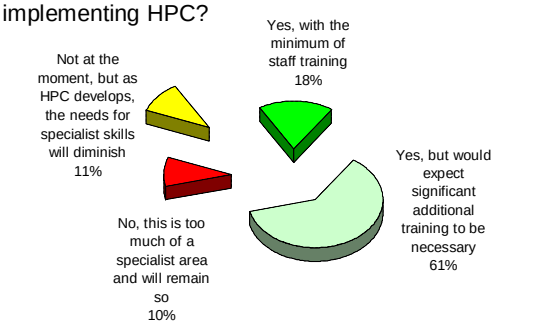

*Architecture and design skills for networking and storage are top of the list when it comes to implementing high-performance computing.*

*Figure 15*

This reflects what we have seen in other studies [1] about getting the physical architecture right – aspects which were further emphasised in the freeform responses from participants. When we asked what we might have missed in terms of special measures and investments, for example, respondents paid particular attention to the importance of power and cooling when designing and provisioning HPC systems:

> **"You need to consider cooling and power measures, this is where a lot of expense has to go."**
>
> **"Power and facilities. Have you any idea of what a sizeable grid/cluster of PCs consumes and how much heat it emits?"**
>
> **"Cooling. Cooling. Cooling. Did I mention Cooling? Oh, and ELECTRICITY SUPPLY. The real skill and the bit no one has training in... making sure those systems don't all starve from lack of power, or die of heat death."**

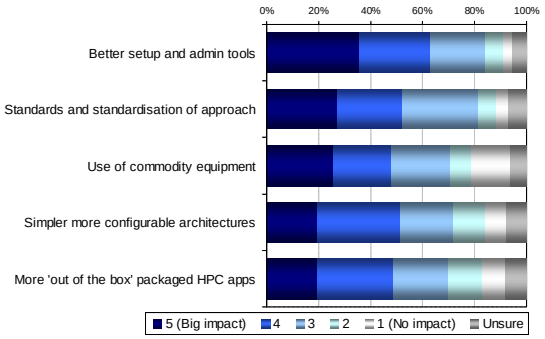
Remembering the 'career' point above, further evidence of the gap closing is the decreasing dependence on specialist skills. One in five say only a minimum of cross training is required (Figure 16), and a further 11% feel the need for specialist skills will diminish. Meanwhile, at the other end of the scale, a hard core of respondents (10%) still believe that the high priests of the supercomputer or Linux cluster will prevail.




*61% of respondents see additional training required on top of general computing skills.*

*Figure 16*

From the standpoint of closing the gap, clearly it is important to reduce the dependency on hard-to-get skills. This is one area where better automation and configuration tools may help – this may in turn benefit from the familiar virtuous circle between standardisation, commoditisation and tooling, in which tools are more likely to be made available as standards are adopted and underlying technologies commoditise accordingly (Figure 17).




*Better tooling would certainly help reduce the dependency on specialist HPC skills, as would more standardisation and commoditisation.*

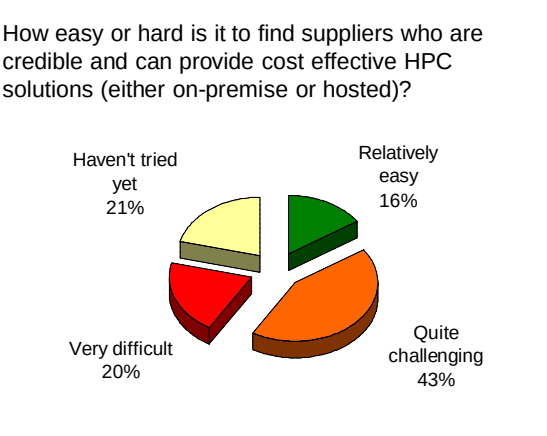
*Figure 17*

## Conclusion

While HPC may have started out (no doubt through necessity) as a specialist area requiring highly customised kit, there is clear evidence that this is less and less the case. This can only be seen as a good thing – compute-intensive workloads of all types will benefit, and there are clear business gains to be had from turning batch jobs into interactive tasks.

However, the observation that smaller organisations are struggling much more than larger ones in this whole area raises a flag about the accessibility of the technology in this space. While the technology area may be commoditising and skills are becoming more aligned with general purpose computing, it's clear that HPC is still a specialist domain with many variables to consider.

If there is anything lacking right now, it is momentum. There is nothing to stop HPC from becoming a more mainstream discipline, to the benefit of companies of all sizes. However progress is slow. In an ideal world, the supplier community would be able to help with this, but finding credible partners is clearly a challenge at the moment (Figure 18). While evidence is emerging that the realities here are changing, there is perhaps a call to action for the vendor and professional services community to pay more attention to generic as well as specialised HPC needs.




*Almost two thirds of respondents see it as a challenge to obtain cost-effective HPC solutions.*

*Figure 18*

While this may appear a down note to some, it's actually an opportunity. Businesses across the board currently depend on running certain tasks in batch mode because there is no other choice. When you think about what batch jobs are in your organisation – be they complex data processing and business intelligence, or business modelling and simulation, or whatever they happen to be – you can take heed of the fact that opportunities may exist to make these more interactive. In addition, the broader availability of more compute-intensive capabilities may open up new opportunities for data processing that were not previously cost-effective.

At the same time, this commoditisation is clearly not happening in isolation – indeed, it is driven by developments in other areas and this will no doubt continue. In particular for example, both virtualisation and cloud computing may accelerate the commoditisation of HPC, as may the use of Gigabit Ethernet in the data centre. Whatever your exact situation with respect to compute-intensive facilities, it may be worth reviewing it, not only in the light of what might now be possible, but also if you are considering adopting any of these technologies.

Whatever your plans, however, do ensure that you take into account the lessons learned by the HPC community over the years.

## References

All the reports referenced here are available for free download at www.freeformdynamics.com


[1] Server Virtualization in Context — Freeform Dynamics, 2009

[2] The BI Inflexion Point — Freeform Dynamics, 2007


The research sample is provided in Appendix A.

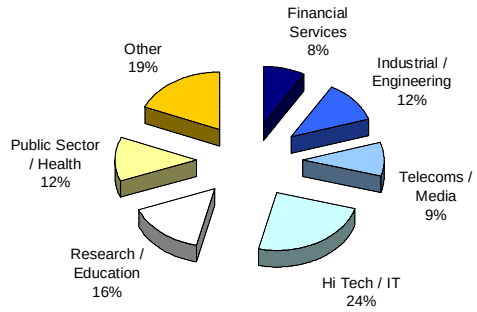
# Appendix A

## RESEARCH SAMPLE

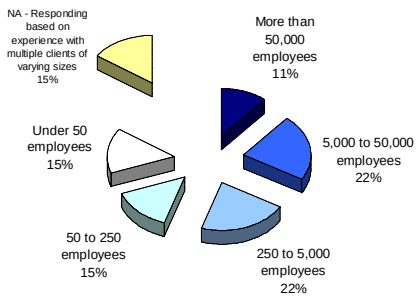
The findings presented in this research note are a subset of those from a larger study conducted in November 2009 exploring the general domain of high end server computing. The study was designed, executed and analysed on an independent basis by Freeform Dynamics.

An online Web based questionnaire was used to gather information and in total, feedback was received from 254 respondents, predominantly IT professionals with direct or indirect experience of high end server computing environments. Sample distribution by industry and organisation size was as follows:

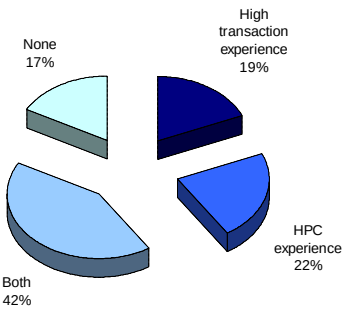
**Participants by industry sector**




**Participants by organisation size**

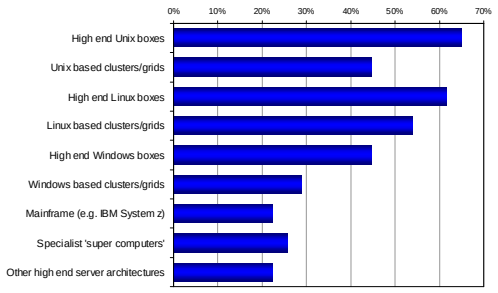



Experience level summary






Which of the following specific architectures do you have experience or knowledge of?






## Acknowledgements

Our thanks go to all those who participated in the study from the readership of *The Register*, whose feedback has been invaluable to provide insights into practicalities as well as opportunities in this interesting but complex area. You help is greatly appreciated.

## About Freeform Dynamics

Freeform Dynamics is a research and analysis firm. We track and report on the business impact of developments in the IT and communications sectors.

As part of this, we use an innovative research methodology to gather feedback directly from those involved in ITC strategy, planning, procurement and implementation. Our output is therefore grounded in real-world practicality for use by mainstream IT and business professionals.

For further information or to subscribe to the Freeform Dynamics free research service, please visit www.freeformdynamics.com or contact us via info@freeformdynamics.com.

## About Microsoft

Founded in 1975, Microsoft (Nasdaq "MSFT") is the worldwide leader in software, services and solutions that help people and businesses realise their full potential.

For more information on Microsoft's virtualization solutions, please visit http://www.microsoft.com/virtualization/.